MICHAEL PREMINGER

The Uexküll Approach

Evaluation of Multivariate Data Organizations for Support of Visual Information Retrieval

■

ACADEMIC DISSERTATION
To be presented, with the permission of
the Faculty of Information Sciences of the University of Tampere,
for public discussion in the Auditorium Pinni B 1097,
Kanslerinrinne 1, Tampere, on August 23rd, 2008, at 13 o'clock.

UNIVERSITY OF TAMPERE

ACADEMIC DISSERTATION
University of Tampere
Department of Information Studies
Finland

Distribution
Bookshop TAJU
P.O. Box 617
33014 University of Tampere
Finland

Tel. +358 3 3551 6055
Fax +358 3 3551 7685
taju@uta.fi
www.uta.fi/taju
http://granum.uta.fi

Cover design by
Juha Siro

Acta Universitatis Tamperensis 1333
ISBN 978-951-44-7407-1 (print)
ISSN 1455-1616

Acta Electronica Universitatis Tamperensis 748
ISBN 978-951-44-7408-8 (pdf)
ISSN 1456-954X
http://acta.uta.fi

Tampereen Yliopistopaino Oy – Juvenes Print
Tampere 2008

Abstract

The focus of this dissertation is to develop an algorithmic approach to the evaluation of visual retrieval in 3D from document databases represented as multidimensional vector spaces.

We are presenting the Uexküll approach, which entails the representation of document databases as multidimensional vector spaces. From such a space 3D projections are selected and downloaded by users. A downloaded projection is defined by three named axes, and it presents documents (possibly also index terms) that pertain to these axes as objects of some form, e.g. droplets, in a 3D scene. Users may navigate among these objects in relation to the axes, scrutinizing and retrieving documents of interest.

The evaluation approach is purely algorithmic. The aim is to evaluate the ability of a multidimensional vector space (termed a *data organization*) to facilitate the visual retrieval of relevant documents, placing such documents prominently along coordinate axes of downloaded projections. The evaluation approach is based on the Cranfield evaluation model, transforming the retrieved projection into a ranked list of documents. The Cranfield model is augmented with measures that gauge aspects of usability that are important for successful retrieval within an Uexküll environment. These aspects are the visibility of relevant documents and the separation of such documents from the non-relevant ones.

The presented evaluation approach is applied to data organizations of varying dimensionalities based on the singular value decomposition (SVD), rotated both orthogonally and obliquely, showing the effect of rotation and dimensionality on the quality of the data organization and the extent to which it facilitates Uexküll-based retrieval.

Acknowledgements

The creation of this dissertation has been a long process. Several institutions, colleagues, friends, and close relatives, have taken part in its conception and helped bringing it to a conclusion.

I would like to thank the Faculty of Journalism, Library and Information Science at the Oslo University College, its leaders and my colleagues there, for believing in me, exhibiting patience, facilitating and encouraging me in this work. I would also like to thank the the Department of Philosophy at the University of Oslo for their support in facilitating my education. Especially, I would like to thank the Department of Information Studies, and the FIRE group at Tampere University, Finland, for making me feel at home during the entire process and particularly towards the end of it.

I wish to thank my friend prof. Sandor Darányi from the University College of Borås, Sweden, for "selling" me the ideas that form the basis for this dissertation and for reading and commenting on my work. I also wish to express my gratitude to professor Inge Helland at the Department of Mathematics, University of Oslo and professor Ole Christian Lingjærde from the Biomedical Research Group at the Department of Informatics, University of Oslo, for their good advice and willingness to read, comment and criticize where necessary.

Special thanks also to my superiors in this period, Inger Cathrine Spangen, Egil Fossum, Liv Gjestrum and Øivind Frisvold for their support. Many colleagues have supported in different ways. Particular thanks to Ragnar Nordlie and Ragnar Audunson for good advice on various aspects of the process.

Thanks also to prof. Herman Ruge Jervell at the Department of Linguistics and Scandinavian Studies for his supervision and support.

Professor Kalervo Järvelin has been following the project through several years, both informally and formally. Thanks a lot, Kal, for your support, patience, hospitality and friendship.

Professor Paul Kantor and Dr. Birger Larsen have been the assessors of the manuscript. I wish to thank them for their comments which have undoubtedly strengthened the outcome.

Last but not least my wife, Inger and my children Iris, Natalie and Daniel deserve the greatest gratitude for enduring the life with a doctoral student that at times has also been a husband and a father. Thank you!!!

Oslo, 20 July 2008

Michael Preminger

Contents

I The Uexküll Approach **13**

1 Introduction and problem definition **14**
1.1 Dimension reduction of term-document matrices 15
1.1.1 The full form vector space 15
1.1.2 Obtaining reduced multidimensional spaces 16
1.1.3 An example of dimension reduction 17
1.1.4 Using the reduced dimensionality for visualization . . 20
1.2 The focus of the dissertation 21
1.2.1 Access to systems 21
1.2.2 Indexing 22
1.2.3 Contents of databases 23
1.2.4 Interpretable data organization 24
1.3 A summary of the idea and the name "Uexküll". 24
1.4 The structure of the dissertation and the research questions . 25
1.4.1 Focus and research questions 25
1.4.2 Summary - the structure of the dissertation 27

2 Approaches to information retrieval and visualization **28**
2.1 System-oriented information retrieval 29
2.1.1 The Cranfield studies 30
2.1.2 TREC 31
2.1.3 XML and INEX 31
2.1.4 Summary 32
2.2 Knowledge space and the vector space model 32
2.3 The three revolutions 33
2.3.1 The cognitive approach and ASK 34
2.4 Human intermediary-based retrieval as a model of automatic retrieval 36
2.4.1 Expert systems for query formulation 37
2.4.2 User and system modelling: the MONSTRAT and MEDIATOR models 38

2.4.3 User revealment . . . 39
2.4.4 Uexküll in the light of intermediary based modelling . . 40
2.5 Visualization in information retrieval . . . 41
2.5.1 Types of visualization within IR . . . 43
2.5.2 Research efforts within IR visualization . . . 43
2.5.3 Visualization in information retrieval - summary . . . 45
2.6 Summary . . . 46

3 The Uexküll visualization 47
3.1 Aims of the Uexküll system design . . . 47
3.2 Main principles of design . . . 48
3.2.1 Determining the threshold . . . 48
3.2.2 The direction metaphor . . . 49
3.2.3 Parameters determining the functionality of the interface 50
3.3 Interaction with the prototype . . . 51
3.3.1 The upper menu . . . 51
3.3.2 The scene . . . 52
3.3.3 Navigating within a scene . . . 54
3.3.4 Retrieval and presentation of documents . . . 55
3.3.5 Navigating among scenes . . . 55
3.4 Concluding remark . . . 57

4 Multidimensional methods for dimension reduction within IR 58
4.1 Introduction . . . 58
4.1.1 A brief historical overview . . . 59
4.1.2 Latent variables and the factor model . . . 60
4.1.3 Result interpretation . . . 60
4.2 Factor analysis as a model and a family of methods . . . 62
4.2.1 The linear factor model . . . 62
4.2.2 Initial solution and rotation . . . 63
4.2.3 Rotations and interpretation in vector space-IR . . . 64
4.2.4 Using rotations in this research . . . 65
4.2.5 Drawbacks of factor analysis methods . . . 65
4.3 Latent semantic indexing and the singular value decomposition . . . 66
4.3.1 The singular value decomposition . . . 66
4.3.2 Latent semantic indexing . . . 67
4.3.3 Methods derived from SVD/LSI . . . 69
4.4 LSI and related methods used in the Uexküll approach . . . 71

4.5 Treatment of binary term-document matrices 72
4.6 Points for further research 73
4.6.1 The true dimensionality of a term-document matrix . . 73
4.6.2 Methods that could take larger data sets 74
4.7 Limitations of using multivariate methods for Uexküll visualization 75
4.7.1 Properties of axes 75
4.7.2 Number representation errors 75

II Evaluation of the Approach 77

5 Evaluation of the Uexküll approach 78
5.1 Introduction 78
5.1.1 Evaluation in IR 78
5.1.2 Approaching the evaluation of Uexküll 79
5.1.3 Main idea and problems 80
5.2 The subject of evaluation: data organizations 81
5.2.1 Decompositions 81
5.2.2 Rotations 82
5.3 Using the best-match paradigm in simulation experiments . . 82
5.4 The test collection 83
5.4.1 The Cranfield II collection 83
5.4.2 Documents and requests 84
5.4.3 Indexing effort 85
5.4.4 Characteristics and use of the Cranfield collection in this dissertation 86
5.5 Summary of the evaluation approach 90

6 Scenarios of user simulations 91
6.1 Viability of non-interactive investigations and simulations in IR 91
6.1.1 Heine's approach 92
6.1.2 Magennis and van Rijsbergen's approach 93
6.1.3 Leouski and Allan's approach 94
6.1.4 White's approach 94
6.1.5 Lin's approach 95
6.1.6 The role of simulation in the present project 96
6.2 Deriving a uni-dimensional location model for the simulations 96
6.2.1 A location model 96
6.2.2 Relatedness to an axis represented by loading 96
6.3 Combined axes and combined loadings 97

6.3.1 The max model . . . 98
6.3.2 The sum model . . . 99
6.3.3 Setting the threshold . . . 99
6.3.4 Choosing a model . . . 100
6.3.5 Using a query in representing a user need . . . 102
6.4 Scenarios based on a location model . . . 103
6.4.1 The notion of relevance . . . 104
6.4.2 Simulation scenario 1 . . . 104
6.4.3 Simulation scenario 2 . . . 105
6.4.4 The accumulation of retrieved documents within simulation scenarios . . . 108
6.5 Strengths and weaknesses of the approach . . . 111

7 Measures of retrieval effectiveness 113
7.1 Introduction . . . 113
7.2 Centroid emphasis . . . 114
7.2.1 Suitability of an organization - the interpretation power of the OCP . . . 115
7.2.2 Initial ranking of data organizations by the centroid emphasis . . . 118
7.3 Ranked list measures . . . 118
7.3.1 Recall and precision . . . 120
7.3.2 Some related measures . . . 122
7.3.3 Discussion of the ranked list measures . . . 122
7.4 Measuring visualization support . . . 123
7.4.1 The need for measuring visualization support . . . 123
7.4.2 Measures modelling different kinds of users . . . 126
7.4.3 Separation-rewarded exposure . . . 126
7.4.4 Separation-rewarded precision . . . 130
7.4.5 Summary: significance of the two measures . . . 132
7.4.6 Testing the measures in boundary situations . . . 133
7.5 Additional measures and statistics . . . 136
7.5.1 Limitation of the hitherto discussed measures . . . 136
7.5.2 Numbers of queries attaining prescribed levels of recall 138
7.5.3 Diversity of access: axis usage statistics . . . 138
7.6 Significance of performance differences among data organizations . . . 139

8 Experiments and results 142
8.1 Introduction . . . 142
8.1.1 Presenting retrieval results . . . 142

8.1.2 Terminology . . . 143
8.2 Outline of the experiments and presentation of results . . . 143
8.2.1 Outline of the experiments . . . 143
8.2.2 Procedure . . . 144
8.2.3 Outline of the result presentation . . . 146
8.2.4 Using the versions of the collection . . . 147
8.3 Results for the automatically indexed version . . . 149
8.3.1 Centroid emphasis . . . 149
8.3.2 Results for simulation scenario 1: ranked list measures 154
8.3.3 Results for simulation scenario 1: visualization support measures . . . 160
8.3.4 Summary measures: discussion . . . 162
8.3.5 Results for simulation scenario 2: ranked list measures 162
8.3.6 Results for simulation scenario 2: visualization support measures . . . 164
8.3.7 Diversity of access: distribution of concept axes among queries . . . 168
8.4 Some results for the manually indexed version . . . 169
8.4.1 Selected precision recall results . . . 169
8.4.2 Summary measure statistics . . . 169
8.5 Different rotations . . . 174
8.6 Summary - axis interpretability . . . 174

III Summary and further research 177

9 Discussion 178
9.1 Measures and results . . . 179
9.1.1 Centroid emphasis . . . 179
9.1.2 Precision characteristics of data organizations . . . 180
9.1.3 Recall characteristics of Uexküll . . . 180
9.1.4 Summary measures . . . 181
9.1.5 Diversity of access . . . 182
9.1.6 Implications of the results . . . 183
9.2 The viability of the simulation approach . . . 183
9.2.1 Validity of the information provided by the simulations 184
9.2.2 Using both scenarios . . . 184
9.2.3 Viability of the simulations . . . 184
9.3 Further research . . . 185
9.3.1 Further work following the same research path . . . 185
9.3.2 User evaluation of the IR interfaces . . . 187

9.3.3 Other research directions to pursue 188
9.3.4 Open questions 189
9.4 Conclusion 190
9.4.1 Research question 1: interpretability of the LSI 190
9.4.2 Research question 2: axis interpretability, rotations and dimensionality 191
9.4.3 Research questions 3: intellectually indexed collections 193
9.4.4 Research questions 4: the potential of simulations . . . 194
9.4.5 Concluding remarks 194

References **196**

Appendices **205**

Appendix A Behavior test for the SRP and SRE mesures **206**

Appendix B Result details for summary measures **210**
B.1 Simulation scenario 1 210
B.1.1 Significance statistics for summary measures: sum model 210
B.1.2 Significance statistics for summary measures: max model 213
B.1.3 Behavior curves for summary measures: max model . 215

Appendix C Recall properties **217**

Appendix D Simulation scenario 2 **225**

Appendix E The relevance judgements of the Cranfield collection **227**
E.1 The README-file accompanying the relevance judgements for the cranfield collection 227
E.2 Recall bases for different dichotomization levels 228

Appendix F Some results for the manually indexed collection **229**

Part I

The Uexküll Approach

Chapter 1

Introduction and problem definition

This dissertation proposes and discusses a novel retrieval approach, called Uexküll. The aim of this approach is to enable users to visually seek information and retrieve it from databases, with minimum of typing. In this approach the retrieval system[1] is intended to take much of the responsibility of presenting the content of the database to the user, and the user part of the interaction is limited to:

- choosing among topics proposed by the system
- visually navigating in 2-dimensional and 3-dimensional[2] scenes, that are created in response to choosing those topics, where documents and terms are represented by graphical droplets.

At the same time, such a system is expected to provide better support for serendipity, supporting "making happy and unexpected discoveries" (Foskett, 1996, p. 26).

For the purpose of this dissertation, the Uexküll approach to information retrieval is defined as retrieving documents through navigation along coordinate axes of 2D and 3D projections of vector spaces brought about by data reduction methods.

A system that implements this approach would have two major parts:

[1] below the word "system" will be used instead of "retrieval system"

[2] Even though computers can facilitate for navigation in higher dimensionality, 2 and 3-dimensions where chosen because that corresponds closely to human perception.

- Terms and documents are ordered and represented in multi-dimensional spaces. This part will be referred to as *the data organization part*. It must support the construction of 2 and 3 dimensional scenes in the interface.

- Graphical user interface that displays scenes generated from these spaces, with which users can interact. This part entails many aspects of human-computer interaction and is not evaluated as such in this dissertation. A VRML-based[3] prototype is implemented for the sake of some rudimentary testing and illustration of the approach.

This dissertation therefore concentrates on the data organization part. The aim is to identify data organizations that provide for effective retrieval via scenes. For this purpose, an evaluation approach is necessary, that allows the comparison of different data organizations, assessing their performance using carefully chosen criteria.

This chapter serves as an introduction, presenting some concepts and ideas that are fundamental for understanding and analyzing the approach, starting with multidimensional vector spaces and the way they are used in IR (Section 1.1), continuing with a presentation of some premises and aspects crucial to the present work (section 1.2) and offering a summary of the idea and explanation of the name Uexküll (Section 1.3). The research questions and the structure of the dissertation are presented in Section 1.4.

1.1 Dimension reduction of term-document matrices

1.1.1 The full form vector space

The representation of document collections as multidimensional vector spaces is a well established practice in information retrieval (Salton & McGill, 1983; Deerwester et al., 1990). Technically this representation uses a set of either manually assigned or automatically extracted *index terms* that are used for constructing queries. An index term is

> a pre-selected term which can be used to refer to the content of a document (Baeza-Yates & Ribeiro-Neto, 1999, p. 444).

[3]VRML stands for Virtual Reality Modelling Language (Ames & Nadeau, 1996), and is a standard named ISO/IEC 14772-1:1997 (International Organization for Standardization, 1997) for 3D multimedia and shared virtual worlds on the internet.

For brevity, "term" will be used below instead of "index term".

Every document in the database has these terms assigned to it with a weight for each term. The simplest weighting scheme is the logical (or dichotomous or binary) one, by which a term either indexes the document or does not, coded by 1 or 0 in the *term-document matrix* (see Figure 1.2) respectively. Other weighting schemes exist where this matrix has real valued entries expressing different levels by which terms are assigned to the documents. Either way, a *term-document matrix* is normally sparse, most of its entries equalling zero, meaning that each term relates only to a few of the documents, and each document relates only to a few of the terms. Typically such vector spaces are defined by thousands of terms, while the illustration (Figure 1.1) shows only two.

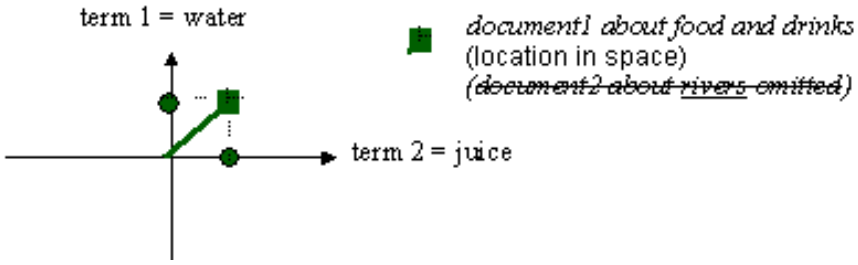


Figure 1.1: *A "full form" vector space representation of one of the documents from the matrix in Figure 1.2*

The best known way to construct a *knowledge space* out of this term-document matrix is the "full form"(Salton & McGill, 1983), where the index terms are regarded as the coordinate axes of the space (see a simple illustration in Figure 1.1). The documents are represented as vectors in that space, with start points at the origin, and end points determined by the weights on the axes (in the binary case either 1 or 0)

Traditionally, queries are represented by vectors of term weights in the same space, and retrieval is done by calculating the cosine of the angle between a query vector and all of the document vectors, returning the documents in descending order of the cosine score (C. J. van Rijsbergen, 1979).

1.1.2 Obtaining reduced multidimensional spaces

Dimension-reduced multidimensional spaces, and their interpretation in terms of latent variable models, are a common way of analyzing complex problems in many fields, notably within the social sciences (Bartholomew & Knott,

1999, p. 2). In IR, multivariate dimension reduction has also been successfully experimented with (Borko & Bernick, 1963; Deerwester et al., 1990; Hull, 1994; Ando, 2000; Blom & Ruhe, 2001).

If we reduce the dimensionality of the vector space (as described in Subsection 1.1.3), the terms no longer constitute the axes of the vector space, but become objects in a space of smaller dimensionality, together with the documents. As an example, a collection indexed by 3000 terms, thus having a dimensionality of 3000, may be reduced, with some loss of information, to dimensionality 150, where the 3000 terms no longer constitute the axes, but are objects in 150-dimensional space. Ideally, the loss of information due to such a reduction of dimensionality is such that persistent or dominant associations among documents, related to the intellectual content of those documents, are kept, while spurious associations due to typing errors, synonyms, homonyms and other such phenomena, are lost or suppressed.

Below, dimension reduction is illustrated by reducing a full form of 3 dimensions into 2 dimensions.

1.1.3 An example of dimension reduction

Consider a collection of three documents indexed by three index terms, as depicted in Figure 1.2. Applying a multi-variate transformation[4] to the matrix above, we obtain a knowledge space representation of two dimensions, as depicted in Figure 1.3.[5]

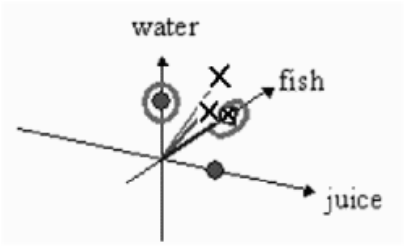
Regarding Figure 1.3, note that:

- All terms and documents have coordinates on each of the axes.
- Items that are relevant to a concept have their coordinates farther away from the origin in the positive direction. Document D_c (fish biology) is more relevant to *fish* than document D_b (rivers). Document D_b (rivers) is slightly more *water* relevant than the other two documents.
- A uni-dimensional (orthogonal) projection of this reduced space (Figure 1.4) will also have all of the items represented on it.

[4]Multivariate transformation are explained and discussed in Chapter 4.

[5] Note that spaces of up to three dimensions are directly visualizable, a fact which facilitates this illustration. Higher dimensionality is, principally and conceptually exactly the same, except for our inability to directly visualize it. However, these figures are only illustrations of dimension reduction, and do not refer to any visualization for end users.

Documents / Terms	food and drinks	rivers	fish biology
water	1	1	0
juice	1	0	0
fish	1	1	1



Document about food and drinks (Da)

Document about rivers (Db)

Document about fish biology (Dc)

Documents' spatial representation

Figure 1.2: *A collection of dimensionality 3 in "full form" knowledge space representation.*

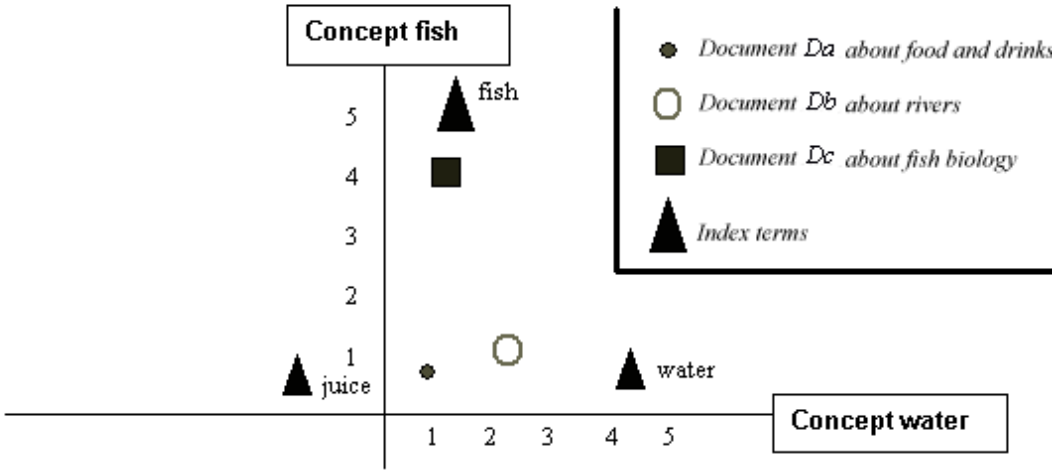


Figure 1.3: *A (fictitious) reduced knowledge space of the collection in Figure 1.2. The concepts are named by the term located furthest along their respective axes.*

- For both representations, objects in the vicinity of the origin are considered less important.

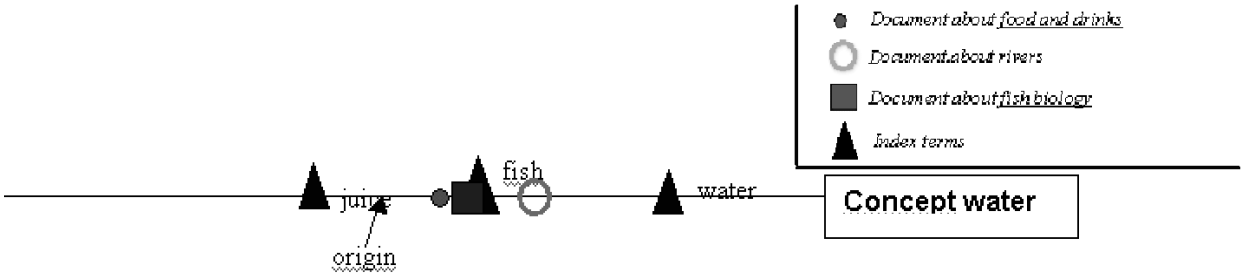


Figure 1.4: *A uni-dimensional projection of the 2-dimensional vector space*

One advantage of dimension-reduced representations in relation to the "full form" vector space is that in the former, objects are related to each other not only through first order associations, sharing common terms, but also through higher order (transitive) associations. If we draw pairs of documents, at random, from a real database, examining each pair separately, most pairs will not have any terms in common, and therefore apparently nothing in common. Consider such a pair of documents, consisting of, say D1 and D2. Two index terms, T1 and T2, indexing D1 and D2, respectively, may have a

third document, D3 which both of them index. D3 provides a higher order association that, in the reduced representation, would draw D1 and D2 closer together.

Furthermore: The coordinate axes of the reduced space comprise "latent variables", that the axes of the non-reduced space are explained through. The projection of the term on the axis representing the latent variable relates to the degree of association of that variable with the latent variable. Low positive coordinates mean lack of interdependence between them, and a negative coordinate value means that the term is negatively correlated with that variable.

Dimension reduction of representations within IR has a relatively long history, done for both classification and retrieval purposes. Here are some examples:

- In the early 60's Borko and Bernick (1963) used factor analysis in order to classify document collections, conceiving the coordinate axes as document classes. This line of research seems to have been abandoned at the end of the 60's or beginning of the 70's.
- Latent semantic indexing (LSI) (Deerwester et al., 1990), is probably the best known effort in this realm of research. Singular value decomposition (SVD) is used, in order to obtain a vector space of reduced dimensionality in an unsupervised manner, i.e. without a-priori regarding the inner structure of the data. Retrieval from the reduced form is done the same way as retrieval from the "full form" vector space: A query (pseudo document) is compared to all documents using a similarity measure (mostly the cosine of the angle between them), and the documents with the greatest similarity are retrieved.
- Ando (2000), Hoffman (1999) and others have, in different ways, theoretically founded and further developed the computations of LSI spaces.

1.1.4 Using the reduced dimensionality for visualization

The idea of the present project is based on three properties of the above mentioned dimension reduction:

- Each document or term has an extension (loading) on each axis.
- A document or term that is highly related to an axis has a high positive coordinate value on the axis.

- Following this, it is conceivable that a term that has the highest coordinate value along an axis may be a candidate to become the name of the axis[6].

If we choose any three axes (out of k axes in a k-dimensional space), that in some way represent a user interest, and only regard the objects (terms and documents) that have high loadings on these axes, we may get documents and terms that are related to the axes chosen and hence to the user request that initiated the choice. A selection like this will hereafter be referred to as a 3D orthogonal projection (Anton, 1987) or, in short, a projection. A projection for which

- coordinate axes have names,
- objects (documents and terms) with low loadings are discarded and
- only highly loaded objects are retained

will be referred to as an Uexküll group.[7]

Such an Uexküll group, being only three dimensional, can be visualized using a 3D computerized visualization system, where objects can be represented by figures (cubes or droplets) within a coordinate system.

In the next section the motivation and criteria for intended use of the spaces of reduced dimensionality will be introduced.

1.2 The focus of the dissertation

This section looks at a number of important aspects that constitute the motivation (easier access to retrieval systems), the premises (indexing and types of text collections) and the task of the present dissertation (evaluation of suitability of data organizations).

1.2.1 Access to systems

Users who search databases are not in pursuit of documents. They are in pursuit of information about topics (this will be elaborated in Chapter 2). More often than not users have problems of communicating their pursuit to other

[6]Axis naming schemes is beyond the scope of this dissertation. The mentioned naming scheme is a simplistic approach to be used in the development of the ideas.

[7]See section 1.3 for an explanation regarding the name Uexküll .

people (Taylor, 1968; Ingwersen, 1992; Nordlie, 2000), let alone to a technical retrieval system. For these cases in general a retrieval system that requires a precise communication of needs through keywords is not satisfactory.

The purpose of this research effort is to design and evaluate a partial remedy to this problem by letting the system display its content to the user by *constituting concepts* (important terms) that capture the content of a certain database, and allowing the user to represent his request in terms of these concepts by selecting a few of them. The system thereby takes more of the responsibility presenting the content to the user, freeing the user from having to express his information need through terms that may, or may not be a part of the system's vocabulary.

Foskett (1996) divides users (whom he calls readers) into two main groups (p. 15):

> Normally, readers will be satisfied with a few items, so long as these contain the sort of information wanted... But there will be situations when readers will require high recall - as much information as possible - even though this means they will have to look through a lot of items that will turn out to be of little or no value.

This dissertation will look at both types of users, and try to assess the potential of the Uexküll approach to satisfy either of them.

1.2.2 Indexing

Indexing is an important way of generating entry points for retrieval of documents. Historically, indexing of collections has been done intellectually, where experts *assigned* index terms to documents, often using controlled vocabularies. With the advent of full-text collections, automatic indexing, by *extraction* of index terms, has become more and more usual.[8]

Whereas intellectual indexing often is binary (an assignable index term either indexes a document or does not), automatic indexing is often graded, meaning that a term may be more or less important as an entry point to a document. The weighting is often positively related to the frequency of occurrence of a term in a document, and inversely related to the frequency of occurrence of the term in the entire collection. Other factors than the

[8]Experiments on automatic indexing by assignment have also been performed (Lancaster, 1998), but will not be dealt with here.

frequency (such as whether the term appears in the title, abstract body or captions), may also be used to assign weights to extracted terms.

For the purpose of the present effort, both types of indexing are of interest. Automatic indexing has traditionally been used for preparing databases for latent semantic indexing, both because the real numbered weights in the term-document matrix lend themselves more easily to the creation of continuous subspaces, and because the co-occurrence pattern is relatively rich. The problem with this type of indexing is that, being accomplished by extraction and subjected to stemming and normalization, the terms (single words, often only roots or word-stems) are not very suitable as axis designators (concepts). Even though the latter problem can be remedied by using of meaningful terms assigned intellectually or automatically after the fact, the absence of the intellectual effort when indexing the documents in the first place is harder to alleviate.

Intellectual indexing (indexing by assignment) most often uses binary weighting, which does not lend itself equally easily to the creation of continuous subspaces, because of the sparseness of the matrixes and the fact that the entries are not dichotomized from normally distributed data. On the other hand, the terms used in such indexing are *assigned terms*, often contextually meaningful phrases, not always occurring in the documents themselves, that would constitute much better candidates to the naming of axes. With the latter hypothesis as a motivation, this project also looks at dimension reduction of binary indexed databases, using the same methods as for reduction of the real numbered data.

1.2.3 Contents of databases

An assumption concerning the Uexküll approach is that it will not be suitable for very specific databases in topics within (e.g.) physics and chemistry. It is assumed that such databases are searched by specialists and have a limited vocabulary known by both information providers and searchers, and allows for efficient text based search. For these types of databases, Uexküll would not represent an added value. The approach will also not be suitable for very large databases, like large segments of the internet or the entire internet, simply because the technology and methods are currently not capable of organizing such enormous data sets.

This leaves us with databases that are sufficiently small, and at the same time sufficiently diverse, for which an approach like Uexküll represent an added value. We assume that the approach would be particularly helpful for

searching document corpora that have distinct and different groups of users. Public, governmental information is an example of the type of material for which this user-specific method is particularly suitable.

Lawyers, civil servants, politicians, specialists of various disciplines, and laymen in search of information are types of users that will look for different things in the same corpus of documents. Different conceptual entry points to the same material as represented by the Uexküll groups may assist these user groups and facilitate effective retrieval for each of these groups for its own purposes.

1.2.4 Interpretable data organization

The data organization used in an Uexküll based system should facilitate the identification of topically relevant Uexküll groups, within which relevant documents are highly loaded. In addition the relevant documents should be effectively distinguished from documents that are less relevant, so that users can select them for further consideration. Obtaining a suitable data organization for the purposes described so far, is tightly associated with the problem of interpretability of the axes.

Interpretability of a data organization is here defined as the extent to which that organization associates documents and index terms with named axes. In multivariate analysis interpretation is often improved by rotating the objects in relation to the coordinate axes, in a way that makes as many as possible of the objects highly associated with few axes, and consequently disassociated with the remaining axes. Rotations are discussed more closely in Chapter 4.

1.3 A summary of the idea and the name "Uexküll".

The idea behind a system implementing Uexküll may be summarized through the following steps:

- After reducing the dimensionality of a collection, we typically obtain several hundreds of dimensions and each dimension has a name assigned to it.
- A user that wishes to search this collection is presented with a list of all those names, and is prompted to select 3 of them (e.g. "History", "Norway" and "Middle ages") that are most likely to represent his "need".

- The system creates a 3 dimensional scene, in which only documents and terms with high loadings pertaining to these axes are downloaded and placed. (filtering effect)

- The user can navigate in the scene, expecting to find material more relevant to "Norway" the farther along the "Norway" axis he moves.

In principle any three of the dimensions can be used as named axes. This is true provided that our dimensions are interpretable. [9]

Inherent in the approach is that any such projection might correspond to a user-specific, individual knowledge subspace. This approach has its roots in biology, going back to the Estonian biologist Jakob von Uexküll (1864-1944), with reference to whom the approach is named. In his Umwelt (environment) theory, Uexküll makes the distinction between Umwelt, the subjective construct of any biological perceiver, and its objective environment.

In this approach, the entire dimension reduced space is analogous to Uexküll's objective environment, and the 3-dimensional projections represent subjective constructs (Uexküll's Umwelts).

1.4 The structure of the dissertation and the research questions

1.4.1 Focus and research questions

Summarizing the previous sections, a central problem underpinning the potential of the Uexküll approach is the possibility of the interpretation of axes within multidimensional organizations of index terms and documents. Chapter 4 discusses models and methods for this task, and reviews related work within information retrieval. Though important for the usability of an Uexküll-based system, the attention given to the question of intellectual indexing, is confined to some experiments that test the potential. This is because the real effect of intellectually assigned concept names needs to be tested with user participation.

The current evaluation approach is developed (Chapters 5-7), and is based on *user simulations* that try to capture important aspects of the intended

[9]Even though LSI normally assumes 200-600 dimensions for the purpose of cosine (or inner product) based similarity retrieval, we do not believe that all of these dimensions have enough interpretability to be effective in the Uexküll setting.

usage of the system. These simulations run queries. The queries are based on requests, for which sets of documents judged relevant have previously been assigned. Retrieval performance is decomposed into a number of aspects:

- how well are documents judged relevant ranked in relation to other documents in the scene?
- how well are documents judged relevant visually located in the 3-dimensional scenes under the various data organizations?
- how well are documents judged relevant distinguished from other documents under the various data organizations?

The latter two points will be termed "visual support", and attempts will be done to characterize and measure it. The evaluation draws largely on the traditional laboratory based evaluation approach within IR (Salton & McGill, 1983), but transcends it as it also tries to measure aspects of usability beyond the traditional approach. We term these aspects "visual support", and they are summarized mainly by second and the third forms above, meaning that relevant documents are prominently located in the scene (high loadings on axes), and well separated from non-relevant documents.

The development of this evaluation approach is also an important focus of the dissertation.

The questions this dissertation will try to answer, based on the user simulations, are:

1. Given the definition of the Uexküll approach in page 14, and measured in terms of overall ranking and visual support, how well do traditional methods of data reduction, particularly LSI, support visualization based on the Uexküll approach? How interpretable are the axes of subspaces derived by LSI?

2. Measured in terms of overall ranking and visual separation, how well will rotation of LSI-derived data organizations support retrieval?

   (a) How do traditional methods of rotating axes influence the axis interpretability, and thereby the support, within the Uexküll approach, for retrieving relevant documents?

   (b) What is the effect of the dimensionality of the decompositions on the interpretability of the rotated spaces?

   (c) To what extent will the methods presented here support the diversity of access-entries to presented material, where an access-entry

is defined as a named axis? Would the data organizations enable different users to access materials processed and presented in an Uexküll based system through different access points?

(d) What is the potential of the approach in helping the two types of readers defined by Foskett (1996) and quoted in Subsection 1.2.1

3. What is the effect of the type of indexing. How well would the Uexküll approach support retrieval of documents in intellectually indexed databases?

4. What are the advantages and the disadvantages of using simulations of the type performed in this project? To what extent do the simulations capture the properties of the approach, and what is the potential of using simulations in similar approaches?

1.4.2 Summary - the structure of the dissertation

The rest of the dissertation is structured as follows: In Chapter 2 the need for an approach like Uexküll is motivated by some central problems within information seeking and retrieval. Before the technical treatment of the approach, we shall discuss some relevant developments and problems in information retrieval, and the potential of the Uexküll approach to aid in resolving some of these problems. Chapter 3 describes a prototype of an Uexküll -based system, developed in order to look at the practical feasibility of the idea, and to facilitate a more practical discussion on how to implement the idea.[10] Chapter 4 discusses methods for obtaining multidimensional representations of text collections used in information retrieval, and also describes rotations that increase the interpretability of the axes. Chapter 5 present the overall approach to the evaluation of the data organizations. Chapter 6 starts with a discussion of simulation in information retrieval, before it develops and describes two simulation scenarios for the evaluation of the approach. In Chapter 7 we discuss measures of effectiveness, both traditional and novel, to be used in the evaluation. Novel measures are developed in order to evaluate aspects of the approach (visual support) not supported by the traditional ones. Chapter 8 presents and analyzes the results of the simulation experiments, and Chapter 9 discusses the limitations of the evaluation approach and the future of the approach in light of the results.

[10]The prototype has been very helpful in identifying how documents or terms appear to the user, thereby supporting a more realistic design of the simulations than would be feasible without such a prototype.

Chapter 2

Approaches to information retrieval and visualization

"Information retrieval (IR) deals with the representation, storage, organization of, and access to information items." (Baeza-Yates & Ribeiro-Neto, 1999). The history of automatic IR spans 5 to 6 decades through which tremendous development has taken place. Naturally, the focus in early years was retrieval of documents from relatively small, limited databases and resources. With the advent of the internet, and the enormous growth of availability of information items, focus has changed towards treatment of large, heterogenous resources. This will not be the focus of this dissertation.

In this chapter we are reviewing some developments in IR that are relevant as motivation to this work and, where applicable, motivate the more specific topic of the dissertation within those developments.

"Knowledge" and "information" are two important terms that are in considerable use in this chapter. "Knowledge", or change in knowledge state (Belkin, Oddy & Brooks, 1982), is often perceived as the abstract entity that is the utmost goal of engaging in information retrieval, whereas information is conceived as a constituent unit of knowledge structures.

Losee (1997) regards information as *values of characteristics* of input and output to certain processes, that do not require the presence of human beings. Knowledge is, in the same article, perceived as the highest entity in the hierarchy also containing perception, observation and belief.

The problem of defining "information", as well as the general use of the word is also addressed by Wersig (1971), who creates a taxonomy of several different uses of this word. Though not directly addressed here, the concept of information lies at the heart of information retrieval research and its

development.

> ... the focus for a concept of information has moved from the areas of generated messages (contents of texts), over the message itself (not its meaning), to its meaning (e.g. to recipient or sender), and ending in the form of reduction of uncertainty in the mind of the recipient. (Ingwersen, 1992, p. 27).

The conception of information has evolved during the decades of IR development, and is different in different schools/approaches to information retrieval.

2.1 System-oriented information retrieval

The approaches discussed in this section (the so called "hard-core" IR) implicitly associate information with text objects, like sets of index terms, terms occurring in documents and the documents themselves. This implicit notion was equivalent with the belief that the potential of better retrieval was in the improvement of the retrieval algorithms internal to the *retrieval systems*. "Retrieval system" is here understood as the computerized "black box" containing the database, the user-interface and the search program. This had consequences not only for research into IR, but also, of course for how information systems were evaluated. Evaluation was based on pre-composed requests, directed at a test collection. For these requests, expert-judged sets of relevant documents, taken from the collection, were devised. Systems were assumed the better, the more closely the document sets retrieved as response to those requests matched the expert-judged document sets.

With this conception of information, it was tacitly assumed that users had stable, well articulated information needs, and these information needs were properly expressed by *stated requests*[1]. Moreover it was assumed that "Indexers and users share similar vocabularies and that science and needs develop slowly or ought to be of a rather static nature" (Ingwersen, 1992, p.51).

This approach to IR was seldom challenged before the late sixties (see for example Taylor (1968)), and is still quite prominent in the design of retrieval tools.

[1]One example why this latter assumption may be problematic is the "label effect": Empirical studies have shown that users tend to express their needs through a label which consists of very few concepts. The description of the need is thus compromised, and is potentially not conveyed with high fidelity to either a machine or an intermediary. This is particularly true for "Muddled topical needs or ill-defined information problems, i.e. the user wants to explore some new concepts or concept relations *outside known* subject matter"(Ingwersen, 1992, p. 117).

2.1.1 The Cranfield studies

Perhaps the most important milestones in the history of the system oriented approach were the Cranfield experiments, Cranfield 1 and 2, conducted at the end of the fifties and the beginning of the sixties, respectively. They mark a tradition, and the name Cranfield is used to designate this tradition.

Cranfield 1 started out as a comparison of a number of indexing languages.

> The essential features of the project were [...] that it was a comparative one, focusing on indexing languages, and seeking to identify and control other system factors bearing on the study variable. (Spärck Jones, 1981, p. 258-9)

One of the goals of Cranfield 1 was to test whether new indexing languages gave better results than established ones, and how indexing depth and other variations on indexing affected retrieval performance. However, at the same time other issues, like composition of requests and evaluation of results, emerged as important. Among these issues were the measures of precision and recall[2] and their mutual relations.

Large collections pose problems for the evaluation of recall, because in principle all documents in the collection need to be assessed for relevance to every request. That is why the Cranfield 2 project used a collection consisting of as few as 1400 documents and 279 requests (see Subsection 5.4.1 for more details). Briefly stated, the collection was assembled on the basis of queries formulated by authors of selected papers, to which this paper was a direct answer. Other documents, taken to be answering the question, were later on added, comprising the question's "recall base" together with the first document.

This strategy was criticized for not representing real life when the relationships between user requests and documents were concerned (Vickery, 1966). This critique was dismissed by Sparck Jones (1981) as inadequate, as Cranfield was a comparative laboratory experiment. A methodological critique raised by Rees (1967) was that

> The problem of a criterion measure remains in that Cleverdon's measure reflects the overall or ultimate performance of the system or subsystem tested. The sources of variation affecting performance are not adequately pinpointed, and small indication is provided as to how to optimize performance (Rees, 1967, p. 68)

[2]Precision and recall measure the ratio of numbers of relevant document retrieved to the total retrieved and the ratio of relevant documents retrieved to the number of relevant document in the collection, respectively.

2.1.2 TREC

The other formative effort of system-oriented IR evaluation is TREC (Text Retrieval Conference). TREC is annually arranged by NIST (National Institute for Standards and Technology in the USA) *"...to encourage research in information retrieval from large text collections" (TREC, 2001).*

In its first years (starting at 1992) TREC concentrated on two primary tasks, the ad-hoc task "searching for documents on some topic on some particular occasion" (Spärck Jones, 2000, p. 43), which was, for several reasons, the most important task, and the routing task, where new documents are matched with existing profiles, in a manner similar to *selective dissemination of information* (SDI). As opposed to the ad-hoc task where queries are matched with documents already in the collection, the routing task is is aimed at the evaluation of distribution of new documents acquired by the collection.

For the large sets of documents involved in TREC, the problem of recall bases mentioned above in connection with the Cranfield 2 experiments is of course much more significant, and cannot be solved on a document by document basis as with the 1400 documents of the Cranfield 2 collection. Here a pool approach has been used: The various systems participating contribute to a pool of relevant documents as returned by their systems for the queries. This varying input to the pool ensures that the pool is "as good as possible" in both coverage and lack of bias. In order to accommodate future research, each request has a "narrative", which specifies what makes a document relevant to this request.

In order to accommodate developments in both interest and practice within text retrieval not covered by the main tasks, TREC has developed tracks (instituted first at TREC-4), dealing with various issues. These ranged from *cross-language retrieval* and *very large corpus* to *high precision* and *spoken document retrieval*. An important track, relevant to this dissertation is the *interactive track*, where users' actions in developing their needs into requests and into queries while interacting with a system, are carefully studied.

2.1.3 XML and INEX

XML is a metalanguage that supports the development of new markup languages. These languages are used for encoding content, but also for encoding semantic interaction between nodes over the internet and providing services (called web services). Each XML document is a set of nodes, ordered in a hierarchy. The nodes are instances of *elements* possibly qualified by *attributes*.

The elements/attributes for each document written in an XML-based language are chosen from a set of elements/attributes defined for that language in a *DTD* (document type definition) or an *XML-Schema*.

INEX, (INitiative for the Evaluation of XML Retrieval) is a project, coordinated from the University of Duisburg in Germany.

> As part of a large-scale effort to improve the efficiency of research in information retrieval and digital libraries, this project initiates an international, coordinated effort to promote evaluation procedures for content-based XML retrieval (Fuhr, 2006).

INEX is largely inspired by the TREC conferences, and has adopted similar organizational means of collaborative testing of different systems and methods for XML-retrieval. Like TREC, also INEX is organized in a constantly growing set of tasks, meant to capture various aspects of XML-retrieval.

2.1.4 Summary

It is interesting to note that one of the more important things that TREC (particularly with its ad-hoc task) has achieved, is consolidating and establishing the strategies from the first Cranfield experiments, importantly with large data collections. As Karen Spark Jones puts it: "Though some retrieval strategies found valuable in TREC were first suggested 40 years ago, one of the major TREC contributions has been to establish them not only through many individual tests but with very large files and with full text data"(Spärck Jones, 2000, p. 83).

2.2 Knowledge space and the vector space model

In the mid-seventies Meincke and Atherton (1976) introduced the concept of Knowledge Space. This was done partly in response to alleged flaws of prevalent classification and indexing schemes, which were not fit to accommodate "growth of knowledge and our changing constructs of it". One important flaw of such schemes was the dependence on more and more information, e.g. more and more descriptors in authority lists. Another flaw was the lack of "a way of interacting with the structures that allows for imperfect awareness"[3]. As a partial remedy, they propose a "multidimensional space model

[3]imperfect awareness means that it is impossible to gain perfect knowledge about the world, because the world dynamically interacts with our perception, and our perception

for knowledge structuring". This model was quite ambitious, being able to represent both concepts, persons' state of knowledge and information items.

A decade ahead of the publication of that paper, the vector space model and the "SMART" system had made their first steps. The vector space model is explained in Subsection 1.1.1. The connection between Knowledge Space as introduced by Meincke and Atherton and the vector space model is discussed in McGill (1975) and McGill (1976).

Along with the *probabilistic model* (Maron & Kuhns, 1960), the vector space model has played a major role in the development of best-match IR. The vector space model is a family of interpretations of the term-document matrix (see Subsection 1.1.1), often denoted by $\mathcal{D}$. Its standard interpretation regards the terms as orthogonal axes, represented by the rows (or columns), and the documents (columns or rows, respectively) as vectors having coordinate values on these axes[4]. The elements of the matrix are *weights*, calculated by a *weighting function*. The weighting function expresses the extent to which the corresponding term describes the corresponding document, based on its frequency of occurrence in the document, its frequency of occurrence in the entire collection, and possible other parameters. In this interpretation a query is a *vector of weights* of its constituent terms, which is the same representation as that of a document. This means that it is possible to establish measures of similarity between a document and a query. This interpretation assumes that $\mathcal{D}$ holds the only information available.

Alternative interpretations have been proposed, some of them also assume other sources of information in addition to $\mathcal{D}$ (Raghavan & Wong, 1986).

Some models of reduced dimensionality have also been proposed. The idea has been that both terms and documents are (generally non-orthogonal) vectors in a vector space of reduced dimensionality. The best known research in this realm has been latent semantic indexing (Deerwester et al., 1990), (see also 1.1.3). Through the years following 1990, LSI has been thoroughly reviewed, and both modifications and alternative models have been proposed (Hull, 1994; Ando, 2000; Blom, 1999; Blom & Ruhe, 2001; He, Cai, Liu & Ma, 2004). Some of these alternatives are discussed in Chapter 4.

2.3 The three revolutions

Robertson and Hancock-Beaulieu (1992) summarize three revolutions that IR had undergone in the 70s and the 80s:

needs to be interpreted.

[4]The choice of rows or columns as terms representatives varies in the literature.

- The *relevance* revolution - the growing acceptance that relevance should be judged in relation to *user needs* rather than in relation to *stated requests*.
- The *interactive* revolution. The possibility, brought by computer and network technologies to interact with a database within a very short period of time (as opposed to the need to wait for an overseas mail reply after submitting a query to an institute holding a database) made it practically possible to make systems interactive.
- The *cognitive* revolution. Information seeking was now described in cognitive terms. The authors use the ASK model of Belkin (Belkin et al., 1982) as an example, where "information need is seen as a reflection of an anomalous state of knowledge on the part of the requester". Among other consequences, this revolution also imposed a change in what was regarded as "the system". What was earlier regarded as the "retrieval system" (see above) is now reduced to "mechanisms", and the system now includes the user's activity while interacting with the mechanism. It was a new type of system that needed to be evaluated.

It was quite early observed that the notion of "user request = user need" was problematic. Belkin et. al. (1982) point at the unlikelihood of the assumption that information needs and documents are equivalent: "A document, after all, is supposed to be a statement of what its author knows about a topic...[whereas] the expression of an information need is a general statement of what the user does not know".

One attempt to resolve this problem within best-match[5] retrieval was to use *relevance feedback* from the user to modify the query and bring the query closer to what the user would find useful in terms of document interrelations in the context. This was, as Belkin claimed, not a real solution to the problem, as the ultimate query still did not take account of the "user need", rather of the "user request".

2.3.1 The cognitive approach and ASK

The cognitive approach became more important in IR research in the late 70s, the 80s and the 90s (Ingwersen, 1999). The most important contribution of this approach may be summarized in the following points:

[5]In best-match retrieval we try to quantify the relatedness of a document to a query based on the context, as opposed to exact match where we try to find documents containing exactly the logical combination of words used in the query. In this text, unless otherwise stated, we speak of best-match retrieval.

- the conception of information: information does not consist of static objects like words in documents, but a process that causes change of knowledge in humans;
- the concept of information need, which is situation-dependent, and can change during a single interaction with an information system;
- only the user can judge the relevance of retrieved documents.

An important contribution to the cognitive approach was Belkin's hypothesis of ASK (Anomalous State of Knowledge). According to the ASK hypothesis "an information need arises from a recognized anomaly in the user's state of knowledge concerning some topic or situation and that, in general, the user is unable to specify precisely what is needed to resolve that anomaly" (Belkin et al., 1982).

This inability by the user to specify his or her needs places limitations on the interaction with retrieval mechanisms through query specification. ASK was inspired by, and partly based on, experiments done in the 70s with a new type of a retrieval system, the THOMAS system (Oddy, 1977), that was not based on query formulation, but on dialogue. This system was not further developed, and the idea did not gain popularity.

The THOMAS system was interactive, interactivity meaning that each retrieval instance depended on the user's evaluation of the previous retrieval instance. The special thing about this system was that at the same time as the user was constructing a model of the system and the database, the system was constructing a model (or "image") of the user, expressed by "a network of associations between documents, authors and subject terms: any pair of like or differing types, may be linked" (Oddy, 1977, p. 5).

The most important feature of THOMAS was that it did not force the user to formulate queries. The only thing the user did to initiate the dialogue was a hint to the system, with a couple of words, about his domain of interest. The system then constructed its first image by a set of associations, and based on this image suggested documents and terms that might be interesting. Roughly described, the user responded by stating which of the terms or documents were of interest. THOMAS used this feedback to modify the image, and based on this updated image, offered new documents (or terms) to be selected or rejected by the user. In this way, and probably as its most important contribution, THOMAS was meant to cater for the changes in the user's conception of his information need during the interaction.

Several reasons may have contributed to the lack of follow-up of the THOMAS approach.

- At the time the THOMAS system was tried out, library catalogues were still searched by intermediaries (e.g., librarians), who in many instances knew the material well and would probably not benefit greatly from such an approach.

- The THOMAS system was tried on MEDLARS, a database system used by practitioners of medicine and related professions. It might be speculated that such experts, with good knowledge of the material again, did not benefit greatly from such an improvement in interactivity.

- Building user models is a difficult, complicated task, and the research rather advanced in other directions.

The contribution of the THOMAS system lies in its inspiration to the further research of the cognitive paradigm.

2.4 Human intermediary-based retrieval as a model of automatic retrieval

It is nowadays widely accepted that the ultimate aim of document retrieval interaction is not the retrieved documents themselves or the "information" in them, but their effect on the knowledge gap that the user wishes to fill by engaging in retrieval. This means that the success or failure of retrieval should be judged not by the retrieved documents, but by the extent to which the user has come closer to filling this knowledge gap. This knowledge gap is not measurable. In many instances the information seeker is not even able to articulate this knowledge gap (Taylor, 1968; Ingwersen, 1992; Nordlie, 2000).

Articulation of the knowledge gap is one obstacle. Once an articulation has been accomplished, the potential knowledge has to be expressed in terms that are understood by the retrieval system. In this context we have already mentioned Belkin's problematization of matching queries and documents. For example in Boolean systems, based on metadata retrieval, articulation of the knowledge gap requires the construction of complicated Boolean expressions and a knowledge of the vocabulary of the system, e.g., which terms have been used to index the collection.

In the author's opinion, one problem is that systems often do not present themselves to the user in an understandable way. Moreover, text-based information systems place much of the responsibility for the interaction on the

users' shoulders. Shneiderman (1998, p. 511-513) reviews some of the obstacles users, particularly first time users of information exploration systems meet.

Libraries have successfully used human intermediaries to help users overcome this knowledge gap (Nordlie, 2000, p. 132). Several efforts have been undertaken to build or model systems that act in a manner similar to human intermediaries, or at least imitate some features of such intermediaries. This we will refer to as *intermediary interaction inspired automatic retrieval* (IIIAR).

IIIAR, as we see it, can be subdivided into system-specific, automatic expert intermediaries like the CONIT (Marcus & Reintjes, 1981) and CANSEARCH (Pollitt, 1984) systems, and expert system models like MONSTRAT (Belkin, Seeger & Wersig, 1983) and MEDIATOR (Ingwersen, 1992). Among other approaches to IIIAR we find *user revealment* (Nordlie, 2000), that sees the intermediary as an agent in the interaction with the user.

Large parts of this branch of research seem to have been abandoned during the 90s. Nevertheless, it inspired the holistic cognitive IR-research at that time. some of the issues and solutions have even "gained new relevance in the new millennium in the research and development towards the Semantic Web" (Ingwersen & Järvelin, 2005, p. 162). The research efforts within IIIAR have been an important inspiration and motivation for the present study, although our approach to the solution is different. In the following sections we will discuss several important efforts within this paradigm, and discuss our own approach in light of these.

2.4.1 Expert systems for query formulation

CONIT and CANSEARCH are examples of expert systems that were devised to, in different ways, facilitate users' access to retrieval systems. CONIT was built to allow users access three different retrieval systems in a transparent way. CONIT translated user requests into the systems' search languages and translated the system responses into a single language. CANSEARCH was a menu driven system that helped users create queries regarding clinical cancer therapy against the MEDLINE database (a large database of medical articles), using the MESH (Medical Subject Headings) vocabulary. Those queries could then be directed at MEDLINE.

CONIT and CANSEARCH, though old and outdated, are here referred to as examples of systems simulating well defined functions of human intermediaries. CONIT, operating on the query - communication level, represented

the human intermediary's knowledge of different retrieval systems and the ability to communicate with the user, whereas CANSEARCH, operating on the communication - query - request model level, represented the intermediary's knowledge of a *limited field* and its vocabulary, while helping the user to create his requests through an interaction.

An example of a newer tool of a similar kind is Metalib by Exlibris (Sadeh, 2004). This is a so-called Metasearching system, mediating between users and quality controlled, heterogenous resources. Such a system consists of a resource metadata repository that holds metadata about resources, and a retrieval software component. The retrieval software is responsible for

- ranking or selecting the appropriate resources, based on its knowledge about the user or the user's affiliation,
- converting and normalizing search results from the heterogenous resources, that can be presented to the user in a unified manner.

2.4.2 User and system modelling: the MONSTRAT and MEDIATOR models

Based on the analysis of 31 dialogues between users and human intermediaries in a library environment, MONSTRAT (Belkin et al., 1983) was developed as a set of independent, but cooperating *experts*, each of which was responsible for one of the underlying functions extracted from the analysis. Such functions were "Problem State", "Problem Mode", "User Model", "Problem Description", "Dialogue Mode", "Retrieval Strategy", "Response Generator", "Explanation", "Input Analyst" and "Output Generator". All these experts communicated through a virtual "Blackboard", so that they did not have to be aware of each other directly. Unlike the CONIT and CANSEARCH approaches, which are technical in nature, MONSTRAT's perhaps most important feature was user modelling - trying to build a model of the *user* (not only the request), including the user's state of knowledge.

MEDIATOR (Ingwersen, 1992), a further development of the MONSTRAT model, was aiming at augmenting the latter with theoretical and empirical advances in IR made since MONSTRAT's conception. At the same time, aspects of MONSTRAT taken into MEDIATOR were modified, to make them more general. The purpose of MEDIATOR has been "adaptively to bridge a user with an information requirement" on the one hand, and "[bridge the user

with a] local and/or remote IR systems" on the other hand. "MEDIATOR demonstrates a corresponding symmetry of functionalities".

An important contribution of MEDIATOR is the *system model* related functions that are only implicit in the MONSTRAT model. The failure of MONSTRAT to explicate this family of requirements is attributed to the user-oriented paradigm within which the MONSTRAT model was conceived. MEDIATOR claimed to move in a more cognitive direction (Ellis, 1996b, p. 76). As a whole, MEDIATOR should be regarded as modelling the process of: "how users think when interacting with a computer which dynamically builds up its models of that user as well as of computer IR systems." (Ingwersen, 1992, p. 205).

"Ingwersen conceives of the model being useful in three ways - for analysis and design of intermediary mechanisms, for assessment and for education and training" (Ellis, 1996b, p. 76). Mediator is thereby not a retrieval system. It is a scheme, whose main purpose is to give designers of retrieval systems a tool for looking at what it is they need to think about when designing a system for possibly different purposes.

Designing intermediary based mechanisms into systems following the MEDIATOR model presupposes domain specific empirical research in order to uncover relevant problems/factors to gear the design towards.

2.4.3 User revealment

In his Ph.D. dissertation Nordlie (2000) pursues the observation of the successfulness of *intermediary aided interaction* as opposed to the failure of *automatic retrieval system interaction* in leading users to materials that may fill their knowledge gap.

Unlike the intermediary models like CONIT and CANSEARCH described above, which are technical in nature, Nordlie's approach is more user oriented, in the sense that it does not see the intermediary exclusively as an owner of topical or technical expertise, but also as an agent in extracting the user's knowledge of his own needs. This is in line with the MONSTRAT idea, but a step further, employing social and discourse theories.

In user intermediary interaction, Nordlie identifies a process to which he refers as "user revealment", or how, with the help of an intermediary, the user overcomes the difficulty in articulating his or her own information need. He offers some recommendations for the design of automatic retrieval systems that might improve these systems' capabilities in this direction. The recommendations are listed below:

- Provide context-sensitive help based on an interpretation of the user's current problem.
- Assist users in the disambiguation and specification of their problem statement, for instance by identifying, highlighting and suggesting the different aspects of a topic or the different topics which are represented in a set of retrieved references.
- Present the "system model" to the user by allowing him to see and browse through the structure or organization of the data (...) and by exemplifying and demonstrating search features and search options.
- Explain search results so the user may see how the system treats his query and why each item is retrieved.
- Develop a set of disambiguating questions, based for instance on thesauri or classification schemes, to help the user focus his query and pose it at the relevant level of specificity.

One problem that Nordlie points to is that, as a consequence of the absence of a user model, online systems have high recall as one of their design objectives, whereas intermediary interaction has high precision as a goal.

This last point is interesting, and provides a specific sub-goal that system design may use as a measure of performance improvement towards the performance of the intermediary:

- Is the design of system X more precision promoting than a normal online system? What do we lose in order to attain this specific objective?
- What new system features (or flaws) can be directly attributed to achieving this goal?

2.4.4 Uexküll in the light of intermediary based modelling

One important difference, as viewed by the this author, between the present work and the intermediary based retrieval modelling efforts is that the Uexküll approach does not entail building a model of the user, at least not explicitly. The emphasis in the Uexküll approach is exploration of the *contents* of databases, attempting to facilitate it for different user groups.

Comparing to approaches represented by efforts like CONIT and CANSEARCH, supporting users in a way that an intermediary would, we claim that the

Uexküll approach represents a modification, as it does not depart from queries formulated by the users. With Uexküll we actually wish to free the user from the need to formulate queries in the first place.

This point is also relevant when characterizing the Uexküll approach in light of the conclusions/recommendations brought forward by Nordlie (2000). Nordlie envisages the use of external aids, like thesauri and classification system to explicate a query. Though this is also possible to implement in the frame of the Uexküll approach, (and could be a fruitful direction of further research), the approach itself will ideally do without such aids.

In light of endeavors like MONSTRAT and MEDIATOR, the Uexküll approach does not represent an attempt at catering for different user groups as a consequence of empirical study of human behavior. We believe in giving the users an opportunity to use their own resources by facilitating *the contents of databases* for navigation. The lack of specific knowledge of the user groups is compensated for by the attempt at facilitating of many entry points, which could accommodate for needs of different user groups.

Indeed, an interesting point we find in Ingwersen's description of the philosophy behind the MEDIATOR model is the strong belief in users' own intelligence and capabilities. Ingwersen believes that

> Solid background models may serve the user better [than asking the user many specific question - MP] when applied to support him making him use his own associative capabilities (Ingwersen, 1992, p.206)

In our approach both the facilitating of the data (organizing it for the purpose of presentation and interaction), and the intended interaction, are exploratory. In line with Ingwersen, we believe in using the users' own resources. Through the Uexküll approach we wish to attain the usage of these resources by presentation of contents to users, making users independent of prior knowledge of the system's attributes.

2.5 Visualization in information retrieval

When using the term visualization, a distinction between *scientific* and *non-scientific* visualization is common. Williams, Sochats and Morse (1995) classify scientific visualization as a method of computing that transforms results of analysis into representations that are easy to scrutinize, and non-scientific visualization as a method of extracting highly abstract features from complex

high-dimensional data, in a way that appeals to human perceptual capacity. As an additional example of this distinction, they provided a televised weather report as an example of scientific visualization, and the "explorer"-style[6] directory structure of the files in an operating system as an example of non-scientific visualization.

Williams et al. (1995) motivate visualizations indicating that graphical means appeal to the human processing capacity differently than linguistic means do(p. 167):

> Humans can discriminate among a large number of colors, textures, distances, sizes, and changes in position and patterns. Humans can also discriminate among sound attributes such as pitch and loudness. Computers can be utilized to present data to humans in such a manner that the flexibility of representation and speed of processing can be coupled with human information-processing capabilities to permit new dimensions of presenting and analyzing data.

Börner, Chen and Boyack (2003) present a number of aspects that well designed visualizations should support (p. 209):

- An ability to comprehend huge amounts of data on a large scale as well as a small scale.
- A reduction in visual search time (e.g., by exploiting low level visual perception).
- A better understanding of a complex data set (e.g., by exploiting data landscape metaphors).
- Illumination of relations otherwise not noticed (e.g., by exploiting perception of emergent properties).
- A data set to be seen from several perspectives simultaneously.
- Facilitation of hypothesis formulation.
- Effective sources of communication.

In the following overview, we wish to discuss some efforts in IR-visualization which, in our opinion, contextualize the Uexküll visualization.

[6] referring to the file organizing tool integrated in the Microsoft Windows system.

2.5.1 Types of visualization within IR

Visualization within IR may be differentiated along various axes. Such are "dominant metaphors, underlying metrics, functions on behalf of users, interaction capabilities, relation to retrieval engines, suggested evaluation points"(Citing Rorvig (1996)), and whether the visualization is created on-line [....] or offline" (H. D. White & McCain, 1997).

All the visualizations discussed in this section are on-line, offering some sort of interactivity. Within this domain H. D. White and McCain (1997) distinguish between "(1) the user's interest model of the literature...[and]...(2) the system's graphical summary of some corpus of documents" as types of underlying models for the visualization. They offer an extensive review of the topic, based, among other things, also on this distinction. The treatment in Subsection 2.5.2 is more focused on efforts having bearing on the present work.

2.5.2 Research efforts within IR visualization

In this subsection we attempt to present some major research efforts that represent important aspects. We categorize them by the way in which the visualization is created from the data representation.

2.5.2.1 The Korfhage paradigm: reducing dimensionality while displaying

This paradigm has been mainly developed by Robert Korfhage's group at Pittsburgh University. The idea underlying all the related projects was the usage of different metrics that may be extracted from the vector space representations of document collections (angles, distances or combinations of these). These metrics are transformed into visual navigational representations. In the center of these latter representations lie *reference points* or *points of interest*, that represent index terms or documents known to be relevant to the current user request. Other objects, whose vector space locations render them as potentially relevant, are placed on the display in relation to these *reference points*. This relation is a transformation of the spatial relation among the objects in the source representation. The dimensionality of the visualization space is determined by the number of reference points. Notable implementations within this paradigm (in no particular order) are GUIDO (Nuchprayoon & Korfhage, 1994), TOFIR (Zhang, 2001) and VIBE (Olsen, Korfhage, Sochats, Spring & Williams, 1993). An interesting further

development is the VR-VIBE (Benford et al., 1995), that added a dimension to the original VIBE system, and implemented the 3D display in VRML.[7].

2.5.2.2 Reducing dimensionality prior to displaying

This aspect is characterized by the use of complex dimensionality reduction algorithms prior to displaying the space the user may interact with.

In the BEAD project, Chalmers and Chitson (1992) have used iterative numerical procedures simulating physical processes, to create a low-dimensional representation used for visualization. Documents in high dimensional space are represented by simulated particles that are brought into a low dimensional spatial configuration that, as much as possible, approaches the *a priori* known distances between document pairs in a vector space (a sort of multidimensional scaling). The numerical procedure groups the particles into "voxels" (volume cells), and places the latter in a low dimensional hierarchical configuration. The low dimensional configuration should retain as much as possible of the multidimensional document structure.

Another example reducing dimensionality before displaying is SPIRE^{TM} (Wise, 1999), whose aim is the visualization of whole document collections, extensively using metaphors, such as galaxies and terrain formations. In the SPIRE project the visualizations were created by projecting high-dimensional document representations onto a plane, via clusters. As the dimensions of the space that were used prior to visualization represent terms, users may, at retrieval time, change the relative weights of terms in order to get a more user specific display.

Daránvi, Zawiasa and Hajnal (1996) have used principal component analysis (PCA) in order to identify and visualize the three most important aspects drawn from a corpus. These aspects are represented by the 3 principal components associated with the highest eigenvalues of the term-document matrix, and may be named using terms that are highly loaded on these axes.

2.5.2.3 Dimensionality reduction of search results

Characteristic for the class of projects discusssed below is that a query is submitted to the system in a "traditional" manner, using keywords, with or without an explicit search language. The result set is then analyzed, and a graphical display visualizes an ordering of the retrieved documents

[7]Card (1996) classifies this type of tools as sensemaking tools, that are used to "help users understand information by associating and combining it."

using relative locations, color-code, tabulation or a combination of these. The ordering may express topics, kinds of publication, relevance ranking, similarity among items, and so on. The LVis system (Börner, Dillon & Dolinsky, 2000) clusters retrieved images based on a latent semantic analysis (LSA) algorithm followed by a hierarchical clustering technique that is based on item similarity. Items are then visualized in a highly metaphorical display, where both the entry into the configuration of the items and the items' configuration itself are apparently designed to create a virtual environment also in a mental interpretation of the word. Two interfaces, both a 2D and a 3D, have been experimented with.

> Both interfaces give users access to three levels of detail: They provide an overview about document clusters and their relations; they show how images belonging in the same cluster relate to one another; and they give more detailed information about an image such as its description or its full size version (Börner et al., 2000, p. 79).

Allan, Leouski and Swan (1997) use an approach to dimension reduction similar to BEAD (see 2.5.2.2 above). Unlike BEAD (where the approach is used for reducing the dimensionality of an entire collection prior to visualization), they apply this reduction to *retrieved document sets*. Additionally they modify their algorithms to enhance the separation of relevant documents from the non-relevant. They report experiments with dimension reduction down to both 3-d, 2-d and 1-d, finding no significant improvement of the 3-d layout as opposed to the 2-d. In a later publication, Leouski and Allan (1998) report a simulation based evaluation of the system, which is discussed in Subsection 6.1.3.

2.5.3 Visualization in information retrieval - summary

In this section we have presented a number of approaches to visualization of document data, focusing on the data reduction element. Related to these approaches, the Uexküll visualization actually reduces data twice: once before visualization, where a number of dimensions (lower than the original dimensionality of the data) is constructed, and during visualization when a subspace of this lower dimensionality representation is selected by the user. This is inspired by some of the approaches discussed above. The main contribution is that the document configuration the user obtains as a result of an initial interaction may be pursued using a direction metaphor. This is explained in more detail in Chapter 3.

2.6 Summary

In this chapter we have been reviewing trends in information retrieval that we consider relevant to the present project. The focus was on problems, acknowledged by the community through decades, to which Uexküll proposes possible remedies through visualization. Prominent here is the dynamic nature of users' information need and users' difficulties in expressing these needs in words, particularly in a way that matches the vocabularies of the retrieval systems/document databases. We have also tried to identify trends in IR visualization, a much younger branch than IR, also here choosing projects that are related to, and inspire, the current project.

Chapter 3

The Uexküll visualization

This chapter discusses the visualization as seen from the user's side, as well as the considerations that this visualization imposes on the parameters that govern the data organization[1].

While working on the project, it became important to get more intimately acquainted with what interaction with an implementation of such an approach would be like. To this end, an interactive VRML-based prototype (Preminger & Darányi, 2000) was implemented, that made it possible to directly experiment with the interface. Screen captures from this prototype are used in the present chapter in order to make the explanations more concrete. It is important to stress, though, that this is only one way of implementing the visualization. The evaluation described in the rest of the dissertation is more general, and seeks to characterize the approach as a whole, rather than a specific implementation.

Along with the description of the way of working with the interface, some parameters influencing the quality of the visualization are discussed, as well as considerations of functionality.

3.1 Aims of the Uexküll system design

With the Uexküll approach we wish to combine several features:

- Visualization techniques.

[1]From Chapter 1:The criteria can be summarized as following: The data organization should facilitate the identification of documents that are relevant to a user need (request). In addition, the relevant documents should be effectively distinguished from documents that are less relevant, so that the user can pick those up and select them for further consideration.

- Placing more responsibility on the retrieval system of presenting the database content to the user than is usual in traditional systems.
- Creating an understandable metaphor for the user.

There are probably many possible approaches to the combination of these elements. This dissertation concentrates on using the vector space model for visualizing associations between topics and documents in a way that allows users to navigate. Below, some design principles will be discussed for a system design implementing Uexküll.

The scope of the work, which contains user simulation, but not direct full-scale user tests, constrains the possibilities of evaluating the features of the visualization. Fulfilling the more long-range goal of the study will, of course, require user tests.

The treatment in this chapter is based the principles introduced in Section 1.1, breaking down these principles into the system design considerations.

3.2 Main principles of design

The Uexküll visualization is based on two successive dimension reductions: the one prior to interaction, provided by the multivariate analysis, and the other, at the time of interaction, taking a 3D orthogonal projection based on the named axes chosen by the user. During the second reduction only documents related to this projection (as perceived by the data organization) are downloaded. The limit of this relatedness is determined by a threshold, below which objects are not regarded as related to the projection.

The downloaded objects (documents and/or terms) are components in a scene, in which navigation/retrieval can be pursued by "moving" in the *direction* of axes. This section discusses these points, and in the end breaks them further down into parameters that control the quality of the visualization.

3.2.1 Determining the threshold

Since all the documents have loadings on all the axes, an important part of constructing an Uexküll group is the determination of some threshold loading, below which documents are not considered to be related to an axis. Documents for which the loadings do not exceed this threshold for any of the coordinate axes of an Uexküll group, are not considered members of

that group, and are subsequently not downloaded for visualization when this Uexküll group is invoked by a user.

The threshold may be determined by various strategies. The simplest strategy is the choice of a constant threshold. A constant threshold may be set for a certain database. A constant threshold may also be set, by a user, for a certain session. This may not be an intuitively appealing strategy, as Uexküll groups are different, and thresholds that leave too many documents in one Uexküll group may leave too few documents in another, but may be a fair strategy to pursue within an evaluation, when we wish to keep a variable constant, thereby reducing the number of uncontrolled variables.

A modification of this strategy is the choice of a threshold that is proportional (or otherwise related) to the size of the scene, for example a constant fraction of the highest loading of any object in the scene.

A different strategy is leaving the threshold as a real-time controlled parameter that the user interactively changes if too many or too few items are shown. Such a strategy would also demand some kind of initial threshold, both for performance control (e.g. avoiding the downloading of too many items when creating the initial scene), and for cognitive load control (e.g. not to show too many items in the initial scene unless absolutely necessary).

3.2.2 The direction metaphor

A metaphor may be defined as "One thing conceived as representing another"[2].

We define the the direction metaphor as "navigating, by a user, in a 3D scene on a computer screen, following coordinate axes, conceived as moving in a 3D space in a physical world, indoor or outdoor". The motion may also be interpreted as following a physical direction with ones eyes. Even though the constituting directions of the physical world are orthogonal, they need not be conceived as such by the user. The user follows labeled directions on the screen the same way he would (physically or visually) follow directions in streets, corridors and the like.

The direction metaphor is actually a utilization of a substantial property of factor analysis, namely that the relatedness of cases or variables to an axis is expressed by the extension of that object along the axis. Following this principle, the farther out an object is along the axis, the more likely it is to

[2]metaphor. (n.d.). Dictionary.com Unabridged (v 1.1). Retrieved June 26, 2008, from the Dictionary.com website: http://dictionary.reference.com/browse/metaphor

be related to that axis: Moving along the axis named "History" in a scene of some hypothetic collection, in its positive direction, one meets objects that are (contextually) more and more likely to be "History" related.

The navigation described here obviously presupposes a good naming scheme of the axes, and that the name assigned to an axis is such that objects conceived more relevant to the name of the axis will also be more related to the axis itself, and what it topically represents.

Regardless of the underlying model that places documents and terms in the multidimensional spaces, and along the coordinate axes (see Chapter 4), the user is assumed to perceive relations between document locations in the scene in an approximately linear manner, meaning that documents twice as remote from the origin are perceived twice as related to an axis. This is of course dependent on the user's ability to project documents angularly remote from the axes onto the axes (an object that has a high loading on more than one visible axis will not lie angularly close to any of these axes, and its exact projections may be difficult to perceive).

3.2.3 Parameters determining the functionality of the interface

The quality of the visualization is dependent on a number of closely related parameters:

Topicality By topicality we mean that documents that are relevant to a projection (Uexküll group) are indeed placed within this Uexküll group when this group is invoked.

Pertinent documents load high on Uexküll group axes When a user chooses projections (concept names) that represent topics he is interested in, documents more pertinent to these topics are expected to have their loadings (coordinates) far out on one, two or all of the chosen axes.

A monotone relation between pertinence and loading In addition to having pertinent documents high up the axes, we would ideally also wish the placements of documents to express their pertinence to an axis on the Uexküll group on a (preferably linear) continuum. A fulfillment of this would also render the first and second conditions fulfilled.

Good naming of axes The naming of the axes is an important issue that needs to be attended to provide good interaction.

3.3 Interaction with the prototype

A typical session starts with the user being presented 3 drop-down menus, all containing the same list of axes (see Figure 3.1). The user is expected to select a concept from each list, and the concepts chosen are assumed to represent the user's interest. Having chosen 3 concepts, along with some additional parameters, the user submits his choice, and a scene is constructed based on coordinate data downloaded from the coordinate database. We will refer to this process as "downloading a scene". The scene (see Figure 3.2) contains a set of axes, and a HUD (head-up display), which may be seen as a compass, providing information about the orientation of the scene and can be used to change this orientation. Terms pertaining to the scene are represented by green 3D droplets, and documents are represented by red droplets.

When the scene has been built, the user may start navigating within it in pursuit of relevant documents. The available navigation acts are:

- clicking on terms that seem relevant, so that related documents appear in their vicinity
- moving along the scene in any direction
- changing the user's orientation in and viewpoint against the scene
- clicking on documents to retrieve these.

In the following, the interaction is described in a more detailed manner, while introducing the screen layout components.

3.3.1 The upper menu

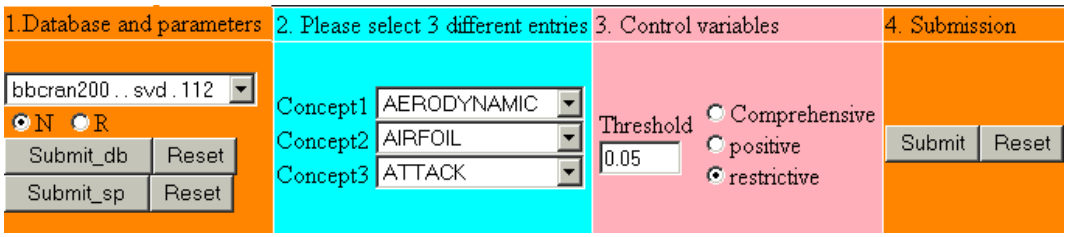


Figure 3.1: *The menu (upper frame) of the initial screen*

The upper menu has four sections that are listed and briefly explained below:

Choice of database Choose the database and the data organization. This is probably not a typical choice a user would be interested in, assuming that the database searched is a part of the system start parameters. (The radio button options "N" and "R" are experimental and give a choice between rotated and non-rotated version of the coordinates where applicable).

Choice of Uexküll group concepts (axes) The user is presented with three identical lists of concepts to choose from. This choice represents the user's interest in the material, and determines which axes will comprise the Uexküll group. It may be argued that for a typical organization a list of 100 to 300 concepts may be too long to search in. To this end some experiments have been made to provide an additional level of navigation, starting with a view of *concept groups*, the choices among these reduces the number of concepts displayed in the menus. This is future work (see Chapter 9), and is not further addressed in this chapter.

Control variables and their meanings The user can manually determine the level (loading) below which an object is not taken to be pertinent to an axis. An initial threshold can of course be determined by the system, but the user should have a way of modifying this threshold should he be dissatisfied with the initial scene results (see Subsection 3.2.1). The radio button choice near the threshold entry has the following meanings:

Comprehensive Objects that are loaded higher than the threshold for at least one of the 3 axes are displayed

Restrictive Objects that are higher than the threshold for all the 3 axes are displayed

Positive Object that have a positive loading on all of the axes are displayed.

3.3.2 The scene

Upon selecting of 1, 2^3 or (for most of the cases) 3 concepts and submitting the selection, the system returns an initial scene (distinguished from "the initial *screen*" above), that contains:

[3]Selecting 1 or 2 concepts is equivalent to displaying 3 or 2 identical concepts respectively.

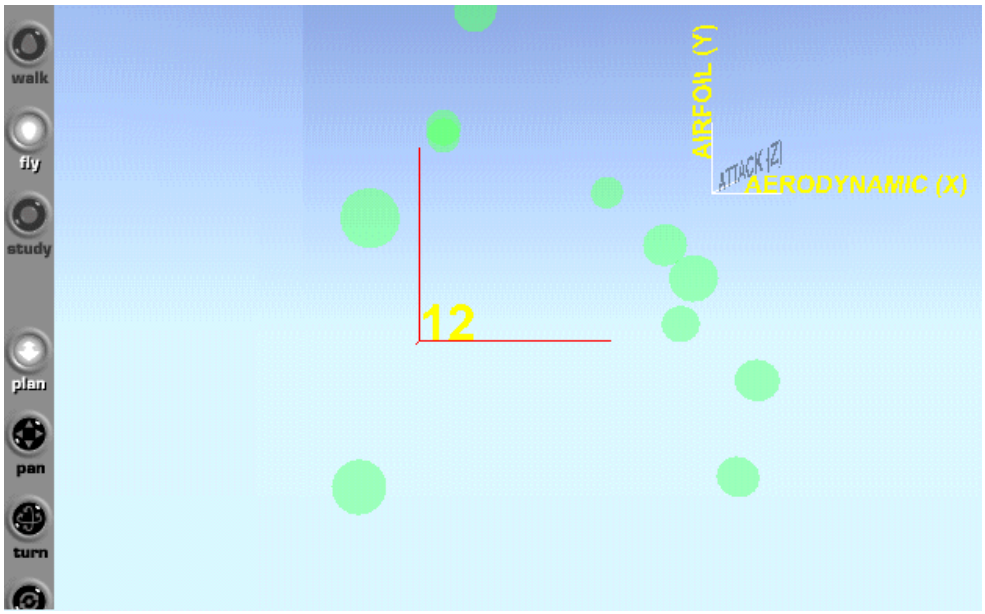


Figure 3.2: *The initial scene the user meets after selecting the first Uexküll group*

Figure 3.3: *The term buttons provide additional access to manipulating the scene. A click on a button reorients the scene so that the corresponding term is put in a central position, and highlighted.*

- head-up display (HUD) of the coordinate axes, on the upper right corner. The HUD is a layout of the axes in the coordinate system, constantly positioned in the upper right corner of the scene, and used as a *compass* for orientation in the scene. The axes on the HUD are labeled with the concept name, and its orientation follows the orientation of the scene. The hidden axis (perpendicular to the screen) is colored black.

- Green droplets that represent the terms pertinent to the current Uexküll group. No text is initially displayed about them on the scene itself, to reduce cognitive load. Text is displayed on demand, when the mouse cursor touches a droplet.

- In addition, a window containing buttons with the names of all the scene-pertinent terms is presented. Pressing a button orients the scene so that the term droplet is visible. This feature provides access through terms. (Figure 3.3).

The initial scene is rotated and tilted, so that one of the three dimensions is perpendicular to the screen and thereby hidden (this will be referred to as an orthogonal scene[4]. By default the 3^{rd} concept chosen will be hidden, as it is assumed most users would choose concepts in descending order of interest. The user can, at any time, determine which of the three dimensions is hidden (perpendicular), by clicking on that dimension's name on the HUD, using the main mouse button.

The user has the option of navigating the scene, moving his viewpoint across the scene in any direction, using the arrow keys or the mouse cursor. The user can at any time tilt or rotate the scene in any direction. A click on one of the axes (as described in the previous paragraph) returns the user to an appropriate orthogonal scene[5].

3.3.3 Navigating within a scene

For navigating within a scene in this particular implementation, we use the idea of a "term bubble". A term bubble is a location of a term with the documents that surround it within a confined radius. Navigating within a scene, the following actions may be taken

[4]VRML does not support orthogonal views (the scene is always viewed in perspective). Therefore the hidden axis will partly appear

[5]This feature may, in the future, be implemented so that the current viewpoint determine the exact offset of the scene for the orthogonal view

- Pursuing objects along the axes (we recall that objects positioned farther out along the positive direction of an axis are assumed more relevant to this axis). This is assumed often to happen in 2D, as it may be difficult to perceive directions that are oblique in relation to the screen. This is a feature inherent in the Uexküll approach, supporting 3D navigation as an feature for users that are used to/have the ability to orient themselves in 3D, but also supporting 2D navigation.
  - Clicking on one of the compass (HUD) axes, the user orients the screen so that the clicked-on axis becomes perpendicular to the screen, and the remaining axes comprise a 2D scene.
  - Using the mouse to rotate the scene, in a way that depends on the current VRML-supported navigation mode (plan, pan, turn or roll, see ParallelGraphics (2004)).
- Closing in on term bubbles: Moving the mouse on a term bubble highlights the bubble and causes the term name to appear in its vicinity (see Figure 3.4). Clicking a term hides it and exposes the documents within the bubble (see Figure 3.5)

3.3.4 Retrieval and presentation of documents

Documents not yet scrutinized/retrieved are represented by red droplets. As the user browses documents (mouse cursor over droplet causes metadata to be highlighted) some of these will presumably be interesting for further scrutiny. When the user clicks on the document droplet, a "retrieved" status is stored for the document, and the droplet changes its color to yellow. Any document already retrieved will appear yellow also in subsequent interactions (within the same session) involving this document.

3.3.5 Navigating among scenes

When a scene is downloaded the user can either navigate within it (see Subsection 3.3.3), or navigate to a new scene.

Navigating from the initial scene is done by modifying the Uexküll group. The currently chosen concepts are still selected on the dropdown menus. The user may substitute one or more of the current concepts with new concepts, that would represent a modification of the path the user goes to pursue his interest - based on the satisfaction/dissatisfaction with the progress in the search.

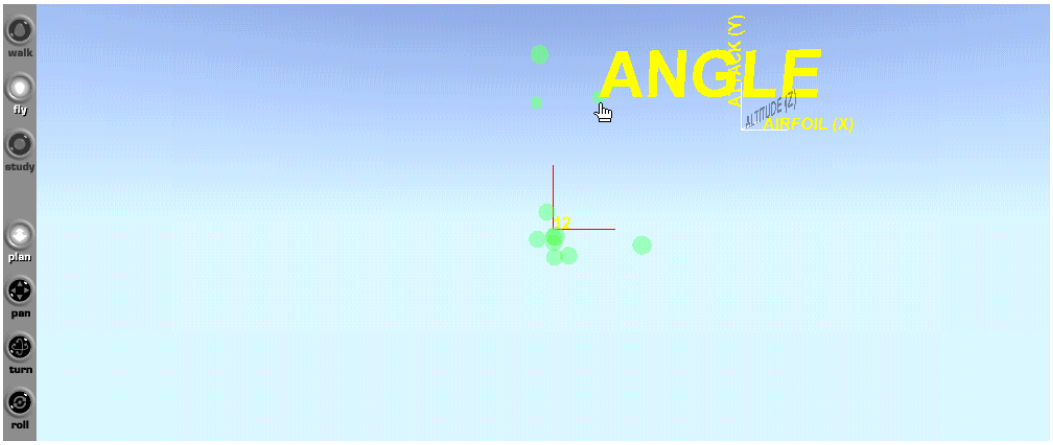


Figure 3.4: *Highlighting a term: Moving the mouse cursor over the term droplet causes the term name to appear*

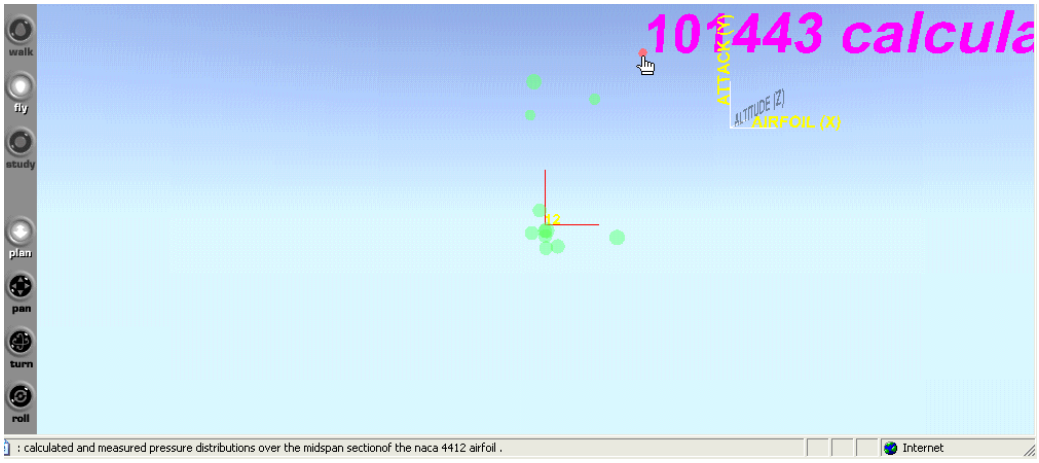


Figure 3.5: *After clicking on a term bubble: The term has disappeared, and a single document (the only one that was pertinent to this term) has appeared up on the right corner. Moving the mouse cursor over the document droplet causes metadata (in this case number and title) to appear in the scene and in the status line*

Changing only one of the current concept axes would represent an adjustment, resulting in a finer modification of the scene. Changing two concepts axes would represent a coarser adjustment, whereas changing all three concepts, would represent a total change of representation of the user interest. Technically speaking, in each of these cases, a new scene is downloaded from the coordinate database[6].

This process can be reiterated until the user is satisfied with the result or decides to terminate the session

3.4 Concluding remark

In this chapter we have described a possible implementation of the Uexküll approach, and through this implementation characterized some properties of the approach as a whole. We wish to reiterate that this is only one possible implementation of the approach. In this prototype the initial scene the user is presented when choosing an Uexküll group displays the terms, and uses terms as gateways to approach documents. This may be regarded as an example of an implementation-specific feature. It is entirely conceivable that a different implementation does not present terms in the scenes. If our database had many more terms than documents, or the terms were stemmed or otherwise truncated words, presenting them would entail much noise. Such indexing warrants an implementation where pertinent documents are presented immediately after a scene is chosen.

[6]For the finer changes one could transfer the current viewpoint, so that the user is "placed" in a recognizable location related to the previous scene.

Chapter 4

Multidimensional methods for dimension reduction within IR

4.1 Introduction

This dissertation addresses the data organization of document collections for visualizing in continuous vector spaces. The organizations experimented with are primarily transformed subspaces of term-document matrices, based on the singular value decomposition (SVD). We are following an exploratory procedure, that tries to impose simple structure on the subspaces generated by SVD, and test the interpretability of the axes following such a procedure, as opposed to the interpretability without such a procedure. We see our effort as an attempt at adding value to the latent semantic indexing model (Deerwester et al., 1990), and it is therefore natural to discuss LSI itself, as well as other efforts that have been addressing and modifying LSI.

This chapter discusses multidimensional methods for organizing document and index term data in vector space by reducing the dimension of the concept space (Kantor, 1994, p. 67). We start by discussing the factor model as a main idea behind the data organization. We proceed with discussing the SVD, particularly the latent semantic indexing (or latent semantic analysis) model, and then discuss the usage of these models/methods within the Uexküll approach. Later sections briefly address the possibilities for treating test collections indexed by human experts within the Uexküll approach[1].

[1]Such collections are almost always indexed dichotomously (an index term either indexes a document or does not), contrary to the occurrence frequency based association between a term and a document, often translated to a continuous association, normal to automatically indexed databases.

4.1.1 A brief historical overview

Factor analysis has been used by formative efforts within data organization/classification. Among these, we note two:

- Borko and Bernick (1963) were attempting to identify categories of documents based on term distribution within the term-document matrix.
- Ossorio (1966) was trying to use technical expressions appearing in documents, and expert statements about the relevance of these expression to *subject domains*, in order to characterize these subject domains. He was factoring statements of experts characterizing the relevance of technical expressions to selected fields of research.

One issue bringing these two efforts close to the present effort, is the interpretation step, using rotations (discussed in Subsection 4.2.2).

Within IR, latent semantic indexing (LSI) and the SVD have been experimented with much more extensively than linear factor analysis, beginning towards the end of the 80's. Here the applications were mostly of pure ad-hoc information retrieval, the type of experiments matching documents already resident in the collection to queries. Efforts within the routing problem, matching new documents to pre-established user profiles (Hull, 1994), have also been noted.

The aim of LSI was the improvement of the matching of queries and documents in the vector space model by suppressing noise, and relaxing the requirement for exact string matching between query and document. This was done by reducing the dimensionality of the data in a way that enhanced similarity through underlying topical relationship, and at the same time filtered away noise and other ephemera due to typing errors, homonyms, synonym and the like.

LSI (also known as latent semantic analysis, LSA) has, in recent years, extended its scope to modeling of processes in the human brain (Landauer, Laham & Foltz, 1997; Landauer & Dumais, 1997). At the same time, LSI has been debated and followed up by the IR-community, resulting in development, improvements and alternatives to the original transformations. This will be briefly discussed in Subsection 4.3.3.

In some sense, the present effort represents a return to the factor analytic path of research practiced in the 1960s, and abandoned in the 1970s. This is mainly due to our putting *interpretability* into focus when reducing dimensionality of document-term relationships (see Subsection 4.2.3). When we, in the present effort, revisit this apparently abandoned path, it is because

our visualization metaphor needs interpretability, and because we would also like to accommodate for the use of intellectually indexed material. The size of the collection is, in this effort, a lesser concern.

4.1.2 Latent variables and the factor model

Many of the concepts that we are using in this chapter, and the dissertation as a whole, are taken from exploratory factor analysis (below referred to as factor analysis, or FA). FA describes a set of *manifest variables* through a set of fewer *factors*, also known as *latent variables*. The most important characteristic of the latent variables is that they cannot be directly observed in real life. They are mathematical constructs. The latent variables comprise a space, and the manifest variables are projected into this space, so that they are expressed in terms of those latent variables. Each and every manifest variable has a coordinate value, also known as **loading**, on each and every latent axis (see, for example Figure 1.3).

Factor analysis as a family of methods (discussed in Section 4.2) is most widely associated with the normal linear factor model. Theoretically, thus, factor analysis assumes linearity in the relationships among individual manifest variables and the latent variables.

Factor analysis is largely associated with Psychometrics. A typical application of factor analysis in psychometrics is the analysis of score data, where *manifest variables* (e.g. school examination scores in various topics) are collected or polled from *objects* (e.g. school children), and on this basis *factors* or *latent variables* (e.g. overall intelligence) are extracted[2].

4.1.3 Result interpretation

An important aspect of factor analysis is the *interpretation* of the results. A variable like overall intelligence is not directly observable, and its existence, although conventionally accepted, may be debated. It is typically a product of interpreting factor analysis results. In order to facilitate interpretation, the results, an array of points in multidimensional space representing the variables, are often *rotated* by some criterion or criteria (discussed in Subsection 4.2.3), so that the researcher, using e.g. scatter plots[3] can conjecture or deduce the meaning of the results: how manifest and latent variables relate to each other and what the latter may signify.

[2]The example is taken from Spearman's experiments from 1904, cited in Mardia, Kent and Bibby (1979)

[3]A scatter plot is a plot locating points in a Euclidean plane

Interpretability and interpretation is a natural bottleneck of the present research. Whereas interpreting results of factor analysis in the social science is done intellectually, with a specific question, problem or model in mind, interpretability in the present context is meant to facilitate retrieval of documents by different users by automatically ordering documents along axes.

In their important article from (1990), Deerwester et al. make two comments that have bearing on the present project: firstly, they explicitly say (p. 395):

> We make no attempt to interpret the underlying factors, nor to "rotate" them to some meaningful orientation.[4].

In a similar context, Parnas (1994) asserts that humans are not assumed to think in orthogonal concepts. This is a plausible observation, that also means that interpretation for its own sake (assigning names to concepts that would entail orthogonal meaning) is not expected to be fruitful. This is not what the present project is trying to achieve. Secondly, Deerwester et al. (1990) say (p. 396):

> Unlike many typical uses of factor analysis, we are not necessarily interested in reducing the representation to a very low dimensionality, say two or three factors, because we are not interested in being able to visualize the space or understand it.

The present author conceives that Deerwester et al. (1990) confine "visualization" to the reduction of the entire space (including all terms and documents) to two or three dimensions.

Further down (p. 396), they state:

> We believe that the representation of a conceptual space for any large document collection will require more than a handful of underlying independent "concepts", and thus that the number of orthogonal factors that will be needed is likely to be fairly large. Moreover, we believe that the model of a Euclidean space[5] is at best a useful approximation. In reality, conceptual relations among terms and documents certainly involve[s] more complex structures, including, for example, local hierarchies and nonlinear interactions between meanings. More complex relations can often be made to approximately fit a dimensional representation by increasing the number of dimensions. In effect, different

[4]"Rotation" is here interpreted as orthogonal rotation.

[5]Given this context, the present author interprets "Euclidean" as 2 or 3 dimensional.

> parts of the space will be used for different parts of the language or object domain. Thus we have reason to avoid both very low and extremely high numbers of dimensions.

It is not difficult to agree with these statements. This project is trying to investigate the limits of implied interpretation of the axes *in higher dimensionality*, for the sake of navigation. Indeed, the structure of both natural language and human perception makes it impossible to assign a "named meaning" to any of the mathematically derived axes, which will fully support orthogonality of meanings. Hypothetically, a result that will manage to assign to each concept only a single axis, will overfit, and will not be very useful.

4.2 Factor analysis as a model and a family of methods

4.2.1 The linear factor model

The factor analytic model explains a set of variables through a *common factor* component and a *unique factor* component.

$$\mathbf{x} = \Lambda\mathbf{f} + \mathbf{u} + \mathbf{e} \tag{4.1}$$

where Λ is the matrix of factor loadings (the loadings are the coordinate values of the manifest variables in the reduced common space), $\mathbf{f}$ is a vector of factors, and $\mathbf{u}$ is a vector of unique contributions of the variables. $\mathbf{x}$ is a vector of zero mean random variables, and $\mathbf{e}$ represents a vector of zero mean independent noise terms.

Assuming $Cov(\mathbf{f}) = I$, where I is the identity matrix, it is also possible to write the factor model in terms of the Correlation matrix:

$$\Sigma = \Lambda\Lambda^{'} + \Psi, \tag{4.2}$$

where Σ is the correlation matrix of the manifest variables, $\mathbf{x}$, $\Lambda\Lambda^{'}$ is the common factor correlation matrix, and Ψ is the covariance matrix of the unique contributions of the variables. Ψ is diagonal: there are no correlations between different variables' *unique contributions*, and the main diagonal of

Ψ features normalized variances of every variable, while the off diagonal elements equal zero.

The factor model is appealing for organization of documents and terms in vector spaces, because it is intuitively appealing to regard a document corpus as a common context for the variables (index terms), but where variables also have some unique contribution each.

4.2.2 Initial solution and rotation

The factors that are reached by estimating the parameters of the model are orthogonal, each one of them independently describes each one of the manifest variables.

The initial solution reached at by factor analysis is determined up to a rotation. The mutual relation of each pair of variables in the reduced vector space is known and unique, but the whole cloud of points may be rotated in relation to the factors.

Giving meaning to a factor in a model is the same as making a factor *highly associated* with some of the manifest variables, and at the same time *highly disassociated* with the rest of the manifest variables. This factor can then be considered as representing a common latent feature of the manifest variables associated with it. This is also called "Simple Structure" (Jackson, 1991, p. 156).

4.2.2.1 Orthogonal rotation

An orthogonal rotation preserves the orthogonality of factors in a space, and may also be thought of as re-orienting the axes of the space. There are several types of orthogonal rotations, differing from each other in the criterion they maximize in order to achieve *simple structure*. Probably the most used is the varimax criterion, that maximizes the variance of the squared elements of the factor loading matrix *within factors* to obtain Simple Structure. Another known criterion is quartimax, maximizing the variance of the squared loadings *within variables*. As the point of departure is an orthogonal matrix, the rotated matrix is also orthogonal, and the correlation between the factors is zero. The orthogonal rotation renders a number of vectors more highly associated with fewer axes, and if the criterion is adequate, it improves interpretability.

4.2.2.2 Oblique rotation

In certain situations, different concepts or latent variables, that by the analysis are technically rendered as orthogonal, are actually known to be correlated. An oblique rotation is a transformation that, in addition to facilitating interpretation, simplifying the structure of the loadings, also tries to reconstruct such correlations. The correlations are very often kept small, because large correlations among two factors may imply that a single factor should be used. Subspaces resulting from such rotations are oblique. Known oblique rotation criteria are oblimin[6], oblimax and promax, the latter being the most popular.

Like varimax, a promax rotation maximizes the variances of the loadings within factors (axes). It is procrustean, meaning that it tries to fit a rotation matrix to a target matrix that is predetermined. The target matrix is brought about by a performing a varimax rotation, thereafter raising all its loadings to a chosen power, often referred to as κ, mostly between 2 or 4[7] (Abdi, 2003). This procedure renders the promax rotation close to an orthogonal rotation, further emphasizing the differences among strong and week components.

4.2.3 Rotations and interpretation in vector space-IR

In the history of IR there are very few attempts at interpreting the axes of reduced subspaces. The earlier attempts are reported by Borko and Bernick (1963) and Ossorio (1966), who used factor analysis for classifying documents, interpreting the axes as classes.

Rotation provides interpretability. A rotation of vectors representing documents or terms in a space will make some of the terms/documents highly associated with some axes and disassociated with other axes. An axis may thus be interpreted in terms of the objects (represented by vectors) associated with it.

In the present context, a great disadvantage of rotations is, that as they increase interpretability, they perform an undue "selection" of documents into those "more important" or "less important". A rotation would probably render quite a number of documents non-relevant to any factor. This would represent a reinforcement of a critique that is already raised by e.g. Ando (2000) in the context of LSI, stating that the LSI may suppress documents

[6]Oblimin with appropriate parameter values, can also be used as an orthogonal criterion

[7]In SPSS, which we are using, κ defaults to 4, which we are also using in our experiments.

in an unwarranted way. It is therefore reasonable to predict that rotation has the potential of increasing precision at the cost of recall.

4.2.4 Using rotations in this research

The following constraints guide the choice of rotation algorithms to be used within the present project:

- According to Chapter 3, documents and terms will be discriminated and retrieved by their extent along axes (factors).
- We are operating with large data sets, and many factors.

The first constraint implies that we are interested in rotations that, by their criteria, increase the variance of loadings within factors. Although many rotation algorithms exist that could possibly suit our purposes, the second constraint leads to the choice of well tested algorithms, that will not produce unpredicted results due to the large number of factors we experimented with. One way of ensuring this is the use implementations available in a package like SPSS. Algorithms like varimax and promax, that optimize by the criterion mentioned above, are fast and efficient, and are obvious candidates for experimentation.

4.2.5 Drawbacks of factor analysis methods

There are many methods by which factor analysis is performed, and different parameter estimation algorithms. Common to all those methods is the use of an estimated covariance or correlation matrix, often referred to as S, the analysis of which results in a multidimensional Euclidean space, within which vectors representing the *variables* are located. The interrelations of those vectors reflect the correlations or covariances between the variables as expressed in S.

One important problem associated with using factor analysis *as a method* is that it poses some demands to the data, in terms of statistical distributions which are hardly met by our term-document matrices. Therefore, Linear factor analysis *as a family of methods* will not be experimented with within this study, but some *principles* and *methods* from the factor model will be taken up. This goes for example for rotations, that theoretically are associated with factor analysis but have been used by researchers to interpret results obtained by SVD (Cheng & Dunkerton, 1995) as well as other methods (for example Konig (2002)).

4.3 Latent semantic indexing and the singular value decomposition

4.3.1 The singular value decomposition

The singular value decomposition (SVD) is a decomposition method, a generalization (for any rectangular matrix) of eigenvector/eigenvalue decomposition.

In the center of SVD lies a representation, feasible for every matrix, that allows the matrix to be written as a product of three matrices of which two are orthogonal and the third is diagonal.

For a matrix X with r rows and c columns ($r > c$)[8]

$$X_{(r,c)} = D_{(r,r)} S_{(r,c)} (T_{(c,c)})'. \tag{4.3}$$

The matrices T and D contain left and right singular vectors, respectively, and the matrix S holds the *singular values* on its main diagonal and zeros elsewhere. Note that S is rectangular, though, rows $c+1, c+2 \ldots r$ having only zero elements[9]. The column vectors of D and T are orthonormal: they are pairwise perpendicular to each other, and at the same time have unit lengths, being the eigenvectors of XX' and $X'X$, respectively. The multidimensional space represented by the SVD is a hyper-ellipsoid, of which the singular values, $S_{i,i}$, are the lengths of the semi-axes.

SVD may be used to *approximate* the matrix X as

$$X_{(r,c)} \simeq \hat{D}_{(r,k)} \hat{S}_{(k,k)} (\hat{T}_{(c,k)})', \tag{4.4}$$

which means that the matrix X is approximated by a matrix of rank k, $k < r$, $k < c$ where the matrices D, S and T are truncated to k columns (to form $\hat{D}$, $\hat{S}$ and $\hat{T}$, respectively), and S is truncated to a diagonal square matrix.

SVD is a *nested* method. It means that the k vectors comprising a k dimensional approximation are also the first k vectors of a $k+n$ dimensional approximation for any k, n so that $k + n <= min(r, c)$.

In the context of IR, the SVD may be seen as a method of approaching the true rank of a term-document matrix. The rank of a matrix is the number of linearly independent rows or columns of the matrix. The low-dimensional

[8] In eq. 4.3 and 4.4 the notation $M_{(r,c)}$ stands for a matrix with r rows and c columns.
[9] Trefethen and Bau (1997) call eq. 4.3 the "full" SVD.

approximation to the treated matrix, obtained by SVD, is the closest possible approximation *given this dimensionality* in the "least square" sense of approximation (Jackson, 1991). It means that the SVD can reveal the "true" rank of a matrix, which is mostly lower than the number of columns or rows, whichever is lower.

One problem with SVD is that it is not always the most *efficient* method of approximating a matrix, because we first have to calculate the entire decomposition, and later truncate the matrices D, S and T, which is wasteful in computer resources. Nevertheless, in this study we try to explore possibilities rather than achieve efficient transformations, and therefore the SVD will be a natural method to experiment with.

4.3.2 Latent semantic indexing

As already indicated in Section 2.2, LSI is the best known application of multivariate analysis within information retrieval. LSI uses the SVD directly. LSI represents terms and documents in the same subspace, generated by the SVD. LSI was designed to be used in partial matching of queries (pseudo documents) and documents, and could also be used in similarity measures of terms with terms, documents with documents, and (in a differently scaled subspace) also terms and documents. The method was developed to improve the measurement, scoring and ranking resulting from cosine or inner product similarities, as done in the vector space model.

The method (particularly under the name LSA) is, in the literature, attributed a certain ability to represent mental processes, for example within learning and knowledge acquisition.

> "LSA provides a potential technique for measuring the drift in an individual or group's understanding of words as a function of language exposure or interactive history." (Landauer & Dumais, 1997, p. 227)

Particularly, this goes for structures of meanings (for example of words).

> "SVD-based learning of the structure underlying the use of words in meaningful contexts has been found capable of deriving and representing the similarity of meaning of words and text passages in a manner that accurately simulates corresponding similarity relations as reflected in several sorts of human judgments and behavior." (Landauer et al., 1997, p. 50)

There are several ways to express term and document vectors for this representation, depending upon the kind of similarity calculation we wish to pursue[10]. Finding the similarity of any two terms is equivalent to taking the inner product between any two column vectors in the original term-document matrix, which is expressed by

$$X'X = (DST')'(DST') = TSD'DST' = (TS)(S'T') = (TS)(TS)', \quad (4.5)$$

where terms are expressed as TS, so that each of the coordinates of each left singular vector is scaled by the corresponding singular value, which, again is the length of the semi - axis of the hyper-ellipsoid representing the space. This will give the most appropriate representation of two term vectors for the sake of computing their inner-product or cosine similarity. A symmetric approach goes for similarity between documents:

$$XX' = (DST')(DST')' = DST'TS'D' = (DS)(S'D') = (DS)(DS)'. \quad (4.6)$$

A single entry, $x_{d,t}$ of the document term matrix expresses the association, or similarity of term t and document d. To approximate this entry in terms of the SVD we write:

$$X = DST' = (DS^{\frac{1}{2}})(TS^{\frac{1}{2}})'. \quad (4.7)$$

$\tilde{x}_{d,t}$ is, by Equation 4.7, calculated as the inner product of a row of $DS^{\frac{1}{2}}$ and a column of $TS^{\frac{1}{2}}$[11].

Equations 4.5 and 4.6 on one hand, and Equation 4.7 on the other hand, provide us with two different types of joint spatial representations, for calculating intra object and inter-object similarities, respectively. The difference between those two types is not very significant, but using equations 4.5, 4.6 and 4.7 for the respective similarities would probably give better results when using similarity (cosine of inner product) based retrieval, than using the unscaled vectors comprising the T and D' matrix. Dumais (1991) reports operating with cosine similarities on unscaled vectors[12], still with substantial success.

[10]Whereas in IR it is conventional to represent documents as columns and terms as rows, we, in this dissertation, express terms (variables) as columns and documents (cases) as rows, in line with factor analysis, where columns are variables. For the SVD, which is symmetric, this is not significant

[11]The expression is arbitrary in that the factor vectors of the term and the the document (the respective loadings of terms and documents on all the factors) take "equal shares" of the singular values. A more general expression of the SVD would be $X = TSD' = (TS^c)(DS^{1-c})'$, where $0 \leq c \leq 1$.

[12]The scaling of the vectors, if done, is not explicitly reported.

4.3.3 Methods derived from SVD/LSI

SVD/LSI has had wide acknowledgement and great impact within the IR community. Throughout the 1990s the community has followed up and debated LSI. Analyses of different kinds, as well as alternative methods in the same realm have been proposed. The focus has also changed, and many of the alternatives or variations now aim at improving the scalability of such methods, i.e. their ability to handle large and rapidly growing collections, as well as more specialized applications.

Interpretability of the axes has so far not been directly addressed by these discussions, but we note that Landauer et al. (2004) are using dimensions in a manner related to what we are doing, implying the interpretability of axes or dimensions.

The methods briefly discussed below represent two approaches to modifying the space in relation to the SVD. One approach is to maintain the emphasis of LSI on the *global structure* of the document collection, whereas the other approach is to increase the discriminating power by more closely treating the *local structures*. The first approach is represented by iterative scaling (Ando, 2000; Ando & Lee, 2001). Iterative scaling changes the way the basis vectors are calculated, equalizing their length, so that documents that belong to minority classes do not disappear in the new space. The second method, locality preserving indexing (He et al., 2004), regards the space as a submanifold, so that for each document the emphasis is the documents that are, prior to the analysis, found to reside in its vicinity.

4.3.3.1 Iterative scaling

Ando (2000) claimed that LSI inherently exposes documents that are "less prominent" in the data material to an undue filtering, diminishing their probability of being retrieved by any request. They argue that documents should be assigned equal prominence, because the immediate goal of a retrieval system is document retrieval, and the retrieval system should not, in advance, on behalf of the users, decide that certain documents are less important than other documents. They propose a modification to LSI, obtained by what they denote Iterative scaling (later iterative residual rescaling). Iterative scaling changes the space so that all documents become roughly equally prominent, at the expense of terms, for which equal prominence is less important. An outline of their algorithm is the following (Ando, 2000, p. 218):

- Start with the term-document matrix, and rescale every document vector by its length, emphasizing long document vectors and diminishing

short document vectors.

- Find the first basis vector of this new matrix using eigen-analysis. This first basis vector will point in the direction of the majority of the long documents.
- Project the TD-matrix onto the (unidimensional) subspace spanned by this basis vector, and subtract this projection from the term-document matrix. In this way we have created a residual matrix which emphasizes the shorter documents (rendering them long because of the subtraction).
- Find the first basis vector of this residual matrix. This will be the second basis vector of the new space. Repeat this process as many times as the desired number of dimensions (k).
- The k basis vectors can now be used to project every document into the space.

The above process represents a stepwise equalization of the document lengths, resulting in a space where inner product or cosine express document similarity.

In a later publication, Ando and Lee (2001) explain the success of this algorithm in terms of what they term topic dominances (roughly corresponding to singular values of the LSI representation), dominances of underlying topics represented in the term-document matrix. They show that the performance of LSI is the better, the more uniformly underlying topics are distributed in the collection. Non-uniformity of topics may be compensated for by boosting short documents. The method performs optimally when the boosting factor is adapted to the degree of non-uniformity inherent in the collection.

4.3.3.2 Locality preserving indexing

Another critique, from a different point of view, raised against LSI in the literature is that LSI identifies and preserves the *global structure* of the document collection, at the expense of the *local structure*. Notably, the first factor of the SVD (the singular vectors corresponding to the largest singular value) will point in the direction of the most often occurring terms in the collection[13], which are mostly non-discriminatory terms. The more significant a dimension (the larger the singular value corresponding to it), the more

[13]many of which would be stop-words, had these not been discarded in a pre-processing step.

"general" it is. This means that to reach discriminatory power we need quite many dimensions. This is also in line with the finding of Dupret (2003), that beyond a certain dimensionality SVD has a tendency to discriminate, whereas lower dimensionality acts "conceptualizing". The latter can be interpreted as preserving global structure.

He et al. (2004), trying to find a more discriminatory data reduction, argue for the possibility that the document space is a *manifold* rather than a *subspace*. This means that every document exists in a local projection of the space, and is located in relation to other documents in its close vicinity. He et al. (2004) have developed an algorithm to obtain what they call locality preserving indexing. The algorithm starts by expressing the coarse structure of the document collection. This is accomplished by creating an adjacency graph with documents as vertices. Edges are only drawn between documents that are close to each other in some sense, for example by Euclidean distance, rendering documents remote from each other disassociated. The threshold for closeness to a document d can be determined either by whether or not the related document belongs to d's k nearest neighbors, or by some distance measure defined for vectors of the term-document matrix. The value of the edge expresses the distance between its two vertex documents. This adjacency graph is expressed by a square symmetric matrix, S, where $s_{i,j} = 0$ if documents i and j are "remote" and greater than 0 for "close" documents. This adjacency matrix is then used to scale (or penalize) the relationships that determine the subspace.

An advantage of this method is that the dimensionality of the subspace is inherently lower than for the LSI, and document collections therefore require smaller dimensionality for good representation. The problem is that the dimensionality needs to be properly estimated for good performance. Small deviations in dimensionality penalize performance significantly.

4.4 LSI and related methods used in the Uexküll approach

The point of departure when employing LSI and related methods for rendering documents in the Uexküll approach is that the projected representation matrices of the terms and documents (for SVD T and D) into the SVD space lend themselves to a geometric interpretation, being taken as coordinates of the terms and the documents, respectively (Furnas et al., 1988).

Following the Uexküll approach, using the direction metaphor (see Chapter 3) rather than inter object similarity measures, the most important issue

for all conceivable implementations is that vectors of documents that are related to an axis, have relatively high loadings on this axis, and vectors of documents non-related to this axis are suppressed towards zero. Experiments later in this dissertation do not indicate any significant difference between stretched and unstretched spaces.

4.5 Treatment of binary term-document matrices

Multivariate methods in IR have traditionally been applied to two types of data:

- Count data - raw frequencies of terms in documents
- Transformed count data - raw frequencies are transformed into real numbered weights by *weighting schemes* like TFIDF (Salton & McGill, 1983) or Log-entropy (Dumais, 1991)

The data above are very often the result of automatic indexing: Programs traverse a collection of documents, extract words, count occurrences and perform pre-processing and post processing. Pre-processing is the elimination of stop-words (prepositions, articles, conjunctions and the like, that cannot help in discriminating documents). Post processing may be stemming and confounding of different word forms, automatic confounding of synonyms and the like.

Test collections indexed by human experts (mostly using binary indexing), have, to the best of our knowledge, not been subjected to multivariate transformations. One reason for that is that such collections have traditionally been searched using the exact match model rather than the best-match model. Another reason is that binary data of that form do not (at least theoretically) readily lend themselves to the methods that have been used (SVD and factor analysis).

Binary data are used in many branches of the social sciences, particularly Sociology (Muthen, 1989) and educational theory (Bartholomew & Knott, 1999), but these data are often not as sparse as indexing data is. For example, dichotomous questionnaire responses would include comparable numbers of positive (1) and negative (0) answers. In a term-document matrix, terms tend to index very few documents, so that methods and models developed for dichotomous social science data are not particularly fit to handle this type of data.

On the other hand, some properties of collections indexed by human experts make them potentially interesting for usage in the Uexküll context. One such property is that we have fewer terms, but the terms, often assigned by experts, are contentually much more meaningful than the automatically extracted words (often word stems). For this reason they are potentially more appropriate for automatically creating meaningful axis names than automatically extracted words would be. Intellectually assigned terms are often compound glosses of two or more words, bearing specific meaning about documents or document groups.

4.6 Points for further research

4.6.1 The true dimensionality of a term-document matrix

The methods that have been presented here reduce the dimensionality of the term-document matrix into an arbitrary dimensionality that may be determined by some heuristics. There is no clear cut answer as to the optimal output dimensionality.

For nested methods like SVD (LSI), heuristic methods exist that determine the optimal dimensionality, e.g. with the aid of a scree plot, a plot that shows the variability explained by the dimensions in descending order (Jackson, 1991). A more algorithmic approach (Dupret, 2003) introduces the "validity rank" of a keyword: SVD is computed for the *covariance matrix* of the terms in a term-document matrix. The covariance matrix, S, is then reproduced with different approximations, represented by different truncations of the matrixes of singular vectors V and U, and the matrix of singular values, Σ. *For each diagonal element j (variance of a keyword, or the keyword's covariance with itself), no other keyword should have a larger covariance with that keyword than the keyword itself.* For a certain keyword, a dimensionality reduction that causes the violation of this rule is termed "invalid". Experiments indicate that low dimensionality solutions tend to conceptualize - meaning that they tend to confound keywords with similar meanings, whereas higher dimensionality solutions tend to distinguish meanings of words with a similar occurrence pattern. The validity rank for a keyword is the transition point between the two effects. This result would be different for different terms, but statistics may be applied to find a collection-wise validity rank, or region of diminishing return for increasing the number of orthogonal factors.

An additional approach, Amended Parallel Analysis (Efron, 2005), repeat-

edly draws matrixes that are statistically similar to the corpus matrix, under the assumption of term independence, and generates confidence intervals for each eigenvalue. The true dimensionality, k, is the last eigenvalue of the analyzed matrix, for which the magnitude is above or within the confidence interval of the parallel eigenvalue generated by the simulation. Looking at the Uexküll approach, it is an open question whether any of the above methods would uncover the true dimensionality, i.e. the dimensionality that would render the best Uexküll performance of a matrix.

The experiments conducted as a part of this project are conducted with four dimensionalities, 36, 75, 158 and 309[14]. As summaries of the results will show, most of the best results are obtained for dimensionality of 158 and 75, which may hint at some optimum lying somewhere between these values. Relying on rotations further complicates the matter, as the rotated versions of the same dimensionality perform differently, and have various peaks in different dimensionality. The optional Uexküll dimensionality is an issue that should be subject to further research.

4.6.2 Methods that could take larger data sets

An important property of some of the methods presented here is that they reduce the dimensionality of a data matrix explicitly. Some of them, like the Eigenstructure based methods, e.g. LSI, depend on the computation of a full decomposition, of which some dimensions can be truncated. Hoenkamp (2003) proposes a method of dimensionality reduction which is arbitrarily sparse in storage requirements, and therefore could take far larger data sets. The method uses the Haar transform[15]. The Haar transform is a scaled step function that can be used in averaging (or low-pass filtering).

Consider any discrete function and two adjacent values of it. These points may be represented by a single point being their average. If the function curve has many points, we can average adjacent pairs of points, obtaining a filtered version of the function, where high frequency noise is eliminated at the cost of some loss of information (this is analogous to the way LSI filters the original term-document matrix). The Haar Transform represents a way to reconstruct the original curve. If we perform the averaging of two

[14]The choice was somewhat arbitrary: Dimensionalities were used as codes under empirical experiments testing different methods. The decompositions above represent four different *levels* of dimensionality reduction. The large amounts of data involved in the experiments prohibited a more fine-grained scale of dimensionalities

[15]This effort actually represents a research that tries to utilize and extend advances made in signal processing and image analysis in order to facilitate multivariate IR.

points (of magnitude x and y, $x > y$, $a = \frac{x+y}{2}$) by adding them to $a - x$ and $y - a(= -(a - x))$, respectively, then it is sufficient to store the average, a along with $a - x$.

This line of thought represents a treatment which has the potential of addressing much larger data sets than are allowed by the traditional multivariate methods, but the adaptability of such methods to serve the Uexküll approach is not obvious.

4.7 Limitations of using multivariate methods for Uexküll visualization

In this section we discuss some difficulties that are associated with our treatment of the document and term data for visualization.

4.7.1 Properties of axes

Consider SVD as one example: In SVD the dimension associated with the largest singular value expresses the best unidimensional approximation of the data matrix. The dimension associated with the second largest singular value is, together with the first one, the best two dimensional representation. On its own, this second singular vector is the next best unidimensional approximation of the data matrix (subject to orthogonality against the first solution). This means that the first and most significant singular vector represents terms that appear very often in the collection, and do not discriminate documents very well. As we descend down the dimensions (by their singular values) we may find solutions that do not represent the data matrix well enough, and which are apt to introduce more and more noise. This means that we cannot operate with too many dimensions. Rotating the axes, using some criteria, we probably get better interpretability on more axes, but at the same time introduce other problems, such as undue suppression of documents (see also Subsection 4.2.3).

4.7.2 Number representation errors

One problem of generating high-dimensional solutions the way we do would be the problem of number representation and accumulating error. Some methods may be vulnerable for these problems, and there exist methods/models, whose purpose is to counter-balance such effects. In this dissertation we do not introduce, or implement any explicit treatments of such effects. Likewise,

we do not perform any error analysis of our results to monitor such effects. We believe that the use of modern computers with a large word-size to perform the analysis limit such round-off errors, allowing us to ignore them in the experiments.

We further assume that e.g. the recommendation appearing in Golub and Van Loan (1983, p. 37) that "Whenever the hardware permits, it is always recommended that inner product accumulation is used for matrix-matrix multiplication" is followed where applicable.

Part II

Evaluation of the Approach

Chapter 5

Evaluation of the Uexküll approach

5.1 Introduction

In this chapter we are presenting the general idea of the evaluation of the Uexküll approach. After a brief general introduction to evaluation in IR, we introduce our methods taken from the best-match paradigm to retrieval evaluation, where we, through user simulations, evaluate an approach that is basically not a best-match approach. Thereafter we discuss our test collection, the Cranfield collection, and the reasons for choosing it, followed by a number of technical issues concerning the evaluation. Such issues are the representation of data organizations in coordinate databases, the representation of requests, and the technicalities of the retrieval of documents. Details of the user simulations, as well as presentation and discussion of the evaluation metrics are deferred to later chapters.

5.1.1 Evaluation in IR

As already mentioned in Chapter 2, evaluation of retrieval systems has undergone an evolution since the Cranfield experiments (Spärck Jones, 1981), until the recent phases of the TREC experiments. This evolution may be briefly characterized as beginning with a pure *system centered approach*, using expert selected (sometimes graded) topic-relevant sets of documents (also called *recall bases*) as the only reference for query based performance testing. Within this paradigm, evaluating retrieval effectiveness of *best-match* systems has been done by using a document database, also called a test collection, with a set of predefined requests, for which recall bases have been defined. Those requests, formulated into queries, are processed by each system, and

for each query the *ranked output* of a system is compared to the recall base of the corresponding request.

In recent years, user interaction has become a more prominent component of evaluating IR systems. Two examples of this trend are the interactive track of TREC (Over, 1998), along with the development of task oriented approaches to interactive IR (Borlund & Ingwersen, 1997).

There has also evolved a dichotomy between the approaches, where laboratory based approaches were challenged. Kekäläinen and Järvelin (2002a) have advocated that the laboratory based approach be used where, for development of algorithms, even within the frame of more user-oriented experiments.

5.1.2 Approaching the evaluation of Uexküll

The Uexküll approach consists of two major parts.

- A data organization part, at the heart of which lie databases of term and document locations in various multidimensional spaces.
- A user interface/interaction part that utilizes the data organization for spatial visualization and retrieval.

An evaluation of user interfaces is traditionally done with user participation (Shneiderman, 1998). [1] A total evaluation of the Uexküll approach is also dependent on being performed with user participation, but the outcome of this evaluation is obviously dependent on the data organizations used during the tests. A user test of such a system is a resource consuming endeavor, requiring a lot of participants' (usually scarce) time resources, and very prone to errors of various kinds (Over, 1998). Whatever findings can be produced prior to such a test, reducing the number of unknown factors, will undoubtedly result in a more effective user test. Consequently, an evaluation of the approach with user participation has, at an early phase, been considered out of scope for the current dissertation. However, for the further development of the approach, beyond this dissertation, an evaluation with user participation is, of course necessary.

This leaves us with a separate evaluation of the data organization part. In this evaluation we would like to incorporate as much as possible of the features of an Uexküll based system. Those features of Uexküll are:

[1] Newby (2002) reviews some evaluations of visualization systems.

- visual interaction - navigation rather than keying search terms,
- focus on prominent documents - how good is the support for bringing prominent documents so that they are easy to focus on.

Therefore, the result of an interaction is good if it:

- makes documents judged relevant appear farther out on axes thana non-relevant documents;
- clearly distinguishes relevant items from non-relevant ones;
- where applicable it also places partially relevant documents properly, so that they are distinguished from both the highly relevant and the non-relevant ones[2];
- the user is given a lead to further pursue his search.

The evaluation method presented in the following sections is meant to get an indication of how well these elements are supported by the data organizations.

5.1.3 Main idea and problems

The main idea underlying the choice of method is to evaluate the data organization as a *best-match* system, using a test collection with a set of requests each having a recall base, but into this evaluation incorporate some aspects of the user interaction. An important motivation for using the *best-match* approach is the observation that an Uexküll scene (see Chapter 3) regards documents as being of different prominence. A problem of this approach of evaluation is that functionally an Uexküll based implementation is not a best-match system, as it is not designed to provide linear ranked lists of documents as a response to a query. Looking solely at the number and rank of relevant documents retrieved is not very insightful for the evaluation of an approach like Uexküll. Notions like "relevant documents", "precision" and "recall", in relation to a reference set, are less predictive in an associative

[2]Users, in different parts of their information seeking process may also be interested in documents that may only be partially relevant (Spink, Greisdorf & Bateman, 1998), and it would be advantageous if the data organization could separate those from the totally non-relevant ones. However, one needs to use recall bases with graded relevance judgements (where available) in order to test this (Järvelin & Kekäläinen, 2002). This evaluation is not pursued within this dissertation. This decision was taken even though graded relevance judgments (5 point scale) were available.

context than with traditional keyword search[3]. Nevertheless, given the current state of the development of the approach, and the fact that no real user tests are performed as a part of this dissertation, we need to adopt some traditional principles of best-match evaluation, but do it, as much as possible, adapted to the features of Uexküll. The latter means we also measure some parameters associated with support for user interaction.

5.2 The subject of evaluation: data organizations

The purpose of the evaluation approach is to find out how different data organizations would serve users accessing document collection through an Uexküll based system. A data organization consists of a decomposition method of some dimensionality, and a rotation (see Subsection 4.2.2). The purpose is to get an indication of how different data organizations would support Uexküll-based retrieval. To this end we represent a test collection as a multidimensional space through different decompositions and subsequent rotations. These result in different *data organizations*, that we regard as systems in terms of the best-match paradigm.

5.2.1 Decompositions

Multivariate statistics provides a great number of decompositions that could conceivably serve for the representation of a collection of documents as a vector space of reduced dimensionality. The current dissertation is not going to offer a solution regarding the best possible decomposition, rather it present an evaluation approach that can be used to evaluate such decompositions against each other. The classification in this research regards different dimensionalities of the same method as different decompositions. This is convenient regarding different rotations, because of the rotation's dependency on the dimensionality. LSI/SVD has been extensively experimented with in this approach, and the results we are presenting here include four different dimensionalities. The choice of dimensionalities is discussed in Subsection 4.6.1

[3]Also their more general prediction power for user satisfaction has been questioned (Saracevic & Kantor, 1988; Su, 1992).

5.2.2 Rotations

In Chapter 4 we have presented a number of popular rotations, both orthogonal and oblique. We also mentioned that there exist many more rotations. In the social sciences, rotations are often used with relatively few factors (typically less than 10). The dimensionalities we are working with, tens and hundreds of dimensions, put certain strain on the rotation algorithms. We are therefore dependent on using well tested algorithms, such that are implemented in commercial packages like SPSS. Our strategy is to experiment with popular rotations like varimax and promax, but, for the purpose of comparison, also performing limited experiments with other types of rotation algorithms.

5.3 Using the best-match paradigm in simulation experiments

The best-match evaluation paradigm is about comparing systems. It is based on quantifying, for each system, the correspondence between the average system response to a number of queries and the reference response, represented by the recall bases of the queries.

To apply this evaluation paradigm to the Uexküll approach, we need a test collection of documents and queries with recall bases. Next we need to represent both documents, terms and queries as vectors, corresponding to our data organizations. To this end we store the coordinate values for each term and document in each data organization in a coordinate database. We *simulate* user interaction through queries, letting each query represent a user need. We evaluate the interaction success using recall bases. For each query (under each data organization), we do the following:

1. Convert it into an Uexküll group

2. Download a scene from the coordinate database that corresponds to this Uexküll group

3. Represent the scene as a ranked list of documents, where each document has an RSV (retrieval status value) representing its extension along an axis.

4. Evaluate the list against the recall base of the query.

The section to follow discusses the test collection and its usage in some detail. The conversion of a scene into a ranked list is governed by a "location model" (point 3 above), the details of which are discussed in Chapter 6, along with the details of representing user needs through queries. The evaluation (point 4) is done using traditional, as well as specially designed measures. These are discussed in Chapter 7.

One contribution of this thesis is that it seeks to develop the algorithmic component on the basis of novel usability criteria. In the evaluation discussion that follows we develop an apparatus, which measures features that are beyond traditional retrieval, namely direct visual support – visual access to relevant document as well as visual suppression of non-relevant documents. The dissertation also seeks to develop the evaluation approach. Traditionally the algorithmic component is only evaluated for its R-P performance, which, though being usability criteria, must be supplemented by additional measures to accommodate for this type of usability evaluation.

5.4 The test collection

5.4.1 The Cranfield II collection

In the following we list some of the requirements ideally met by a test collection used in this effort:

- Multivariate statistic may be computationally demanding, both in CPU and storage resources. Testing different rotations of different decompositions (dimensionalies) will require each coordinate value of each term and each document in each rotation and each dimensionality to be stored. Additionally, we would like to accommodate for the test of various algorithms for both transformations and rotations. Such algorithms will not always be optimized for many variables/dimensions. Given the power and storage capacity of modern computers, each of the above is possible to accommodate for with a large collection. Still, in order to keep the experiments manageable, we have found that the test collection should preferably be small.

- The collection needs to have a sufficiently large set of queries with recall bases of relevant documents (Buckley & Vorhees, 2000; Spärck Jones, 1976).

- As mentioned in Subsection 1.2.2, we would like to test the adequacy of the approach used with intellectual indexing. We therefore wished to have an intellectually indexed test collection.

A collection satisfying all three demands was not easy to find. The Cranfield II collection satisfied the two first items above, but the indexing is not of the type we were looking for, as it was done by extraction. Nevertheless, the extraction was performed and post-processed intellectually (see Subsection 5.4.3), which makes it reasonably suitable for testing the approach. Additionally, we believe that the fact that it is within a limited subject area makes it more suitable than it would be if it was, e.g., a collection of news items. It would, of course be an advantage had the collection been more diverse, but we believe that for pursuing the approach described so far in this chapter, the lack of diversity is acceptable.

The Cranfield collection has become an institution in information retrieval, and the name is often used to denote a paradigm (laboratory based evaluation) within IR. The collection consists of 1400 documents about high speed aeronautics and aircraft structure. It also includes a large set of queries, with a set of documents judged relevant for each query. This collection has been used in information retrieval experiments for 40 years, with recent examples dating to the past few years (Tai, Ren & Kita, 2002; Dupret, 2003).

5.4.2 Documents and requests

To create the test collection, 271 *base documents*, at that time recent papers about high speed aeronautics and aircraft structure, were collected, obeying to some criteria. An example of a criterion was that it should have at least 12 references. References in these papers were used as a fan to collect more related papers. Authors of base documents were asked to formulate research questions giving rise to the research producing the base documents, as well as formulate some question that emerged in the course of the research, and could be, or indeed were, directed at some information service. To each of those questions they were to assess the relevance of their own papers and references. An additional list of potentially relevant documents for each question (prepared by post graduate students in the service of the project) was submitted to the authors for final assessment at a later stage. The entire process resulted in 1400 documents and 279 requests with graded relevance assessments as recall bases. The present project uses 225 of the requests that were digitally available with the relevance judgements. The relevance judgements,

originally graded, have been dichotomized (see Subsection 5.4.4.1).[4]

The process of relevance judgements for the Cranfield II project has been subject to debate in the years to follow the experiments. Those debates actually went to the heart of the problems of measuring retrieval effectiveness, and are interesting as such. A central point is the close interdependence between documents and requests, as requests were designed by the authors of some of the documents, directly addressing the core of these documents. Even though the documents giving rise to the requests were excluded from the relevance set of their own request, they may be said to have been closely associated with other documents, due to the method by which the collection was built. Another particular problem associated with this is the existence of "missed" relevant documents, pointed out by Swanson (Harter, 1996), claiming that thousands of documents actually relevant to one or more of the queries were missed.

A very important point, addressed by Ellis (1996a), is that the relevance judgement should be founded on the basis of *user relevance* rather than *stated relevance*, thereby simulating more closely the process of document search by users with a real information need.

> However, the implementation of user relevance judgments in the tests represented a compromise with the real life situation in two key respects. The first was in the reliance on stated requests, the second in the assumption that the information needs and user relevance judgments were static. It was clear that in both these respects the tests were unrealistic (Ellis, 1996a, p. 28).

5.4.3 Indexing effort

The Cranfield experiment was meant to research into how different aspects and properties of indexing affected retrieval performance. Therefore the indexing was designed to allow variation in the extent to which those aspects were applied. This was done by:

- Defining "Devices of indexing", some of which promote recall, others promote precision. These could be applied to a "device-less" index, thereby creating languages and measure retrieval performance when these languages are used.

[4]It would be interesting to use the graded relevance judgements and see to what extent the ranking of the relevant documents is in line with the relevance grades. Originally, the relevance of each document to each request was judged on a 5 point scale, where 1 is highly relevant and 5 is not relevant at all (see Appendix E.1 for a complete description).

- Defining the device-less index to be a very exhaustive, natural language based index: "a crude, elemental index language from which all the other languages (...) would be derivable." (Cleverdon, Mills & Keen, 1966b, p. 48).

The indexing of the Cranfield collection was meant to be used for "searching by subject prescription only", apart from classes established by bibliographic coupling, as an oblique way of getting at subject content (Cleverdon et al., 1966b, p. 41):

> [...] a subject index must provide facilities for adjusting and manipulating its classes; it must allow the index classes examined to be expanded or contracted, and in different directions, until a match with the search prescription is recognized. Index language devices are the agents of this manipulation. They are devices whereby class definitions may be adjusted to meet the requirements of different searchers.

5.4.4 Characteristics and use of the Cranfield collection in this dissertation

In this dissertation we are using two versions of the collection:

- An automatically indexed version, where an extraction process is applied to the abstracts (1398 documents), followed by a normalization.
- An early version, for which index terms were manually extracted from the texts. For this version (where occurrence counts of terms in documents were originally provided) term-document associations are dichotomized, so that any count larger than zero is replaced by one. The purpose is to approach the behavior of an intellectually indexed collection.

In Table 5.1 we provide some data of the versions of the collection used in the present dissertation.

5.4.4.1 Relevance judgments

As already mentioned, the original relevance judgments for the Cranfield collection were made on a 5 point scale, where 1 was highly relevant, and

Table 5.1: *Characteristic of the versions of the Cranfield collection used in the experiments*

Item	Automatically Indexed version			Manually indexed version		
	minimum	average / number	maximum	minimum	average / number	maximum
Documents		1398			1399	
Terms					2686	
Unique terms*		2338			2159	
Queries		225			225	
Relevant documents per query	2	8,16	41	2	8,16	41
Terms per document	11	52,11	155	5	29,5	95
Terms per query	3	8,17	20	3	7,92	17

*after theremoval of contextual synonyms

5 was not relevant at all[5] (see Appendix E.1). In the current project, the relevance judgments are dichotomized such that only documents of level 5 are taken to be non-relevant (0), whereas the rest are taken to be relevant (1). This was decided because it gave a collection with a somewhat larger set of relevant documents for each query (see data about recall bases at different dichotomization levels in Appendix E.2. It was also assumed that searchers having an in depth user need (see Subsection 1.2.1) would find such documents relevant. It would, of course, be interesting to perform experiments with a different dichotomization (e.g. where both documents of relevance 4 and 5 are taken to be non-relevant), and see the effect.

5.4.4.2 Preprocessing of the automatically indexed version

The automatically indexed version of the collection[6] is provided with stemmed vocabulary. Further normalization included the finding and discarding of *contextual synonyms* among both terms and documents. Contextual synonyms are here defined as terms that index exactly the same documents, as well as documents that are indexed by exactly the same terms. These do not contribute to the rank of the term-document matrix. An important next step was the removal of terms with collection-frequency 1, leaving us with 2338 terms. This is done exclusively to reduce the size of the collection, as-

[5]In addition, the document on the basis of which the request was formulated was marked with -1.

[6]This version has been downloaded from the FTP-site of Cornell University.

suming that such terms do not contribute to the inter-document association structure of the matrix. This was done while executing care not to remove any term that is the sole entry to a document, which would mean discarding of a document from the collection. The left half of Table 5.1 presents the important figures characterizing this version.

5.4.4.3 Preprocessing of the manually indexed version

A digitized version of the collection was downloaded[7] (our gratitude to Dr. Mark Sanderson). This distribution included a version of the indexing (2686 terms as opposed to 3096 in the most exhaustive index, see also Subsection 5.4.3).

Skimming of this version, superficially comparing it to the term list in Cleverdon, Mills and Keen (1966a, p. 220), we found it to confound word forms, which was actually one of the indexing languages implemented by the Cranfield study (see Subsection 5.4.3). This language was given in the report to consist of 2541 terms Cleverdon et al. (1966a, p. 59) as opposed to the 2686 terms found in the version we used. The discrepancy may be due to the fact that not all word forms of a group are confounded in this version.

We have not applied any further normalization to this version. The only reduction of the matrix we performed was finding and removing contextual synonyms (see Subsection 5.4.4.2) among both terms and documents.

5.4.4.4 Using the versions of the collection

For each of the two versions of the collection described above, the following were loaded into a database:

- The documents represented by title, abstract and identification number (common to both versions),
- The index terms
- Term and document association (as expressed in the term-document matrix),
- The 225 queries in terms of their constituent index terms,
- Relevance judgements (common to both versions),

[7] `http://www.dcs.gla.ac.uk/idom/ir_resources/test_collections/cran/`

- coordinate values for terms and documents as obtained by the analyses (Chapter 4). This part of the database will be referred to as the *coordinate database*.

Retrieval was done by querying lists of documents ordered descendingly by coordinate values along an axis representing the query (this is elaborated on in Chapter 6). For every instance of retrieval (each data organization/each query) a constant number of documents (40) was retrieved. The number was chosen being near the highest number of relevant documents for any query, which meant that the R-precision measure (Salton & McGill, 1983) could practically be taken for all the queries.

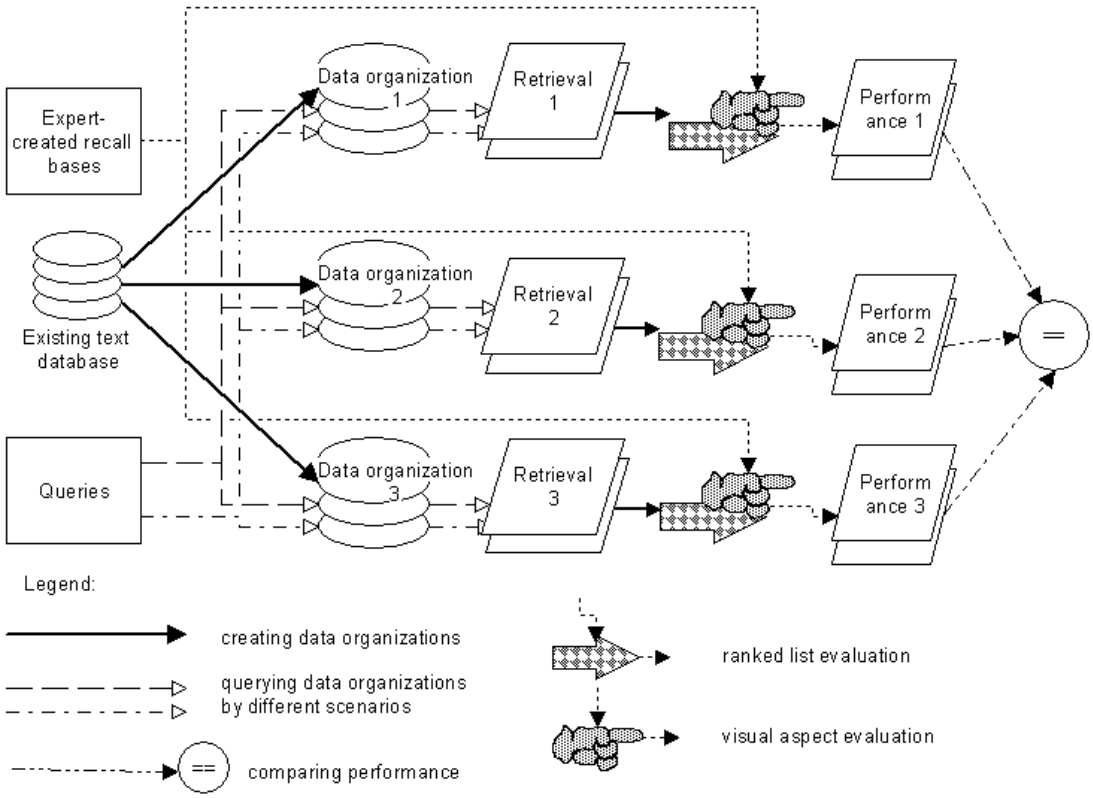


Figure 5.1: *Comparing data organizations, the evaluation approach*

5.5 Summary of the evaluation approach

The evaluation situation is depicted in Figure 5.1. To the left lies a known text database with queries, recall bases and relevance judgments. From this database several (in this sketch three) different data organizations are created (solid bold lines). The same queries are directed at all 3 organizations (dashed lines), and results are compared to the recall bases (dotted lines). The circles with the equality sign in them symbolize comparison situations, and one can see that the performance of the newly created organizations is compared.

The dashed lines with transparent arrow heads symbolize procedures, or *simulation scenarios*, by which queries are applied, and results extracted. These will be discussed in Chapter 6. The pattern-filled arrows stand for the evaluation of retrieval results as ranked lists, in line with the traditional laboratory model. The hand-shaped pointers represent measurement of aspects of visual interaction with the system (functional testing of data organization). Such aspects are, for example, how well relevant documents are drawn up the axes, and how well non-relevant documents are suppressed towards the origin. These differences are further elaborated on in Section 7.4.

Another, important aspect is the extent of the interpretability of the dimensions.

- How many axes participate in the query responses?
- How many items participate in successful queries?

In Chapter 7 we discuss the types of performance measures that lie in the center of the evaluation of our simulation experiments. The simulations themselves are discussed in Chapter 6. The performance measures used for the simulation based evaluation are chosen to best capture the mode of work Uexküll is meant to support, and at the same time to reduce or counterbalance the drawbacks of the "one-collection evaluation", as expressed above.

Chapter 6

Scenarios of user simulations

As already indicated in Chapter 5, the Uexküll approach will not be subjected to user tests within the present effort. The choice was to subject the system to some sort of simulation, where we

- Submit queries with expert judged relevant document sets to different multivariate statistical data organizations generated from the test collection,
- Manipulate these queries in a way that captures aspects of the way a user would retrieve documents based on a user need,
- Using the evaluation apparatus developed in Chapter 5, rank different data organizations according to retrieval results.

In the present chapter we will discuss simulation and IR in light of the literature, describe and argue for simulation scenarios and develop simulations based on *location models* for documents in the Uexküll groups.

6.1 Viability of non-interactive investigations and simulations in IR

In this section we are discussing a number of approaches to simulation within information retrieval. The approaches discussed do not by any means comprise an exhaustive list, but serve to contextualize the simulation approach taken by the present project. Heine's approach (Subsection 6.1.1) is discussed first because it discusses simulation within IR in a wide context (not in connection to a certain project), giving examples of different uses of simulations. Magennis and van Rijsbergen (1997) (Subsection 6.1.2) apply simulations in

a situation where user tests would be natural to use, but such tests are prohibited by practical circumstances, whereas Leouski and Allan (1998) (Subsection 6.1.3) and R. W. White, Jose and Ruthven (2004) (Subsection 6.1.4) have done simulations in a way that is meant to *complement* user tests. Lin (2007) simulates users in a QA environment, drawing on the laboratory evaluation model, augmented with a specially designed performance measure.

6.1.1 Heine's approach

Heine (1981) introduces simulations as either verbal, indicative processes or applied mathematics or computerized techniques, in a way that always presupposes:

- **Optimization:** Identifying something that is "best" according to some criterion
- **Judgement:** Identification of the system's essential components.
- **Input and output:** real artefact, decisions or information as part of the system description.
- **Intervention:** which is a natural consequence of optimization.

Heine further contrasts "simulation in the narrower, mathematical sense" with "investigation and experimentation", where for the latter "an experimental apparatus is needed for them to be implemented [...] and no suppositions are made as to how the information acquired is generated". Simulation, on the other hand, "does not require an apparatus, and does concern itself with how information (data) [sic!] is generated by a system".

Interpreting Heine's terminology in the context of the present effort, a user test would be termed "Experimentation" as it requires an apparatus (a functioning GUI as well as a number of test users and a way of assessing those users' retrieval results). The simulation experiment in question does not require this apparatus. It does, however require a collection with queries and expert judged recall bases associated with each of those queries. As a rule, simulations include an element of randomness, driven by knowledge we possess about the statistical properties of a process.[1]

[1] A classical example of what is known as "Monte Carlo simulation" is the simulation of the flow of customers in a post office. In such an exercise we may define say four "counters" in an imaginary post office, and for a simulated period of time draw "customers" at random (from a known distribution) to occupy the counters. In this case we use our knowledge of

This is, however not always the case in IR simulations, and the random element is not always there (at least not explicitly). Heine provides an example[2] where he uses results from an experiment to draw precision/recall surfaces produced by different conjuncts of a set of terms. Heine wishes to show that "purely formal constructions can usefully be discussed and compared using familiar information retrieval concepts, with no additional definition and dealing only with observables". One way of regarding the present project is that we extend the above mentioned example to the evaluation of the visualization support provided by different multivariate organizations to the Uexküll approach. As our "element of randomness" we use the queries of the collection, assuming they are sufficiently representative (instead of drawing them randomly from a distribution).

According to Heine, we need a formal description of the system that we simulate. Such a description will, in our case, be obtained by expressing the locations of objects in space (we call these "location models" and they are defined and described in Section 6.2). We must also explicate and justify our simplification of the supposed user interaction into what we call "scenarios", omitting parts of the detailed process we expect to happen in reality. Another point Heine mentions (p.195,196) is the connection between a "query" and a "user need". It is important to stress, though, that the question whether e.g. a Cranfield topic can be attributed to some user need, is not a part of the present project.

6.1.2 Magennis and van Rijsbergen's approach

Magennis and van Rijsbergen (1997) simulated users doing multiple iterations of interactive query expansion. They wished to "determine the potential and actual effectiveness of interactive query expansion in a large scale realistic search context"(p. 325). In their experiments, terms for query expansion are selected by the retrieval system and ranked by some procedure. User behavior is simulated by simply letting the machine decide the cut-off (number of

the probability that a customer will enter the post office, in order to generate customers at random for our experiments. The servicing time of a customer at a counter may also be drawn from some known distribution. Then we count the number of customers successfully serviced, estimating the efficiency of the 4 counters. Before estimating the efficiency we may wish to run such a simulation repeatedly and generate statistical descriptive data for our sought efficiency.

[2]example 3 (Heine, 1981, p. 188) (The 'logical surface' and 'document weighting surface' of a set of terms). In this example observed retrieval data created within a retrieval experiment are read into a program that implements an algorithm that plots the probability distribution of precision and recall values regarding different retrieval strategies applied to an information need.

terms added to the original query) "on behalf" of the users. Here there is no reference to any explicit system model, and no apparent stochastic component that will represent variations in the behavior of individual users. Instead, every possible behavior within a range of behaviors (selecting every combination of 5 predefined cut-offs through 4 iterations) was used.

The similarity of this approach with the present effort is that users are simulated explicitly, and that individual differences between users are not modelled.

6.1.3 Leouski and Allan's approach

Leouski and Allan (1998) wished to evaluate interaction against a 3D graphic configuration of documents retrieved as a response to a query. The authors' assumption was that if the rank ordering of a result set is incompatible with the user's need, the user would begin by identifying a known item somewhere in the list, look for items similar to this one, and thereafter items similar to the latter ones, and so on. This assumption leads them to "reshuffle" the ranked list by ordering documents that are "similar" to each other close together in a visible 3D configuration by some algorithm. The user interaction against such a list is simulated by relevance judgements that are made available to the system (as if the user himself had supplied the relevance judgements), and the ordering of the documents in the 3D configuration is updated as a result of these.

Leouski and Allen's effort is similar to the present project in that it uses known retrieval parameters (e.g. relevance judgments) to evaluate an ordering of objects. This effort, in similarity with ours, does not have a proper random component.

6.1.4 White's approach

The effort of R. W. White et al. (2004), done within *implicit relevance feedback*, is a relatively recent addendum to the literature describing simulations within IR. An *implicit feedback model* prescribes how information about user behavior can be used to modify queries without users' explicit intervention. Specifically here, different *document representations* that a user traverses are used as evidence in the selection of query-augmenting terms. The purpose here is to evaluate the performance of different implicit feedback models. The approach uses *relevance paths*, as evidence to be used by an implicit feedback model, to simulate users' conception of relevance of a document to a request or a topic. A relevance path describes a user's traversal of different

representations of a given document. A retrieval system may offer a number of representations for each document, e.g. the title, the summary, the full text. A relevance path will include some of these representations in some sequence. Based on terms from the representations constituting the relevance path and the distance travelled along the path (number of representations a user traverses), the implicit feedback models under scrutiny generate terms to augment queries.

R. W. White et al. (2004) simulated users' choices of relevance paths by drawing at random from a set of given relevance paths for each document. Each topic (query) is associated with a pre-set list of terms (taken from documents relevant to the topics and ordered by normalized occurrence statistics). The term lists brought about by the implicit feedback models are compared to those pre-set lists, using correlation measures. Precision and recall are used to evaluate the retrieval outcome.

In similarity with our approach to simulation, and in line with the laboratory approach, this approach uses topics to simulate information needs (at least implicitly), extending the laboratory approach in the direction of mimicing user interaction. Whereas we use newly devised measures to extend the scope of the laboratory approach, R. W. White et al. (2004) use correlation measures for a similar purpose.

6.1.5 Lin's approach

As another relatively recent approach to simulation in IR we choose to examine Lin (2007), describing simulation of user behavior within question answering (QA). This effort seeks to evaluate how previously evaluated QA and IR systems perform in a QA environment, particularly addressing question series. Question series represent a challenge as they model communication with users within a context rather than mere fact oriented answer supply to a single, context devoid question.

Lin's simulations draws heavily on the laboratory evaluation model, augmenting the evaluation design with explicit *user models*. The user models supply the system component with data and observe the system output.

Lin modeled two type of users, the "QA-styled" and the "IR-styled" users. The QA-styled model is confronted with a number of TREC-evaluated QA-systems, and the IR-styled model with two IR systems. In similarity with our approach, performance is evaluated using a specially designed evaluation measure. Unlike our approach, it is the interaction of user model and system type that is tested.

6.1.6 The role of simulation in the present project

Figure 5.1 in Chapter 5 sketches the evaluation of the data organizations' properties as a part of the Uexküll approach. The simulation scenarios described below are symbolized by the pattern-filled pointers.

Performing the simulations is interesting in itself as a part of our contribution, namely to algorithmically characterize aspects of usability of a graphical retrieval approach.

In addition, the simulations serve to focus and constrain the necessary user tests, thereby (hopefully) rendering these more useful. In other words, user simulations should determine the data organization or organizations against which user tests, at a later stage, shall be performed.

6.2 Deriving a uni-dimensional location model for the simulations

6.2.1 A location model

A location model, in this dissertation, is defined as *a geometric simplification of the interrelations of a group of objects in a 3-D space, based on their location in that space.* It has the following features:

- It tries to capture the way a document would be retrieved visually (by a user interacting with an Uexküll-based system), so that the entire process can be algorithmically simulated.
- It is used as the basis of ranking documents for retrieval based on the query centroid.
- It can therefore be used in algorithmic tests (hereafter: scenarios), that simulate visual retrieval, by representing it as best-match retrieval.

6.2.2 Relatedness to an axis represented by loading

Figure 6.1 (A slight modification of Figure 1.3) exemplifies dimension reduction by reducing dimensionality from 3 to 2. It illustrates that objects that are pertinent to an axis get a high loading on that axis. Sample objects on this figure are not directly related to sample objects used in subsequent figures.

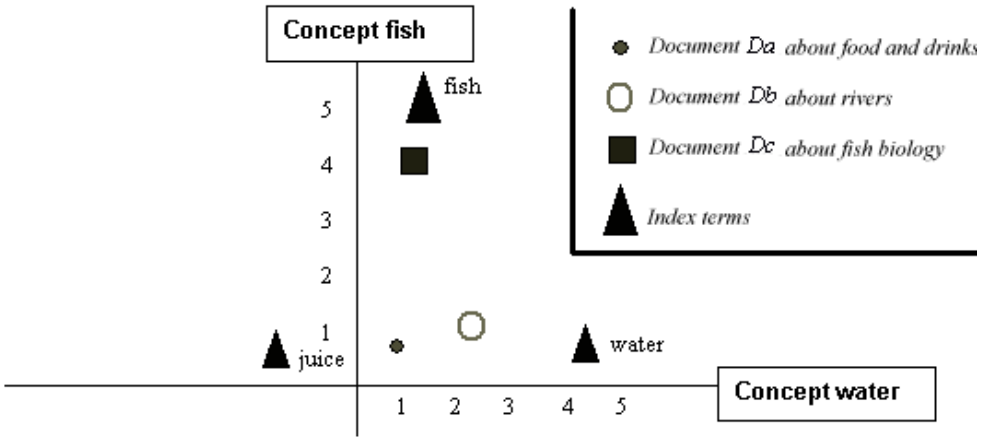


Figure 6.1: *Objects pertaining to an axis have high loadings on that axis*

In the following figures objects marked with Di, Ti, Ai and Ci (i representing any ordinal number) represent documents, terms, axes and concepts, respectively. In Figure 6.2 documents D1, D2 and D3 are located in an Uexküll group for which the drawn axis is one of the axes. Based on the documents' loadings on this axis, document D3 is the most related, and document D2 is less related. Document D1 has a loading below a predefined threshold, and is therefore considered unrelated to this axis.

6.3 Combined axes and combined loadings

Points in a 3D configuration have no inherent linear order. In order to represent each object's three dimensional locations so that it can be a part of a ranked list, we need to represent this location by a single number, in a way that best represents the assumed relative prominence of the object among all objects in the scene.

Since we are opting for a representation which is as simple as possible, and have currently no grounds to assume any non-trivial weight differences among concept axes, two alternative representations are plausible: the sum (or the average)[3] of the loadings, and the highest loading, referred to as the *sum model* and the *max model*, respectively.

[3]since the value of the combined loading is only going to be used for ordinal or scaled ranking, the sum and the average are equivalent, and the average is therefore not presented as a distinct alternative.

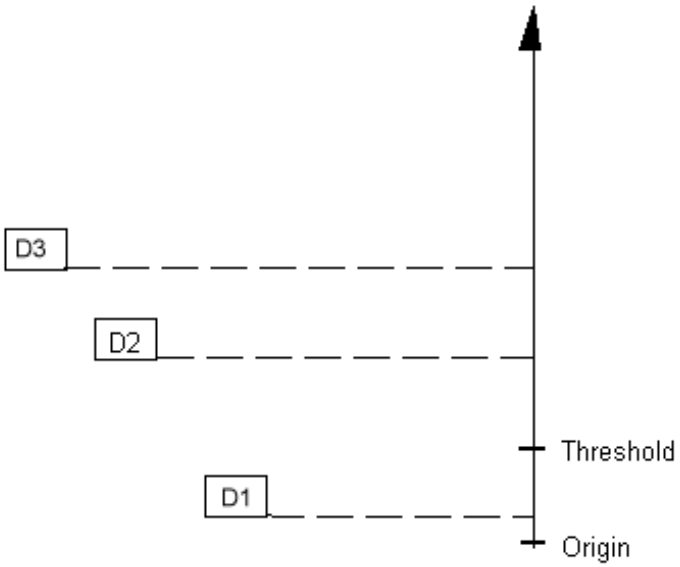


Figure 6.2: *Objects' relatedness to an axis is expressed by their loading on that axis*

Which model that best supports the representation of interaction against a scene, is assumed to depend on loadings along coordinate axes, but also on how that model would capture user behavior. For example, if we knew that users would very often download a scene based on three concept axes, but only pursue one of the axes, then an appropriate model would represent the three loadings of a document by one of these, probably the highest loading. If we knew users more often than not actively use all three axes of a downloaded scene, then a combination of all axes (sum or weighted sum) would seem appropriate.

6.3.1 The max model

In Figure 6.3 we illustrate an Uexküll group and the *combined axis* of an Uexküll group based on the max model. For each document, the combined axis holds the *combined loading*, which is the highest loading on any of the three axes. Thus, whereas the combined loading of D1 equals its loading on A1, D2's combined loading equals its loading on A3.

The max model would reward documents that load high on a single axis equally as high as if the document loaded slightly lower (still high) on an additional axis. The model ignores the fact that two documents may have a reverse order on a single axis than the order they acquire on the combined

axis, and may, in some instances, represent a discrepancy in relation to real life retrieval.

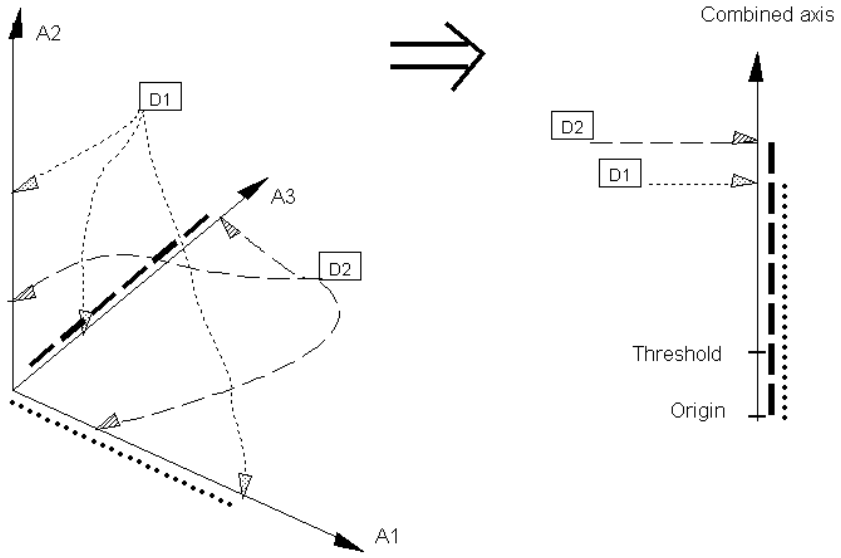


Figure 6.3: *The combined axis of an Uexküll group using the max model*

6.3.2 The sum model

In Figure 6.4 we illustrate an Uexküll group and the combined axis of an Uexküll group based on the sum of the loadings. For each document, the combined axis holds the *combined loading* which is the sum of the positive loadings on all the axes[4].

The sum model represents the entire scene. It rewards documents that are related to more than one of the axes, giving them a higher ranking than documents that are only related to a single axis (with the same or comparable loadings).

6.3.3 Setting the threshold

Another way to explain the max and sum models is through the relationship between setting the threshold on the individual axis and setting it on

[4]Negative loadings are truncated to 0. In factor analysis, a document that is negatively loaded on an axis is assumed to be negatively correlated with the latent variable represented by the axis, but we decided to truncate such loadings to 0, and assume they are not related to the axis.

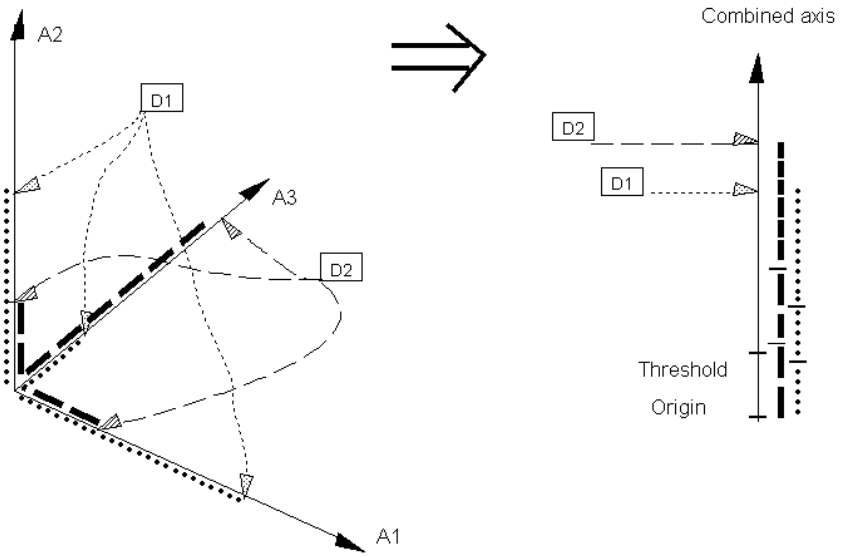


Figure 6.4: *The combined axis of an Uexküll group using the sum model*

the combined axis. Figures 6.5 and 6.6 show two-dimensional views of the threshold setting for the max and sum models, respectively. To avoid confusion, we use different indexes than in previous figures. Documents residing in the white areas are regarded impertinent to the Uexküll group. An interesting difference between the models is manifest through documents D6 and D7, which are below the threshold for both individual axes. The max model renders them impertinent to the Uexküll group, whereas the sum model renders them pertinent, as the sum of their coordinates is higher than the threshold. In practice this is not a significant difference, as it is unlikely to concern a great number of documents. It rather illustrates that the sum model is better at capturing the topicality of the Uexküll group as a whole, rather than the significance of each of the constituting axes.

6.3.4 Choosing a model

As already mentioned, we have not conducted any practical experiments that could indicate how users would behave when interacting with an Uexküll-based system. We are, however, basing our modeling of user behavior on the direction metaphor.

The intended use of an Uexküll-based system, exemplified in Chapter 3, prescribes that users construct Uexküll groups based on three chosen concepts. Choices of two or one are also possible, but we see three distinct concept as

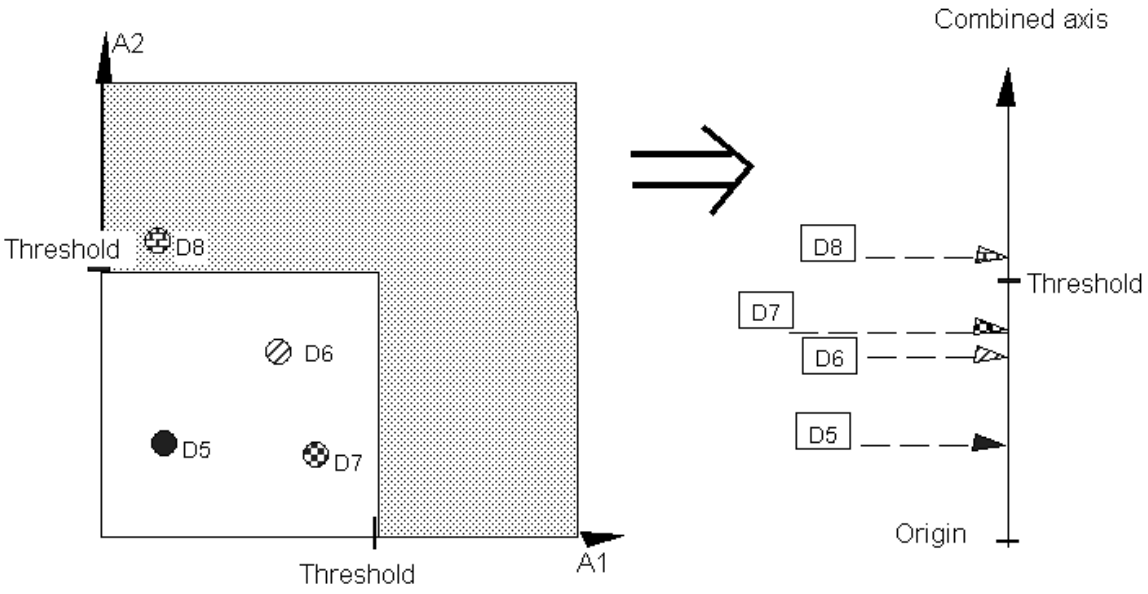


Figure 6.5: *Setting the threshold using the max model*

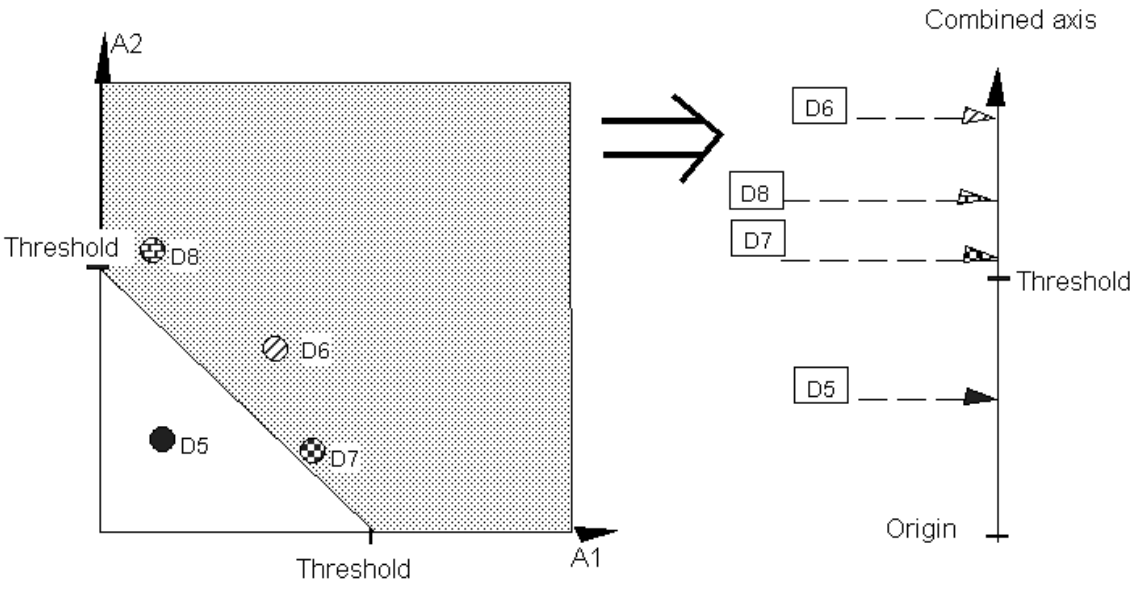


Figure 6.6: *Setting the threshold using the sum model*

a typical case. The discussion in the preceding subsections illustrates that a good representation of this process should take all three concepts axes into account, pointing at the sum model as a better representation than the max model, leading to the choice of this model to use in our simulations.

6.3.5 Using a query in representing a user need

In Figure 6.7 we illustrate the way a query represents the initial information need of a user. The centroid vector of the query term in the multidimensional space is calculated, and the three axes with the highest loadings (OCP, see Subsection 7.2.1) are chosen to comprise the Uexküll group.

The three axes of the Uexküll group are then represented by a single axis, the combined axis, on which any object is represented with the combined loading derived in the previous section.

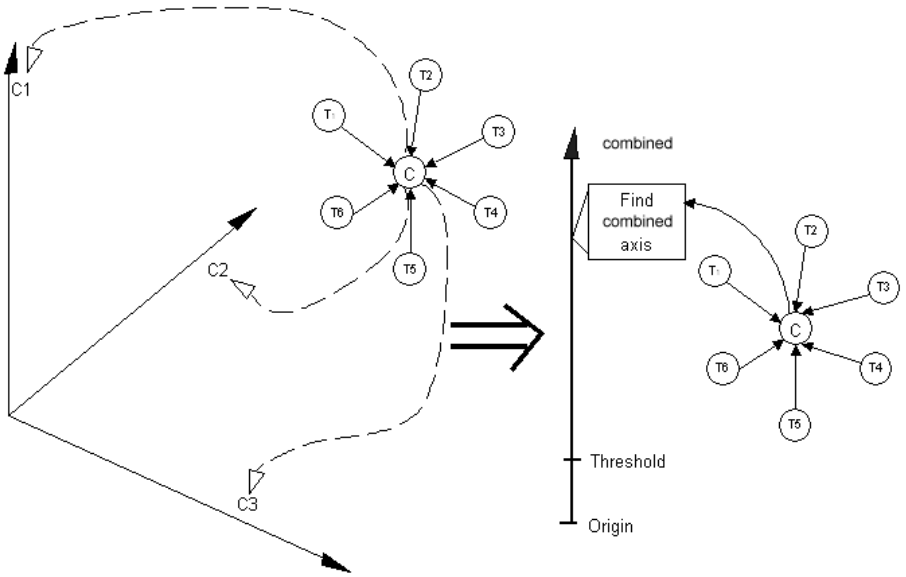


Figure 6.7: *The construction of a combined axis by the centroid of a query*

Figure 6.8 summarizes the location model: Documents having loadings below the threshold on the combined axis (combined loadings below the threshold) are not part of the Uexküll group.

It is important to note the difference between this organization and the traditional use of methods like LSI. A traditional LSI-based retrieval system returns a ranked list which is based on the angular vicinity between a spatial representation of a query and of the documents, ranking the angularly

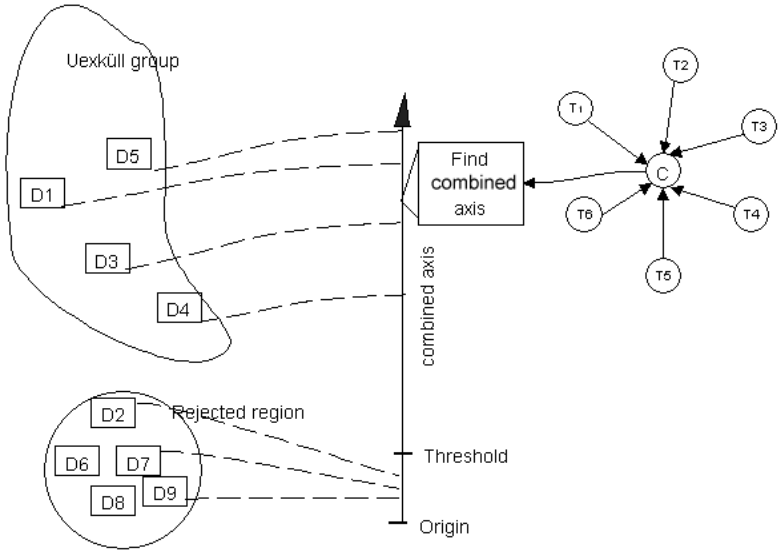


Figure 6.8: *An illustration of the location model to be used in the simulation scenarios.*

closest documents highest. The difference may be illustrated through the effect of an orthogonal rotation on retrieval. This rotation would not affect the traditional LSI ranked list[5]. The reason is that the coordinate axes are not participating in traditional LSI retrieval. On the other hand, the list returned by the location model is dependent on the coordinate axes, and therefore sensitive to rotation.

6.4 Scenarios based on a location model

In this section we develop simulation scenarios that build on the location model depicted in Figure 6.8. The scenarios:

- implement the location models;
- represent (in different degrees of detail) the interaction discussed in Subsection 3.3.3;
- produce ranked lists of documents, where the RSV for each document equals its relative loading. This implies that the highest ranked document in every Uexküll group has $RSV = 1$.

[5]An oblique rotation may change the ranking slightly.

This allows us to evaluate results based on both ranking (precision and recall) and separation of relevant - non-relevant documents, using measures to be discussed in Chapter 7.

6.4.1 The notion of relevance

The role of the simulation scenarios described in the remaining of this section is to represent Uexküll scenes as ranked lists of retrieved documents. In such a context discussing the term "relevance" is inevitable, and care must be exerted.

Characterizing our evaluation environment in Mizzaro's (1997) terminology, our relevance judgements were originally *expressed* by values taken from a five point *rating scale* (p. 814) between a document and a request (p. 812). In line with our location model, our data organizations return ranked lists of documents, where documents are more or less relevant, expressed by an RSV. Here, our organizations, through this model, provide a *numeric estimation (magnitude estimation)* form of *expression* to the relevance of a *document surrogate* to a *query*. As we technically express this magnitude in terms of relative magnitudes, assigning the highest ranked document in a certain response the value 1, and wishing to avoid declaring a document "100% relevant" to a request, we choose to use the term prominence regarding this kind of relevance. We are going to talk about document or term prominence, expressed by the relative loading of a document or a term on an axis, and axis prominence, the most prominent axis in a context being the axis for which the highest loaded object is loaded higher than the highest loaded object on the other axes.

The term *relevance* remains used in its original sense: a relevant document is a document that is *judged relevant* by a domain expert.

6.4.2 Simulation scenario 1

Below follows a step by step procedure of the first scenario (see also Figure 6.9), followed by a short discussion.

1. The query centroid represents the user's initial "information need".
2. The Uexküll group chosen for this query consists of the three axes on which the query centroid has the highest loadings.
3. Documents belonging to this Uexküll group are retrieved and assigned an RSV, which is the *combined loading* relative to the highest combined

loading for that query, so that the highest loaded document gets the value 1.

4. The documents are then ranked by their RSVs.

An Uexküll group pertaining to the centroid is constructed by the three axes of the coordinate database for which the query centroid vector has the highest loading (OCP, see Subsection 7.2.1).

In Figure 6.8, documents D1, D3, D4 and D5 are considered likely to be found by the user, whereas documents D2 and D6 to D9 are discarded, their loadings being below the threshold.

This scenario is a simple representation of the interaction as outlined in Subsection 3.3.3. However, here the choice of Uexküll groups leads directly to documents. The difference is that the description in Subsection 3.3.3 is inspired by a certain implementation of the approach, whereas we, in the simulations, try to evaluate the *approach* rather than a specific implementation. Another simplification here is that this scenario does not include navigating between scenes. The derived Uexküll group thereby represents the entire interaction. Navigation is included in the next scenario, presented in Subsection 6.4.3.

6.4.3 Simulation scenario 2

Simulation scenario 2 is a further development of simulation scenario 1, and is depicted in Figure 6.10. It attempts at bringing the simulation conceptually closer to an interaction situation, corresponding to the implementation described in Chapter 3, including the navigation between Uexküll groups. The idea here is to let the query represent a user need that develops into a search process. Below follows a step by step description:

- The query centroid represents the user's initial "information need".
- The axes for the Uexküll group are chosen as the three axes on which the query centroid has the highest loadings.
- Documents belonging to this Uexküll group are retrieved.
- The query terms are sorted in ascending order of their distance from their centroid[6].

[6]In simulation scenario 2, as implemented here, this point is not important, as documents are not weighted based on how quickly they were found. A further development might draw on this sequence of terms to simulate an interaction that is interrupted at some point, or weights documents differently based on the sequence of the terms.

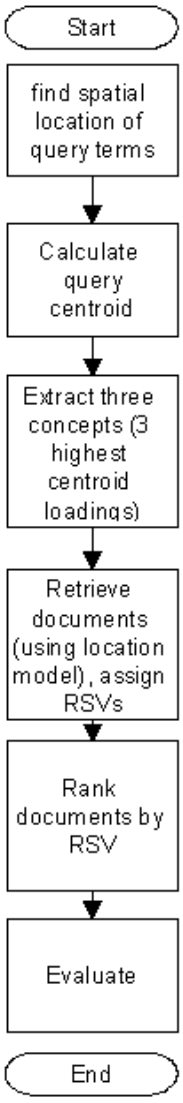


Figure 6.9: *A flow chart of simulation scenario 1*

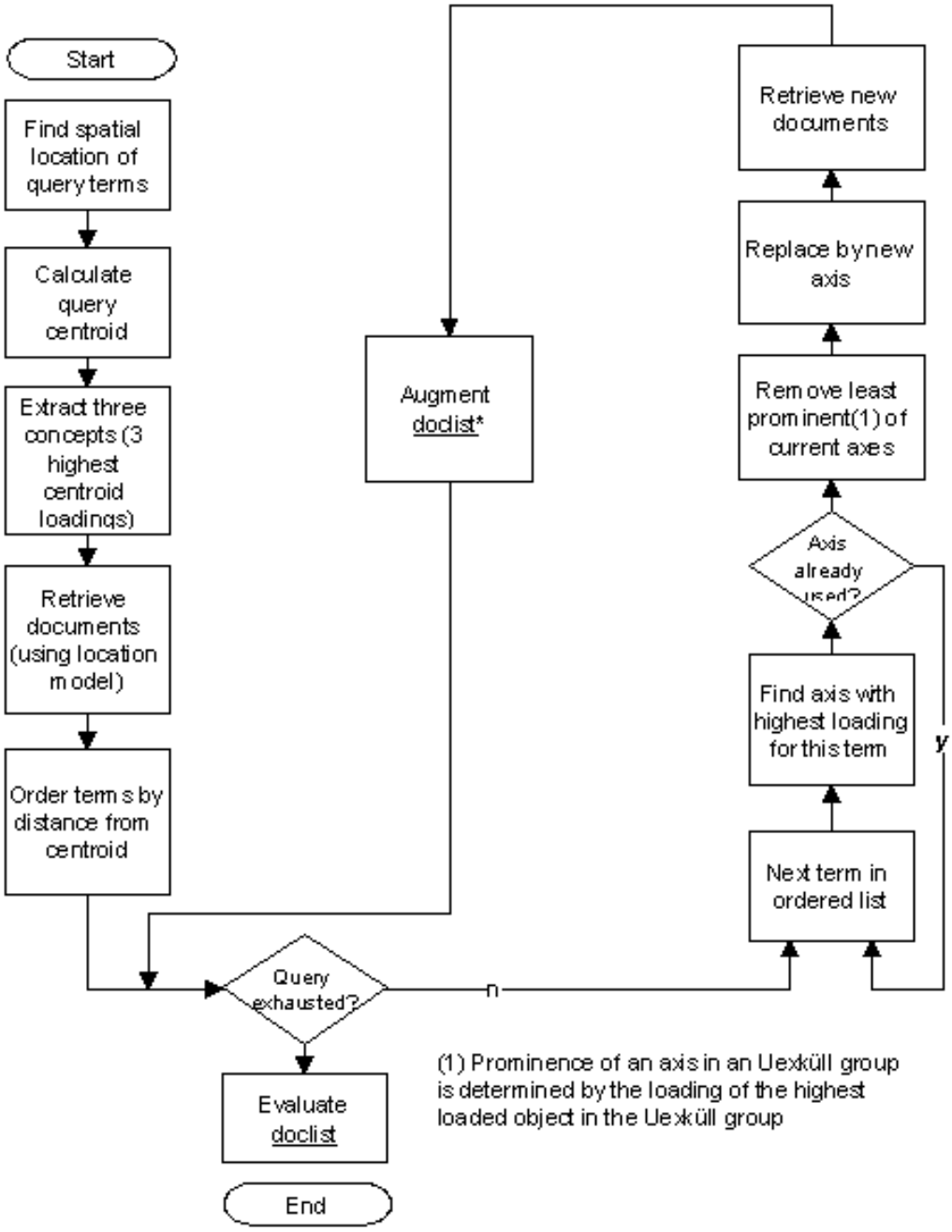


Figure 6.10: *Simulation scenario 2 - a summary view (*see a detailed view of augmenting the document list in Figure 6.12)*

- For each term t in this sequence:

  - If the axis on which t has the highest loading in the entire multidimensional space is not in the current Uexküll group (or any Uexküll group so far used) - substitute it for the least prominent axis[7] in the current group.
  - The two most prominent axes of the previous Uexküll group are kept.
  - Retrieve any document that belongs to the new Uexküll group and is not yet retrieved by the previous interactions).

Figure 6.11 illustrate how axis C3 (assumed to be the least prominent) is replaced by a new axis on which T1 is the most prominent (see figure 6.7 for reference). A similar process occurs for all terms of the queries (if its most prominent axis has not already been used).

This process is repeated until the query is exhausted. The documents collected are subsequently compared to the query's recall base, as a part of the evaluation. This scenario thereby involves more than one Uexküll group, and can be regarded as a process.

The first Uexküll group is created using the centroid of the query terms. The scenario then tries to find the documents most pertinent to this Uexküll group. This step represents the user's initial approach for the search. Then a new projection is defined with the two most prominent axes being kept, and a new axis, substituting the currently least prominent axis, is chosen. The new axis is chosen so that it best represents the next term from the query.

Note that the step whereby the user selects the terms (those terms which in real use eventually lead her to the documents) is in this simulation omitted, as we are evaluating the approach rather than a specific implementation.

6.4.4 The accumulation of retrieved documents within simulation scenarios

As each of the scenarios proceeds, documents with high RSVs are accumulated in a "user query result table", and assigned an RSV figure. For simulation scenario 1 the accumulation is simple, a single list is created in a single interaction. For simulation scenario 2, the accumulation is somewhat

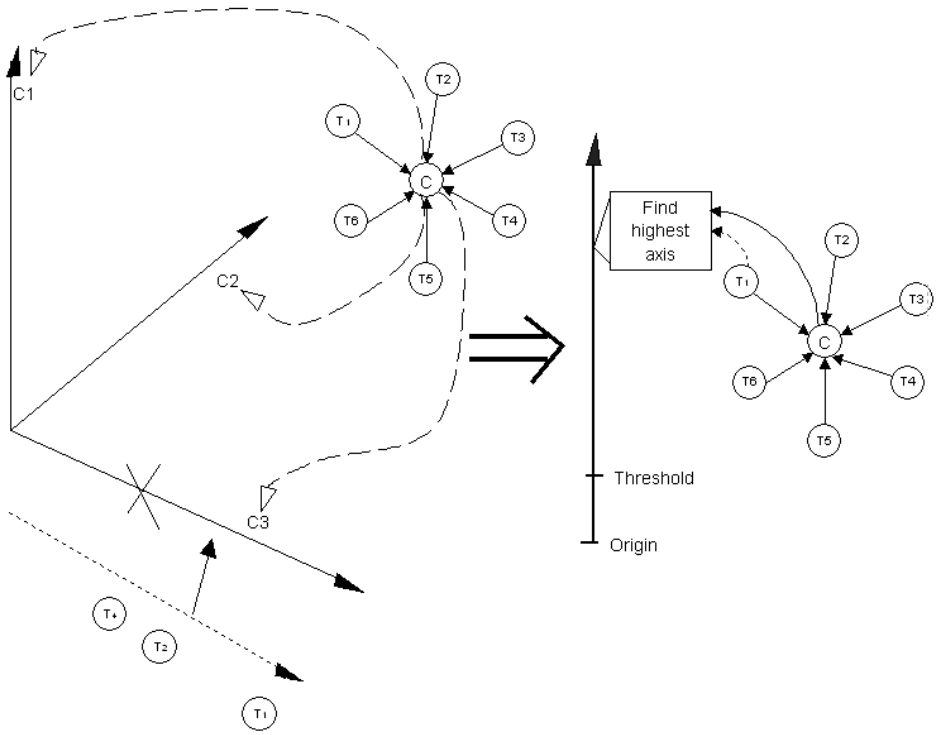


Figure 6.11: *Simulation scenario 2 - updating of axes: C3 is replaced by a new axis on which T1 is the most prominent term. Refer to figure 6.7*

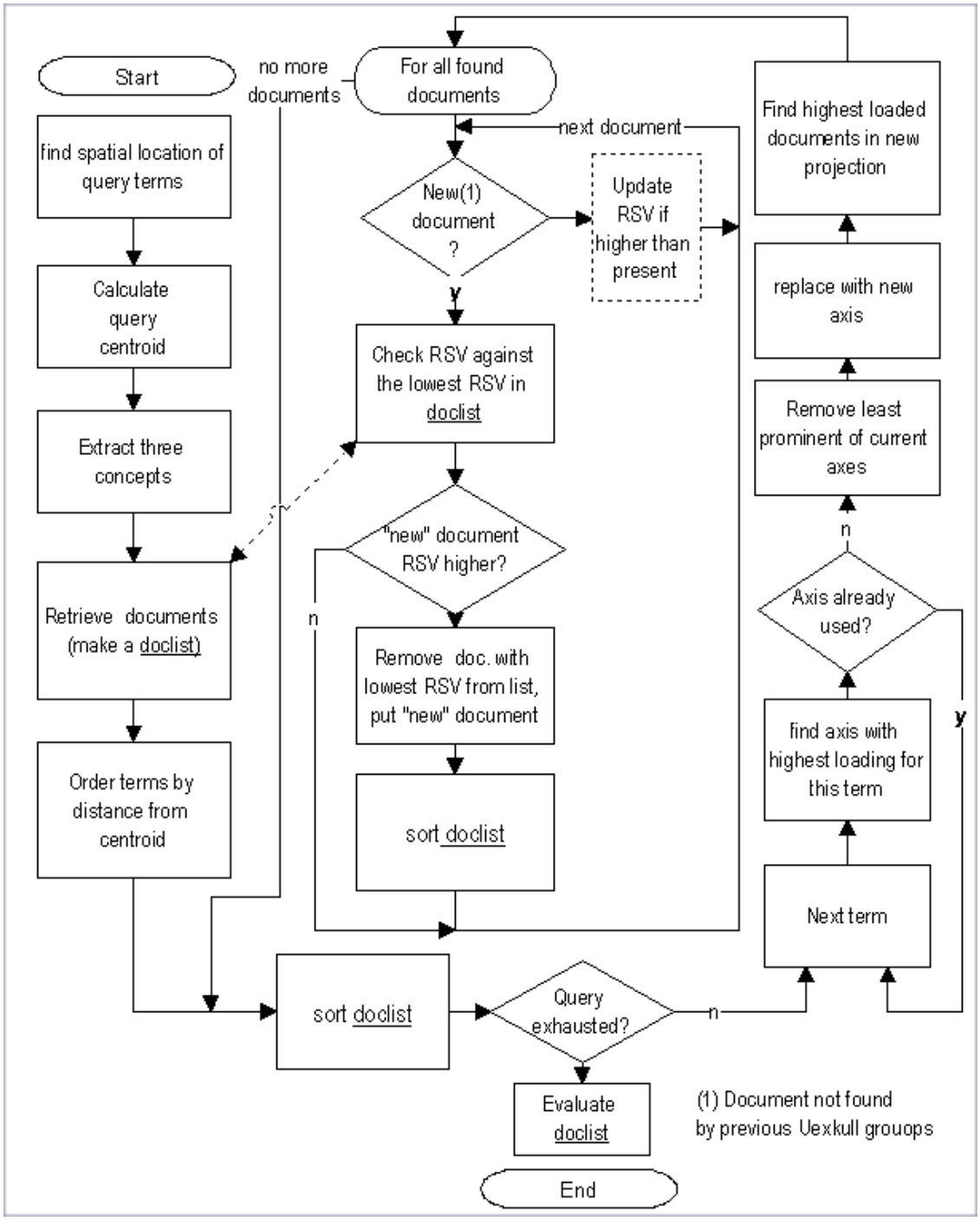


Figure 6.12: *Simulation scenario 2- a detailed view. The list of retrieved documents is being augmented as the search advances*

more complicated, as it simulates a process. The accumulation of retrieved documents is depicted in a more detailed flow chart (Figure 6.12).

Here are some principles governing this accumulation:

- The loading of every retrieved document along the combined axis is divided by the loading (on the same axis) of the highest loaded document. Thus, the highest loaded document in a query response always has an RSV equal to 1.
- In a real search situation Uexküll groups are not mutually comparable, meaning that users would not be able to compare the prominence of a certain document across Uexküll groups. A system implementing the approach, on the other hand, could record documents encountered or selected more than once during an interaction and compare (normalized) RSVs.
- To make the simulations account for that case where a document is retrieved across more than a single Uexküll group, we have considered the following alternatives:

  1. Ignore repeated occurrences, keeping the RSV first found for this document.
  2. Update the RSV if the new RSV is higher than the previous RSV for that document, so that the highest RSV obtained for this document is recorded as the document's RSV for that query. (see dashed box in Figure 6.12)
  3. Add a small increment to the original RSV.

  We chose to pursue option 2, assuming it will best represent the search situation.

6.5 Strengths and weaknesses of the approach

This section summarizes some of the strengths and weaknesses that are associated with the use of user simulations in the evaluation of our approach.

As already discussed in the beginning of Chapter 5, the approach subscribes quite loyally to the laboratory evaluation model, and this alone violates some assumptions about the way an Uexküll-based system is expected to function,

[7]The prominence of an axis in an Uexküll group is determined by the highest loading of any object in the scene on it.

mainly that use of queries or text based requests is not fully in harmony with the way a system implementing the Uexküll approach is meant to work. The simulations will therefore not be able to capture e.g. aspects of serendipity (Foskett, 1996), which are important for the Uexküll approach. We also see that questions can be raised about the fidelity by which simulation scenario 2 is able to represent a user interaction.

Nevertheless, we believe that the simulation approach is viable in this project, because the level of ambition regarding the aspects the simulations are supposed to capture is sufficiently limited - namely the extent to which the placement of documents in the space and in Uexküll groups would promote retrieval of relevant documents during an interaction.

Regarding transferability, we have chosen a text collection with many queries, and are pursuing a conservative interpretation of the result. Even though the topicality of the collection and the queries at hand may not be typical of what we expect all Uexküll users to be interested in, the results are transferable within the level of ambition mentioned above.

We therefore believe that the simplifications that the scenarios entail are reasonable in the context.

Chapter 7

Measures of retrieval effectiveness

7.1 Introduction

Measuring of information retrieval phenomena has presented a dilemma ever since the first retrieval tests (Ellis, 1996a). Unlike measurement of physical phenomena, measuring of retrieval effectiveness has a subjective element in it, that one cannot escape, and that inhibits the objectivity of the measurement. Measures of retrieval effectiveness are also tightly associated with the notion of relevance. Relevance, as already mentioned several places in this text, is a volatile phenomenon, being both subjective and varying with time. Particularly in an associative context as Uexküll, one needs a certain awareness of the various pitfalls the ambition of quantifying relevance entails, particularly when devising new measures deriving from relevance.

The evaluation design sketched in Chapter 5 needs a number of performance measures that can be applied to the simulation scenarios (Chapter 6). This chapter discusses IR *performance measures* in general terms, and in the context of the Uexküll approach. As already indicated, the evaluation design seeks to compare different data organizations and tries to assess

- the individual organization's suitability to serve the Uexküll approach;
- the relative performance of the organizations when used within the Uexküll approach.

This means that for each data organization we will first try to see if it at all supports the Uexküll mode of retrieval (test of suitability, Section 7.2). Two types of measures are then applied to those organizations that are found suitable:

- Measures that are traditionally associated with the effectiveness of best-match IR-systems. Performance is measured in terms of counts and ranking of relevant vs. non-relevant documents in response to queries (Section 7.3).
- Measures that incorporate the visual support of the data organization (Section 7.4), particularly the organizations' ability to separate relevant documents from the non-relevant ones.

It may be claimed that the last mentioned measure, combining the two aspects, should be sufficient when characterizing the data organizations. However, in our analysis we consider the ranked list aspects to be important, particularly for comparability within IR-evaluation following the laboratory model. The second type of measures superimposes novel aspects on the traditional evaluation, that are of importance to this particular project. Following Newell (1969), Sormunen (2000) and Pharo (2002), the combination of these aspects in our evaluation will be *justified* by showing, that the relationship between those types of measurements, and the results they provide, are not trivial. Presenting results of both types illuminates the contribution of these aspects to evaluation of the approach, and thus may provide insight into the properties of the data organizations, as well as the evaluation approach as such.

7.2 Centroid emphasis

An important principle on which Uexküll is based is the traditional interpretation of factor analysis results, where variables that pertain to factors (axes) load highly on them. Uexküll groups have three dimensions. Therefore the Uexküll group that best represents an object in a multidimensional space consists of the three axes on which this object has the highest loadings.

The overall evaluation procedure chosen for this project consists of the following main steps (see also Chapter 5):

- representing a query by the Uexküll group consisting of the three highest loaded factors of the centroid of its constituent term vectors,
- downloading documents that load high on a representative axis of these three factors (see Section 6.3.2),
- evaluating the organizations by observing the distribution of the loadings of the relevant and non-relevant documents on the representative (combined) axis.

We therefore need some measure that, for each organization we use, can indicate that the organizations can be meaningfully analysed by the above mentioned steps.

7.2.1 Suitability of an organization - the interpretation power of the OCP

Consider the centroid vector of a query in some data organization. Like any other vector, and in line with the traditional interpretation of factor analysis, its favorable Uexküll group would consist of its 3 highest loaded factors (axes), provided, of course that the data organization opens for this kind of analysis. We call this Uexküll group the *optimal centroid projection* (OCP). One way of testing whether the projection consisting of the query's three highest loaded axes is indeed the favorable Uexküll group, is to compare the performance of the OCP to that of other, supposedly inferior, projections.

The procedure is simple:

- We extract the centroid of a query and the Uexküll group thereof, retrieve the 40 documents[1] with the highest combined loading in the OCP by that combined loading, and count the number of relevant documents among them. We repeat this exercise for all the queries, and count the grand total of relevant documents retrieved at DCV@40[2].
- We repeat the procedure above for some supposedly inferior projections.
- We summarize the results both graphically (Figure 7.1) and arithmetically.

The extent to which the OCP performs better than inferior projections is denoted the *centroid emphasis*. The CE as a measure is defined as follows:

Let i, j and k represent the three axes that have the i^{th}, j^{th} and k^{th} highest loadings of the centroid point of a query, respectively. $CP_{<i,j,k>}$ is the *centroid projection* based on these axes. CP_i will be used as a short form for $CP_{<i,i+1,i+2>}$. By this notation

[1]The number 40 is motivated by two requirements: 1) downloading as few documents as possible, to simulate as little cognitive load as possible on the user 2) since 40 is the highest recall base of any query in our test coollection, this number allows us to use the R-precision measure.

[2]DCV stands for document cut-off value, a predefined number of retrieved documents at which precision and recall are calculated.

$$OCP \equiv CP_{<1,2,3>} \equiv CP_1.$$

Let $\rho_{q,i}$ be the set of relevant documents retrieved with query q and projection CP_i. Let

$$R(CP_i) = \sum_{q=1}^{Q} |\rho_{q,i}|$$

be the aggregate number of relevant documents retrieved for all the requests $q = 1...Q$, with CP_i. In the case of the Cranfield collection, where Q=225 and 40 documents are retrieved for each query

- $\rho_{1,1}$ would be the number of documents judged relevant of the 40 documents retrieved for the first request under CP_1
- $R(CP_1)$ is the number of documents out of 40*225 = 9000 that were judged relevant when CP_1 was used with each request.

Let

$$A = \{CP_{i_1}, CP_{i_2}, ...\}$$

be a set of centroid projections[3]. The centroid emphasis metric can be expressed as

$$CE = \frac{1}{|A'|R(OCP)} \left[|A'|R(OCP) - \sum_{i_n \in A'} R(CP_{i_n}) \right], \tag{7.1}$$

where

$$A' = \{CP_{i_1}, CP_{i_2}, ...\} - \{OCP\} = A - \{OCP\}$$

is the set of all centroid projections with which the measure is calculated, excluding the OCP, and $|A'|$ is its cardinality.

Equation 7.1 is an expression of the relative superiority of the OCP. This superiority is captured by computing how many more relevant documents

[3]The extra index, i in the equation is needed because CP_1, CP_2, CP_3 are associated with specific centroid projections, and we will not always use all of the centroid projections in an evaluation situation. $CP_{i_1}, CP_{i_2}...$ are arbitrary projections. In our experiments we were using CP_2, CP_3, CP_5, CP_7.

the OCP retrieves as opposed to the supposedly inferior projections. We multiply $R(OCP)$ by $|A'|$ so that we can compare it to the aggregate number of documents retrieved by the inferior projections. This is the same as taking the average of the inferior and subtracting R(OCP). We choose the sum for legibility reasons. Data organizations for which the OCP retrieves any relevant documents and the inferior projections retrieve none, will score 1. If there is no real interpretability of axes, we will expect the number of relevant documents for all projections to be random, which, over a large number of queries will converge towards equal numbers of relevant document for any projection, so that the measure converges towards 0 (Losee, 1998).

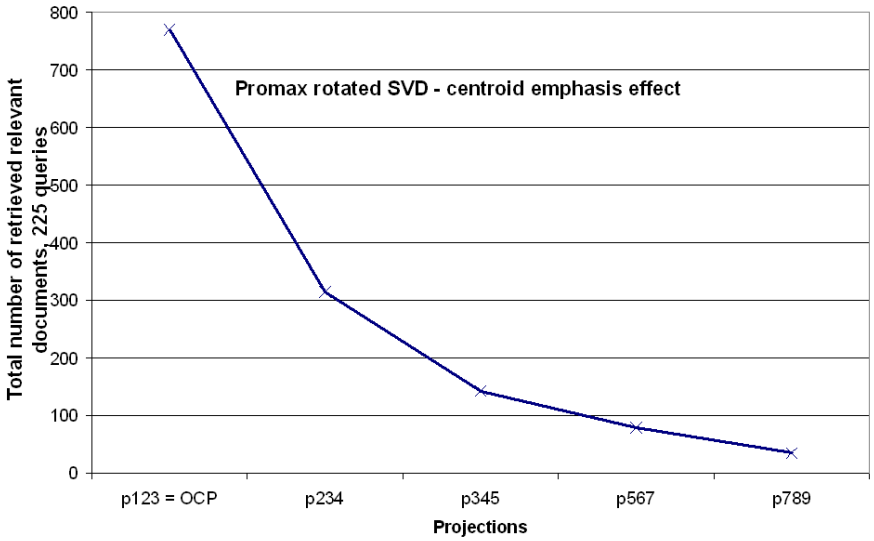


Figure 7.1: *A graphical presentation of the centroid emphasis effect for a data organization. CE is expressed in terms of the aggregate number of relevant documents retrieved for 225 queries at DCV@40 (Cranfield 1400 collection) for* $A = OCP, CP_2, CP_3, CP_5, CP_7$

Figure 7.1 presents an example of 5 distinct projections, and the aggregate number of relevant documents retrieved for each projection over 225 queries. The results may be given in more detail by presenting a recall precision curve for each projection (Figure 7.2).

In an additional example (Figure 7.3), we compare three low-dimensional CE curves to a CE curve generated by a random data organization. It shows that the performance of the SVD spaces is clearly much batter than the random performance. This strengthens our belief in the sensibility of the CE as an

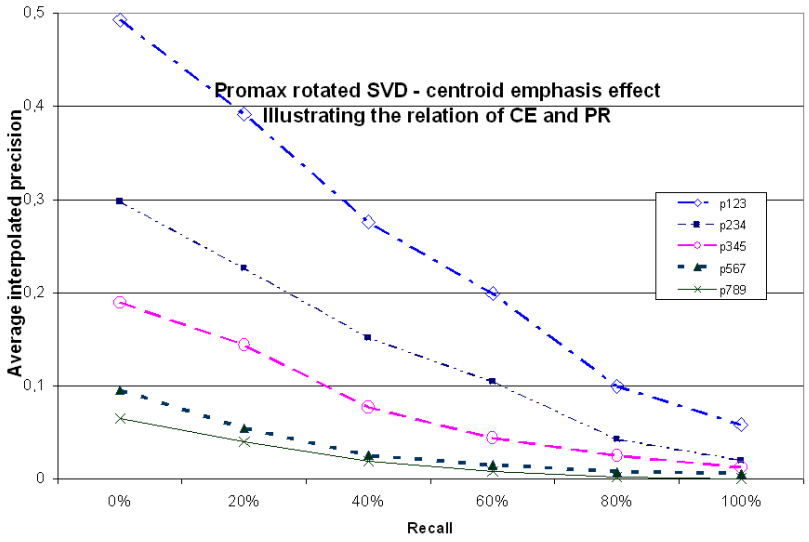


Figure 7.2: *A more detailed view of the centroid emphasis effect- PR curves for the different projections. Each point in Figure 7.1 is represented by a curve.*

indicator, as well as the sensibility of the query construction strategy based on location models.

7.2.2 Initial ranking of data organizations by the centroid emphasis

The main objective of using the centroid emphasis is to assess the suitability of a data organization for usage within an Uexküll context. It is, however possible to provide an initial indication about the ranking of different data organizations by suitability, by comparing CE curves of different organizations (see Figure 7.4). This ranking does not take into account the placement of relevant documents in the lists, nor the visual separation of relevant vs. non-relevant documents. These aspects must be catered for by other measures to be discussed in the remaining of this chapter.

7.3 Ranked list measures

We use the term "ranked list measures" as a common designator for the type of measures traditionally used in the (best-match) IR laboratory model for comparing IR methods and engines that produce linear (strictly or weakly) ranked lists (see also Chapter 2).

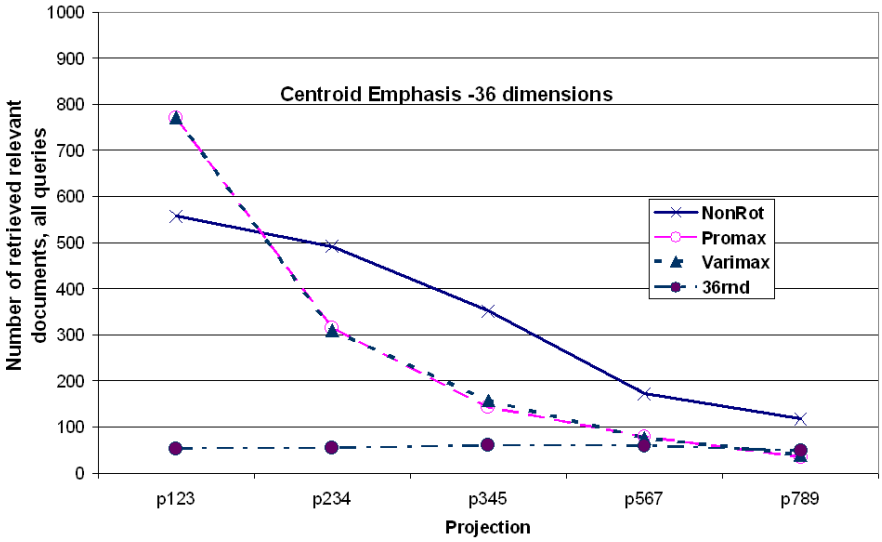


Figure 7.3: *The CE of the random curve exhibits a uniform performance over all the range of projections, as opposed to the descent in performance for inferior projections exhibited by the other, non-random curves.*

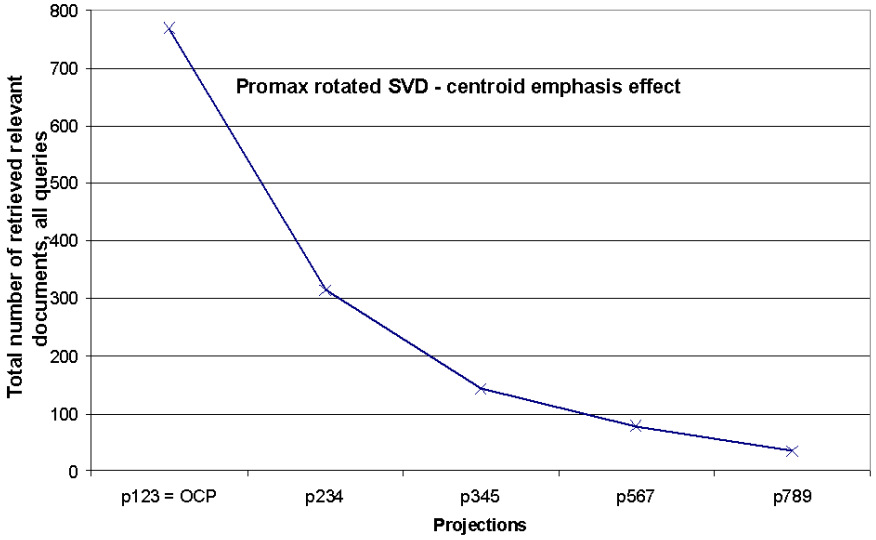


Figure 7.4: *Initial comparison of 3 organizations using CE graphical expression of Figure 7.1*

The list that the measures will be applied to is the list of documents sorted in descending order by the combined loading resulting from the simulation experiments (Chapter 6).

The most prominent measures in this category are precision and recall, but other measures, more or less based on precision and recall have also been proposed. Below we briefly review several performance measures of this type.

7.3.1 Recall and precision

The method used in Cranfield and in the main TREC experiment is that recall bases are prepared for each request (see Subsection 5.1.1). Systems' performance is evaluated by how closely their responses to queries based on those requests resemble these recall bases in terms of recall and precision. Recall and precision are defined as follows (Salton & McGill, 1983):

- $R = r/n = \dfrac{\text{the number of items retrieved and relevant}}{\text{total no. relevant in collection}}$
- $P = r/R = \dfrac{\text{the number of items retrieved and relevant}}{\text{total no. retrieved}}$

Recall and precision have been criticized in the literature for being insufficient measures of effectiveness (Saracevic & Kantor, 1988; Su, 1992; Kekäläinen & Järvelin, 2002a). Still, and particularly used in combination, they are the key measures of effectiveness within classical IR.

It was discovered quite early that high precision is correlated with low recall (subject to some qualifications elaborated by Cleverdon (1972)). The plotting of recall/precision curves has thus become an important tool of evaluation. These curves combine points of precision as a function of recall. Using such curves, the goal when optimizing retrieval performance is to "push" the curve towards higher precision figures for each recall point. Another strategy has been the computation of single valued *summary measures* that somehow summarize the performance in terms of precision and recall in a single figure. Such measures are

- **Average precision** (Buckley & Vorhees, 2000; Keen, 1992; Kekäläinen, 1999) The average of the precision scores obtained after each relevant document is retrieved, using zero as the precision of relevant documents not retrieved or at 10 (11) recall points.
- **Mean average precision** (Buckley & Vorhees, 2004) The mean over average precision readings for all queries. It is based on the averages of

the precision scores obtained after each relevant document is retrieved. It is based on much information but has been criticized in the literature for not being easy to interpret.

- **R-precision** (Baeza-Yates & Ribeiro-Neto, 1999, p. 80), is a precision figure after R documents have been retrieved, where R is the number of relevant documents in the collection judged relevant for the request in question. R-precision is indifferent to the position of the non-relevant documents among the R retrieved.

- **The Normalized Recall measure** (Salton & McGill, 1983, p. 180) contrasts the factual recall of a query against the ideal situation, where all the relevant retrieved documents are ranked ahead of all the retrieved non-relevant documents. In this way also the notion of precision is involved in the measure. As opposed to R-precision, this measure is not indifferent to the position of the non-relevant documents in the retrieved ranked list. The measure focuses on the retrieved documents only.

- **Precision@DCV, also known as Prec(λ)** calculates the precision at *document cut-off values* determined in advance. This measure models a user, interacting with a retrieval system, only willing to browse through a certain number of documents before giving up. Normal cut-off levels are 1,2,5,10,15,20,30,50,100,300,1000. (Buckley & Vorhees, 2000).

- **Generalized recall and precision** Generalized recall and precision (Kekäläinen & Järvelin, 2002b) are used to "handle varying degrees of relevance among the retrieved documents" (p. 1120). They are defined as $gR = \sum_{d \in R} r(d) / \sum_{d \in D} r(d)$ and $gP = \sum_{d \in R} r(d)/n$, respectively, where D is the set of all the documents in the collection, $R \subseteq D$ is the set of retrieved documents in response to a query, n is the size of the retrieved set and $r(d_i)$ is the relevance of document d_i in the range $0.0 - 0.1$. These measures may be used in a similar manner to their traditional equivalents, supporting evaluation against recall bases that use graded relevance judgments, where the aim is to reward systems that present highly relevant documents early in the retrieved list. In this dissertation graded relevance judgments are not being used, and therefore the measure will not be used directly. But graded relevance assessments are available for the Cranfield collection, and gR and gP, as well as the cumulated gain-based measures described in Subsection 7.3.2, may have a role in related future efforts.

7.3.2 Some related measures

The measures discussed here are special measures, not directly derived from precision/recall, that we see fit to present in this context.

- **Expected search length (ESL)** (Cooper, 1968) is a measure that expresses merit in terms of the number of *non-relevant* documents a user browses in a (generally weakly ordered) ranked list, before he or she has viewed all the required *relevant* documents. It calculates the expected number of documents viewed before the last relevant document in a given weak ranking. ESL assumes that the user knows something about the number of relevant documents (alternatively the proportion of the relevant documents among all the documents) he or she requires as a result of a search, and knows where to stop searching. This last point may be a source of critique when choosing this as a measure of efficiency in a user simulation of the kind described here. This is because we assume a user that knows little about the database and the documents in it, and would not generally have decided in advance the number of relevant documents after which to terminate the search.

- **Cumulated gain-based measures** Cumulated gain (CG) based measures (Järvelin & Kekäläinen, 2002) are a relatively recent addendum to the available set of evaluation measures[4]. Like the generalized precision and recall described in Subsection 7.3.1, these measures use graded relevance judgements done by experts, with the purpose of rewarding systems that present highly relevant documents early in the retrieved list. Besides, these measures attempt to avoid the dependence of other measures on the size of the recall bases per query in the test database, and reduce or avoid vulnerability to outliers (relevant documents appearing relatively late in the ranked result set). The CG measures presuppose a strictly ranked list, and graded relevance judgements for the documents. Some ideas in their construction may be useful when designing measures of visual support (Section 7.4).

7.3.3 Discussion of the ranked list measures

In our application the requests are used as simulation means of a slightly different kind of interaction than the one those measures were intended for. When a best-match system returns a ranked list, it is presupposed that the

[4]Some version of this family, the (n)(D)CG is among others used in the INEX initiative (Gövert, Kazai, Fuhr & Lalmas, 2003)

user will browse the documents in the sequence of this list. The response that Uexküll supports is a scene with scattered objects. Some of them are more prominent than others, and it is possible to transform them into a ranked list (which we do in the simulation scenarios). But the order by which *a user* would browse this set of documents is less predictable.[5]

This means that a measure like ESL, that presupposes such a list, and explicitly asks how many non-relevant documents the user *must* go through before encountering any relevant document, may not be very fruitful in this analysis.

Being a single figure measure obtained by only a single point on each query response, the R-precision measure has been praised in the literature for its statistical properties (Buckley & Vorhees, 2000, p. 39), which compare well with measures that are calculated as averages of several calculations pertaining to a run. An important point that may benefit our experiments is the documented stability[6] of the R-precision measure across different query sets. This property may support our results, giving reason to believe that the results, although obtained using a single collection and a single query set, are generalizable.

7.4 Measuring visualization support

7.4.1 The need for measuring visualization support

Being a 3D configuration of object, a user is not expected to conceive the downloaded scene as a ranked list. The support of the Uexküll approach for rendering relevant documents visible to the user, separating them from the non-relevant, is crucial for successful interaction with an Uexküll based system. Our evaluation should thus combine comparison of traditional counting and ranking of relevant documents with measuring the extent to which the system visually separates relevant documents from the non-relevant ones. In other words, even if the ranking of a certain query in a certain data organization is found by the ranked list measures to be perfect, we are not sure that the user will benefit fully from it, due to the locations of relevant and

[5]This means that in our context, the measures, be it recall, precision or others, are simulated measures (simulated precision, simulated recall, a. s. o.). We will still designate those "recall" or "precision", keeping the difference in mind.

[6]In Buckley and Vorhees (2000) stability was measured as the ability of a measure to maintain the same ranking of retrieval systems/methods across query sets that express the same topics in different ways.

non-relevant documents in the scene. The differences of locations may be too small for a visual system to render properly.

Organizing documents for retrieval in an approach like Uexküll is therefore more than a matter of ranking relevant document higher than non-relevant ones. The idea is to have relevant documents

- appear where the user expects to find relevant documents, i.e. high up the axes and
- appear *visually separated* from the non-relevant documents.

An ideal organization would, by our criteria, place all relevant documents far out along the axes, whereas non-relevant documents would disappear into the vicinity of the origin. This is of course not the case in real life, and both in our user simulation situation, as well as in the real search situation, the user needs to browse through non-relevant documents that are placed up the axes (even though the exact browsing sequence is not predictable).

As discussed in Section 5.3, a central concern of the current dissertation is the development of algorithmic procedures that directly measure features of usability, in this case visual support for interaction. To this end we need to be able to predict the utilization of a (hopefully good) result set, based on the special graphical mode of operation of Uexküll.

Below we are developing two such measures: separation-rewarded exposure and separation-rewarded precision. Both measures reward good ranking and good visualization support. Since it is beneficial for a continuous measure of this type to have the range of 0 to 1 for worst and best performance, respectively (Losee, 1998, p. 86), both measures subscribe to this rule.

In Figure 7.5 we have a number of documents distributed in an Uexküll group. The figure is a followup of the location model described in Chapter 6 and depicted in Figures 6.2 - 6.8, this time omitting documents below the threshold. The vertical axis represents all three axes. For each document, it represents the axis in the Uexküll group for which this document has the highest positive loading. Recollecting from Chapter 6, we have designated this the *combined axis*, and the extension of the point on this axis the *combined loading*.

Small white circles and small black circles (connected to grey patterned and brick-wall patterned rectangles), represent combined loadings of relevant and non-relevant documents, respectively.

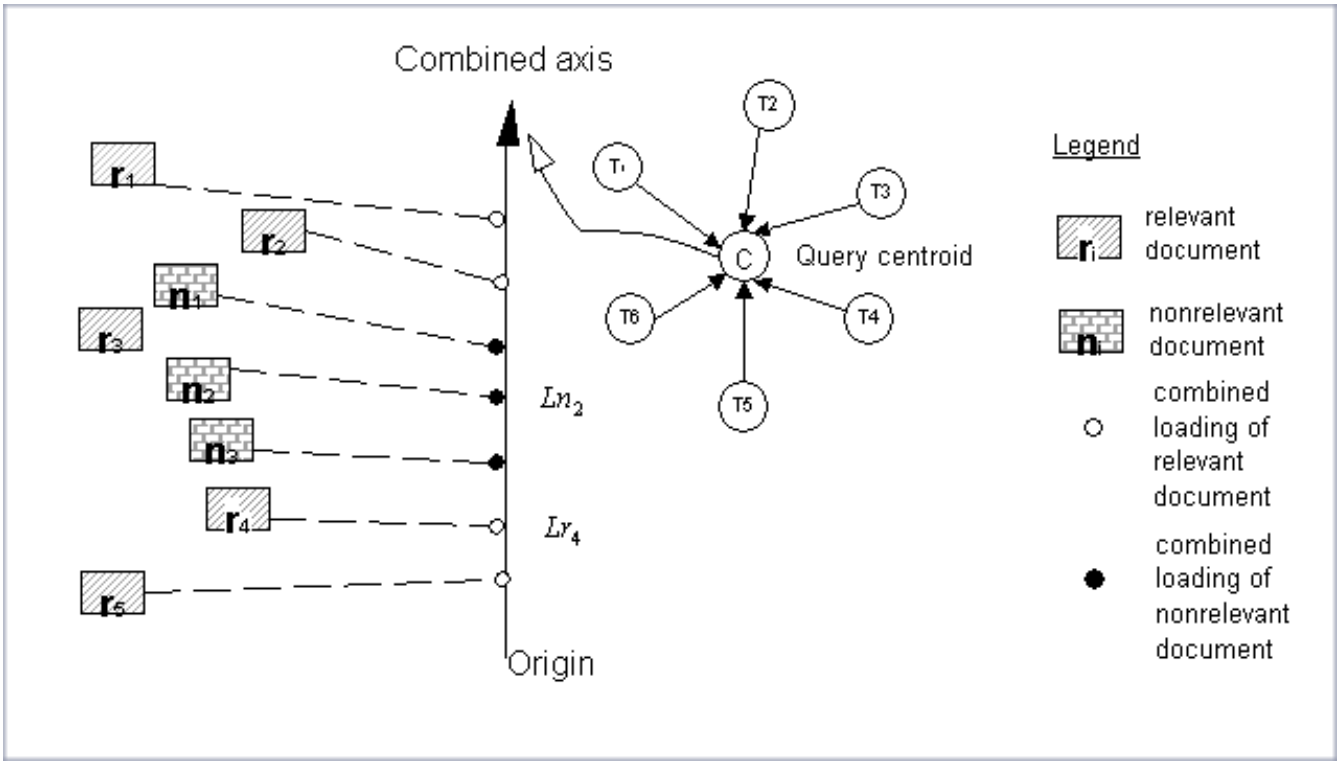


Figure 7.5: *components of the visualization support measures*

7.4.2 Measures modelling different kinds of users

Foskett (1996) discusses two types of users in search of information items (see quotation in Subsection 1.2.1). Users subscribing to those types are assumed to access a system like Uexküll differently.

1. The user in search of very few (possibly known) items will move along the axes, scrutinizing documents. Upon finding a relevant document that he considers satisfactory, this user terminates the search, without regarding other (possibly relevant) documents.

2. The user in search of a large number of relevant documents (e.g. endeavoring on a literature study, not in search of a known item) will scrutinize all relevant documents he encounters.

In order to model how these two kinds of users could benefit from the properties of a data organization, we model each of them by a measure that supports visualization:

1. Separation-rewarded exposure (introduced in Subsection 7.4.3) measures how well relevant documents are exposed from the non-relevant ones, *regardless of other relevant documents*. This measure attempts to model the type 1 user.

2. Separation-rewarded precision (Subsection 7.4.4) regards the search more like a process of collecting relevant documents, quantifying each relevant document's separation from all other documents, relevant and non-relevant, ranked below it. This measure attempts to model the type 2 user.

A discussion elaborating on the measures and their significance follows in Subsection 7.4.5.

7.4.3 Separation-rewarded exposure

As the term *exposure* implies, the measure presented below quantifies the extent to which relevant documents are *exposed*, in the sense that they are not shaded or obscured by more highly ranked non-relevant documents. A reward component (described in Subsection 7.4.3.2) modifies the measure, so that it rewards relevant documents that are far ahead of the non-relevant ones, which would increase their probability of being observed by the user.

The measure is called separation-rewarded exposure (SRE). It is calculated for each relevant document separately, and thereafter an average is taken to constitute the measure for the whole query[7].

7.4.3.1 The exposure component

Let i be a running index over all relevant documents in the scene, and j a running index over the non-relevant documents in the scene, such that r_i is the i^{th} relevant document, and n_j the j^{th} non-relevant document. L_{r_i} and L_{n_j} (see Figure 7.5) denote the respective combined loadings of these documents. For each relevant document, r_i, Let

$$\lambda_i = \{n_j | j = 1 \ldots N \wedge L_{n_j} < L_{r_i}\} \tag{7.2}$$

and

$$\epsilon_i = \{n_j | j = 1 \ldots N \wedge L_{n_j} = L_{r_i}\} \tag{7.3}$$

be the sets of non-relevant documents loaded lower than and equally as loaded as r_i, respectively[8].

To each *relevant* document, r_i, we assign a value - *exposure*, denoted EX_{r_i}, that represents the fraction of the total number of the *non-relevant* documents that *do not* shade r_i (in other words, the fraction of non-relevant documents for which $L_{n_j} < L_{r_i}$). If $|\epsilon_i|$ non-relevant documents are equally loaded as r_i, half of them, $(|\epsilon_i|/2)$, are taken to be *shading* r_i.

$$EX_{r_i} = \begin{cases} \dfrac{|\lambda_i| + |\epsilon_i|/2}{N} & \text{if } N \neq 0 \\ 1 & \text{otherwise} \end{cases} \tag{7.4}$$

If we, as an example, have $N = 10$ non-relevant documents in a scene retrieved as a query response, and for some relevant document r_i

- $|\lambda_i| = 3$ of the non-relevant documents are loaded lower than r_i
- $|\epsilon_i| = 2$ non-relevant documents are equally as loaded as r_i, and

[7] A weighting that assigns more importance to the first relevant documents retrieved could also be used.

[8] The definition of equality is, of course subject to the level of precision of representing real numbers.

- 5 non-relevant documents are loaded higher than r_i (meaning that they shade r_i)

then $|\epsilon_i|/2 = 1$, a half of the number of the equally ranked non-relevant documents, is added to the 3 lower ranked non-relevant ones, to make 4 documents, and

$$EX_{r_i} = \frac{4}{10} = 0.4.$$

EX_{r_i} equals 0 for the situation where a relevant document is behind all the non-relevant documents in the scene.

7.4.3.2 The separation reward component

We now wish to modify the exposure measure with a separation reward. We start by defining some auxiliary variables:

- T is the threshold below which the documents are excluded from the scene.
- V is the length of the visible range of the scene. V is defined as the loading of the top ranked document (relevant or non-relevant) minus the threshold.

$$V = max_{i,j}(L_{r_i}, L_{n_j}) - T. \tag{7.5}$$

- f_{r_i}, which is the fraction of the relevant document's position within the visible range V:

$$f_{r_i} = \begin{cases} \frac{L_{r_i} - T}{V} & \text{if } V \neq 0 \\ \\ undefined & \text{otherwise.} \end{cases} \tag{7.6}$$

- f_n, which is the average loading of all the non-relevant documents in the scene, again as a fraction of V

$$f_n = \begin{cases} \frac{(\frac{1}{N}\sum_{j=1}^{N} L_{n_j}) - T}{V} & \text{if } V \neq 0 \text{ and } N > 0 \\ 0 & \text{if } V \neq 0 \text{ and } N = 0 \\ undefined & \text{if } V = 0, \end{cases} \tag{7.7}$$

  where N is the number of the non-relevant documents in the scene.

For each relevant document, r_i, the *separation reward*, s_{r_i}, rewards a high coordinate value of the relevant document along the combined axis, combined with a low average value of the non-relevant documents, and is calculated as follows:

$$s_{r_i} = f_{r_i} \times (1 - f_n). \tag{7.8}$$

7.4.3.3 Separation-rewarded exposure

For each relevant document, the separation-rewarded exposure measure, SRE_{r_i} is defined as

$$SRE_{r_i} = EX_{r_i} \times s_{r_i} \tag{7.9}$$

and for a query, the measure SRE_q is taken as the average of $SRE_{r_i}, i = 1 \ldots R$ of all relevant documents,

$$SRE_q = \frac{1}{R} \sum_{i=1}^{R} SRE_{r_i}, \tag{7.10}$$

where R is the number of relevant documents in the scene.

The situation implied in Equations 7.6 and 7.7, for which $V = 0$, occurs when all documents in a scene have the same combined loading. In this situation the exposure value is 0.5, and the separation reward is not applicable. This situation is theoretic, and is an artifact of the combined axis/combined loading representation. It represents the situation where the largest loadings of the documents in the scene do not discriminate among documents.

In Table 7.1 we present two numerical examples calculating the SRE for a fictitious query, for which seven documents are retrieved, under two different data organizations. Three of the retrieved documents (number 1, 4 and 6 in ranked order) are relevant and four are non-relevant. Values for each relevant document (where applicable) are presented in the middle part, and the query-wise results are presented in the lower part of the table. The measure rewards the better separation on the right hand side with a higher value, even though the ranking on both side is equivalent.

Table 7.1: *Two examples calculating the SRE measure for a query. The named expressions follow Equations 7.2–7.10*

	non-relevant	relevant						non-relevant	relevant					
Ranking	L_{ni}	L_{ri}	Exposure	f_{ri}	Separation reward (S_{ri})	Separation rewarded exposure (SREri)	Ranking	L_{ni}	L_{ri}	Exposure	f_{ri}	Separation reward (S_{ri})	Separation rewarded exposure (SREri)
1		0,8	1	1	0,773	0,7730263	1		0,8	1	1	0,8059	0,8059211
2	0,4						2	0,3					
3	0,3						3	0,3					
4		0,2	0,5	0,211	0,163	0,0813712	4		0,2	0,5	0,211	0,1697	0,0848338
5	0,1						5	0,1					
6		0,08	0,25	0,053	0,041	0,0101714	6		0,08	0,25	0,053	0,0424	0,0106042
7	0,05						7	0,05					
	Average non-relevant loadings	Threshold (T)	Max loading	Range (V)	f_1	SREq =		Average non-relevant loadings	Threshold (T)	Max loading	Range (V)	f_1	SREq =
	0,2125	0,04	**0,8**	**0,76**	**0,227**	0,2881896		**0,1875**	0,04	**0,8**	**0,76**	**0,1941**	0,300453

7.4.4 Separation-rewarded precision

The measure discussed in this section has two components:

- The precision at every relevant document
- The separation of every relevant document from the documents ranked below it. Documents tied to the relevant document are also accounted for here.

Better overall precision will increase the quality of the expected retrieval result and better separation will give the user a better idea of the boundary between relevant and non-relevant documents. A user that starts at the outskirts of the axes will know "where to stop", and a user that starts the search at the origin (more in line with the direction metaphor) will more easily see where documents begin to be relevant.

The measure is called separation-rewarded precision (SRP). It calculates a precision component and a separation reward at each relevant document, and thereafter combines those into a composite measure. Query performance is evaluated by averaging the individual scores of the relevant documents. In the following treatment, variables designated with capital letters retain their meaning from the previous sections.

7.4.4.1 The precision component

The precision component at the i^{th} document (with $RSV_{r_i} = L_{r_i}/max_{i,j}(L_{r_i}, L_{n_j})$) takes into account all the documents with an RSV higher than RSV_{r_i}, regardless of relevance.

Let k index all documents in the scene, in descending RSV order. Let v_k=1 if the k^{th} document is judged relevant and 0 otherwise. For each document, m, the precision at it is

$$p_m = \frac{1}{m} \sum_{k=1}^{m} v_k \tag{7.11}$$

7.4.4.2 The separation reward

The separation reward for each relevant document, r_i, takes into account all documents either ranked below the current document or tie to it.

We start by defining the set $\{\alpha_i\}$ as the set of all documents, irrespective of relevance, loaded below r_i:

$$\alpha_i = \{d_k | k = 1 \ldots D \wedge L_{d_k} <= L_{r_i}\}, \tag{7.12}$$

where D is the number of documents in the scene.

We proceed by defining

$$a_i = \begin{cases} L_{r_i} - \eta & |\alpha_i| = 0 \\ \\ \frac{1}{|\alpha_i|} \sum_{r_k \in \alpha_i} L_{r_k} & \text{otherwise,} \end{cases} \tag{7.13}$$

which, for each relevant document r_i, is the average of the normalized loadings of the documents ranked below it or tied to it, or for the last document, $L_{r_i} - \eta$, where η is assigned the value of 0.0001.

The multiplicative reward component for each relevant document is defined as

$$s_{r_i} = \frac{L_{r_i} - T - (a_i - T)}{L_{r_i} - T} = \frac{L_{r_i} - a_i}{L_{r_i} - T}. \tag{7.14}$$

The reward will approach 0 for no separation and 1 for perfect separation.

7.4.4.3 Separation-rewarded precision

For each relevant document, r_i, the separation-rewarded precision is

$$SRP_{r_i} = s_{r_i} * p_{r_i}. \tag{7.15}$$

The query-score for the separation-rewarded precision is the average of the SRPs for the relevant documents:

$$SRP_q = \frac{1}{R} \sum_{i=1}^{R} SRP_{r_i}. \tag{7.16}$$

where R is, as before, the number of relevant documents in the scene.

In Table 7.2 we present two numerical examples calculating the SRP for a fictitious query, for which seven documents are retrieved, under two different data organizations. The document response is the same used in Table 7.1, and again the measure rewards the better separation on the right hand side with a higher value, even though the ranking on both side is equivalent.

Table 7.2: *Two examples of calculating the SRP measure. The named expressions are computed as prescribed in Equations 7.11-7.16.*

	non-relevant	Relevant						non-relevant	Relevant				
Ranking	L_{ni}	L_{ri}	Precision	Avgαi	Separation reward (s_{r_i})	Separation rewarded precision (SRP_{r_i})	Ranking	L_{ni}	L_{ri}	Precision	Avgαi	Separation reward (s_{r_i})	Separation rewarded precision (SRP_{r_i})
1		0,8	1	0,188	0,7246	0,724583	1		0,8	1	0,1717	0,7454	0,745417
2	0,4						2	0,3					
3	0,3						3	0,3					
4		0,2	0,5	0,077	0,5767	0,288333	4		0,2	0,5	0,0767	0,5767	0,288333
5	0,1						5	0,1					
6		0,08	0,5	0,05	0,335	0,1675	6		0,08	0,5	0,05	0,335	0,1675
7	0,05						7	0,05					
Threshold	0,04				SRPq=	0,393472	Threshold	0,04				SRPq=	0,400417

7.4.5 Summary: significance of the two measures

In line with other measures of effectiveness, SRE and SRP are based on calculating a score for each relevant document retrieved as a response to a query, and thereafter averaging the scores for the query. The two major

differences between the SRE and SRP consist in how each of them, when scrutinizing each relevant document i, relates to the other documents in the scene.

- SRE ignores other relevant documents in the scene, regarding only non-relevant documents as potentially distorting the scene for i.
- SRP, being derived from the precision notion sees the scrutinized document as a part of a ranked list, and rewards "isolation" above neighboring documents, relevant or non-relevant.

In this way, the measures may be seen as representing two types of users. A user that is in search of a factual answer or a known item, will terminate the search when such an item/a reply is found. Moreover, if he found the answer in the 3^{rd} relevant document he will not care to remember the two other relevant documents. This user is better represented by the SRE. The exposure component represents the ease of finding a known item. The query score represents the *average cost* of arriving at the sought relevant document, or the average prospects of identifying it, seen from the point of view of the current relevant document. Neither "recall" nor "precision" are considered here. Consider Equation 7.4, and suppose that for a certain query, we have a single non-relevant document only, but it is ranked above all the relevant documents. This is a situation that might conceivably give acceptable visual retrieval, but gives zero exposure = 0, and subsequently SRE=0. This situation is consistent with an extremely impatient user, not willing to look at more than a single document before getting the answer or giving up.

A user who is conducting a literature study, will probably remember, store or otherwise relate to all relevant documents retrieved. For such a user, the precision part of the SRP measure represents not only the current relevant document found, but also the record of other hitherto retrieved relevant documents. The separation part, the average of the loadings of documents ranked below the current scrutinized document represents the cost of identifying the current document.

7.4.6 Testing the measures in boundary situations

The SRE and SRP measures are intended for the evaluation of retrieval support provided by different data organization. The measures are designed to evaluate document configurations subject to the location models we have devised in Chapter 6, expressing both ranking and visual separation through a summary score. To gain better understanding of the performance of the

measures, we subject them to different types of document orderings and observe their behavior.

The measures' scores should reflect the retrieval quality for each combination of ranking and document separation the organizations attain as responses to queries. In the following we present 27 configurations of retrieved documents, for which we expect our measures to score correctly. We consider three main types of document configurations:

- Weighted ranked lists with steep descent. Provided good ranking, this represents good overall separation of relevant from non-relevant documents.
- Weighted ranked lists with moderate descent. This would only represent moderate visual separation, and therefore moderate performance although the ranking is good.
- Weighted ranked lists with gaps, representing separation between groups of documents above and below the gap. Provided good ranking, and, in addition, having relevant and non-relevant documents above and below the gap, respectively, this represents the most favorable response.

For each of these configuration types, we wish to test how the measures behave

- when the ranking of documents is ideal (all relevant documents ranked higher than the non-relevant)
- when the ranking is of moderate quality (relevant and non-relevant documents similarly distributed in the sequence)
- when the ranking is poor (all relevant documents ranked below the non-relevant ones).

This amounts to 9 types of ranked lists (below referred to as configurations).

We experiment with 3 variants of each of these 9 list types: a variant with many relevant and few non-relevant documents, a variant with few relevant and many non-relevant documents, and a variant with roughly equal numbers of relevant and non-relevant documents in the list. This sanity test gives us 27 ranked lists of combined document loadings, which we present in Tables A.1 - A.9 in Appendix A.

The procedure of the sanity test is as follows:

- We construct 9 arrangements of *coordinates*, representing the three main configurations in each of the 3 relevant/non-relevant document number relations described above.
- We then populate each of these 9 coordinate configurations with relevant and non-relevant documents, representing good, moderate and poor ranking, respectively. This gives us 27 lists.
- We calculate the measures for each of these 27 lists.

In Figures 7.6, 7.7 and 7.8 we are presenting behavior curves for the measures in the configurations that represent different numbers of documents. The figures demonstrate the sanity of the measures, in that better separation within a configuration results in a higher score of the measures. The configurations of documents are represented as category points on the X-axis. The points are designated $< descent > \_ < ranking\_quality >$, where $< descent >$ is either steep, gradual or gaps, and $< ranking\_quality >$ is either good, poor or mod (moderate).

When there are more relevant than non-relevant documents in a ranked list, or the numbers are similar, the measures rank the configurations in a conceivable manner. The gappy, steeply descending and gradually descending configurations perform best, next best and poorest, respectively. For the configurations that hold more non-relevant than relevant documents, the gappy and the steep configurations are indistinguishable in performance for the part of the SRE measure, whereas for the SRP measure the ranking of the configurations is close but distinguishable in favor of the gappy. According to the analysis in Subsection 7.4.5, the SRE measure is supposed to serve the analysis of support for users interested in few relevant documents. For such a user the gappy and the steep descent may not be significantly different in fulfilling the task of retrieving the best 1,2 or 3 documents. A user who is interested in all relevant documents will be better served by a gappy (well ordered) sub configuration. From this point of view, then, this aspect of behavior of the measure is acceptable.

It is interesting to note, in addition, that where the visual separation of documents is gradual, the SRE has a better potential for distinguishing the organization that rank documents better, even though the scores for those organizations are close. This also reflects the situation that in an Uexküll based system, good ranking does not pay off when the descent in RSV from the relevant documents to the non-relevant is very gradual.

For the SRE measure, all document configurations that have poor ranking (non-relevant documents above relevant documents) score 0. This will also

be the case for configurations that have a single non-relevant document, that scores higher than all other (relevant) documents in the list. This is the effect of the exposure component (see Subsection 7.4.3.1, particularly Equation 7.4). Even though the latter situation conceivably means good visual support, it models well an impatient user that gives up after the most visible non-relevant document.

The situation in the rightmost curve-pair of Figure 7.8, where good, medium or poor ranking in a gradual configuration is hardly distinguishable, expresses how difficult it would be to visually distinguish relevant from non-relevant documents in such a configuration, even though the ranking is (by standards of ranked lists) significantly better at the point marked *gradual_good* then at the point marked *gradual_poor*. The difference in visuality, being very little, is well expressed by the little difference in the scores of both measures[9].

7.5 Additional measures and statistics

7.5.1 Limitation of the hitherto discussed measures

The measures discussed so far in this chapter seek to utilize the high number and the diversity of the queries/topics of the text collection at our disposal, in order to characterize the potential usability of the Uexküll approach, by computing averages of different kinds of query performances across the entire set of queries/topics.

Albeit the novelty of the approach, and the novel elements of this evaluation, seeking to evaluate elements of usability, the evaluation so far follows the traditional way of evaluating IR results, averaging performance across queries/topics. (This is referred to as collection-level evaluation)

Nevertheless, evaluating an approach like Uexküll also requires that we regard performance as more than some average responses over a number of predefined topics. This section will address additional/alternative use of the simulation results, trying to take a closer look at the usability potential of the Uexküll approach.

[9]This point raises some interesting usability issues to be followed up by system design. It is conceivable that situations where chunks of documents have very similar loadings on specific axes could be handled by graphical interference in real time. For example if the system senses chunks of documents that are very close to each other, the system could separate them graphically, retaining their mutual ranking, but making them more distinguishable. Such interference could, of course result in wrong ranking from the point of view of the user, and this possibility is not taken into account in the measures. The measures are, in this sense, conservative.

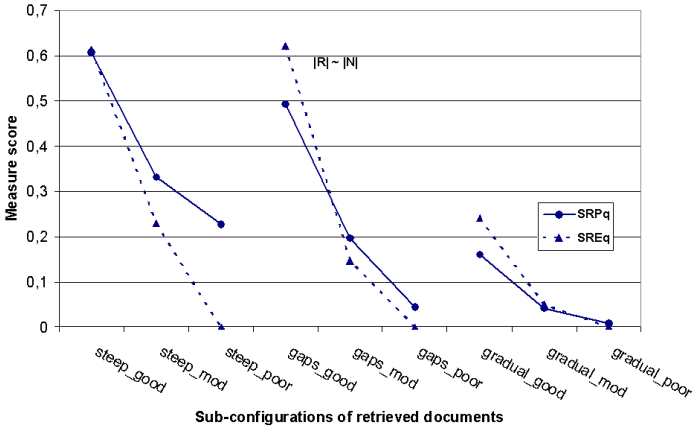


Figure 7.6: *Behavior of SRP_q and SRE_q when roughly equal numbers of relevant and non-relevant documents are retrieved. 9 different configurations.*

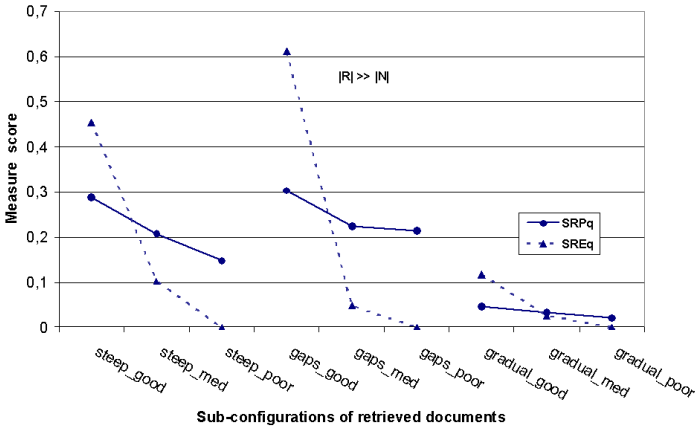


Figure 7.7: *Behavior of SRP_q and SRE_q when more relevant than non-relevant documents are retrieved. 9 different configurations.*

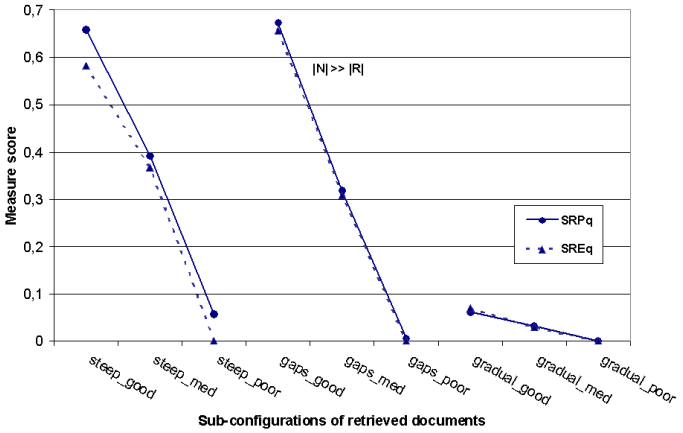


Figure 7.8: *Behavior of SRP_q and SRE_q when more non-relevant than relevant documents are retrieved. 9 different configurations.*

7.5.2 Numbers of queries attaining prescribed levels of recall

A system based on the Uexküll approach will, by design, not offer a ranked list of documents as a response. Retrieving many documents that do not have a natural strict ranked order can impose a high cognitive load on the user. It is therefore necessary to limit the number of documents retrieved, probably causing some potentially interesting documents not to be retrieved. Translated to our simulation situation, for each query that represents a user interest, we retrieve a limited number of documents. With this limitation, not all queries will attain full recall.

We can account for this situation, at the collection level, by observing how many queries attain *prescribed levels* of recall. In addition to being a measure of retrieval effectiveness, in line with PR-curves and other measures, this gauge has an independent usability implication, and may indicate the levels of recall that can be anticipated in a user interaction situation.

7.5.3 Diversity of access: axis usage statistics

A system based on the Uexküll approach presents a list of concepts to the user, from which he is supposed to choose three that best represent his information need, in order to download an Uexküll group. It is one and the same list of concepts that is presented to different users with different interests, and we expect users with different interests to use different concepts as entry points.

Suppose that the *queries* in our test collection represent a wide range of information interests that can be answered by the documents in the collection. Each query is, in the simulations, implemented by selecting the triplet of axes on which its centroid loads highest. A good data organization would support *diversity of access*, different queries would be implemented by different triplets of concepts. In a poor organization, a large percentage of the query centroids would be highly loaded on very few axes (for example the axes corresponding to the largest singular values in an SVD organization) and the other axes would have little or no contents associated with them.

To investigate the diversity aspect of organization quality, we will present curves of distribution of axes among queries, and find out whether all the queries are clustered around very few axes (poor organization) or are many axes distributed uniformly among queries (good organization). Knowing that fewer axes than the dimensionality entails are interpretable, these curves

also indicates the number of interpretable axes we actually get from our organizations.

7.6 Significance of performance differences among data organizations

The experimental design used for the evaluation is a block/treatment design. For each summary measure, results collected for all queries in different organizations are ordered in a matrix of b blocks $\times$ k treatments, where the queries are the blocks (rows) and the organizations are the treatments (columns). Based on each such matrix we can run both ANOVA (a parametric test) and a non-parametric test like the Friedman test (Conover, 1980; Hull, 1993; Kekäläinen, 1999; Järvelin & Kekäläinen, 2002).

Hull (1993) has advocated the use of both parametric and non-parametric tests to assess the significance of retrieval experiment results. For the above described setting (provided that the parameters of data permit), parametric tests, (like *analysis of variance*, also known as ANOVA)) are often preferred, as their use is based on the values of the *data*, so that measured magnitudes and their relations are preserved.

Non-parametric rank tests (e.g. the Friedman test (Conover, 1980)) used in this setting, base their inference on the *rank order* of the measurements, so that the measured magnitudes are lost, and are not used in the significance inference. This fact renders such tests weaker than their parametric counterparts. As suggested by the adjective "non-parametric", these tests do not assume any particular parameters governing the distribution of the data (apart from assuming that errors are independent and come from the same continuous distribution). Using ranks instead of absolute scores, such test also normalize for differences in scale. The trade-off is a lesser inferential power provided by these tests.

Our experiments are comparing different data organizations, which can be viewed as different retrieval systems. Since retrieval data are generally not normally distributed[10], and the distribution of the measures developed in Section 7.4 across the methods/queries is not expected to be normal either, a non-parametric significance test is required (Hull, 1993; Keen, 1992).

[10]The large number of observations could counterbalance the non-normality, but ANOVA, normally used for multiple comparisons in a normal setting, has also the property that queries with great variability in performance among systems influence the results of the test, an effect that is automatically normalized for by the Friedman Test.

The non-parametric Friedman test (Conover, 1980; Kekäläinen, 1999) was chosen to this end, as it supports multiple comparisons. It allows us to test whether (1) the treatments (data organizations) are significantly different from each other, and, in the positive case, (2) control the significance of their ranking pairwise.

The formula that we are using for the Friedman test statistic is taken from Hull (1993) and Kekäläinen (1999), and follows below:

$$F_c = \frac{(b-1)(B - bk(k+1)^2/4)}{A - B}, \tag{7.17}$$

where b is the number of queries, k the number of data organizations,

$$A = \sum_{i=1}^{b} \sum_{j=1}^{k} (R(X_{ij}))^2 \tag{7.18}$$

and

$$B = \frac{1}{b} \sum_{j=1}^{k} {R_j}^2. \tag{7.19}$$

The null hypothesis is that the treatments (data organizations) come from the same population, and we reject it if F_c (Equation 7.17) exceeds the $1-\alpha$ quantile of the F distribution for the applicable number of treatments/number of blocks.

$R(X_{ij})$ is the rank of the cell in the i^{th} row and j^{th} column, and R_j is the sum of ranks in the j^{th} column. When converting the raw data into ranks, average ranks are assigned to tied values. For example if two items share the 3^{rd} smallest value, the ranks given will be 1, 2, 3.5, 3.5, 5, because the two items "share" the 3th and 4th rank, and their rank is set to $(3+4)/2 = 3.5$ each.

Upon getting a significant indication that the treatment column data do not come from the same population, i.e that there are differences within the treatments, the following procedure is adopted: The rank sums of each two treatments, are compared (subtracted from each other) and the absolute value of their difference compared with the expression

$$t_{1-\alpha/2} \left[\frac{2b(A_2 - B_2)}{(b-1)(k-1)} \right]^{1/2}, \tag{7.20}$$

using Student's t-test, $t_{1-\alpha/2}$, where α is the significance level,

$$A_2 = \sum_{i=1}^{b} \sum_{j=1}^{k} \left[R(X_{ij})\right]^2$$

and

$$B_2 = \frac{1}{b} \sum_{j=1}^{k} \left[R(X_j)\right]^2 .$$

Chapter 8

Experiments and results

8.1 Introduction

8.1.1 Presenting retrieval results

This chapter summarizes and analyzes the experiments and the experimental results. We start with a brief discussion about presentation of IR results in general, and the particularities regarding the present project. Thereafter we present the experimental procedure and the results.

Presentation of retrieval results is discussed by Salton and McGill (1983), Keen (1992), Tague-Sutcliffe (1992), and others. When following the traditional laboratory model, the normal practice has been to plot precision vs. recall curves of different systems, methods or other settings of retrieval that are being compared. At times it is also necessary to characterize a system by some summary measures. In Chapter 7, we have referred to some measures that are derived from the precision/recall notions, choosing the R-precision measure as a summary measure to characterize the ranking qualities of our data organizations. In addition we have presented two other measures, developed in Section 7.4.3 and Section 7.4.4. These will be used in line with the R-precision measure, in order to extract and compare visualization properties of the different data organizations.

In the center of the analysis/presentation lies the *axis interpretability* of the data organization. Reiterating from Chapter 1, interpretability is defined as the extent to which that organization associates documents and index terms with named axes to which those terms and documents are relevant. More concretely, interpretability will be characterized in terms of the ability of a data organization to bring documents that are relevant to an information need to prominent positions along coordinate axes. These axes will be chosen

by users to represent that same information need. Axis interpretability is not expressed by a number, but is an aggregate characteristics based on all the applied measures.

8.1.2 Terminology

For convenience we briefly reiterate some of the terminology used in this dissertation, which is applicable to this chapter.

Decomposition or multivariate decomposition: an algorithmic treatment of a term-document matrix, resulting in a raw, arbitrarily rotated vector space with reduced dimensionality. We use the word also to designate the *vector space* resulting from a decomposition.

Rotation Changing the orientation of all the vectors in a vector space in relation to the axes, following some criterion or criteria (see also Subsection 4.2.2).

Data organization The space resulting from a decomposition, possibly rotated.

Combined axis The representative axis of a scene, embodying the transformation of the 3D layout of the document into a ranked list (see Section 6.3).

8.2 Outline of the experiments and presentation of results

8.2.1 Outline of the experiments

Our experiments contain two main aspects:

- The evaluation of the suitability of the data organizations, that we use to the task of visualization. To this end we use the centroid emphasis developed in section 7.2. These tests already provide us with an initial indication of the relative quality of the different organizations.

- The comparison of suitable organizations. This aspect is subdivided into three sub-aspects:

Ranking: How documents judged relevant to a query, as well as other documents, are ranked within the Uexküll group axes (represented by the respective combined axes (see Figure 6.8)) selected for that query. This we cater for using the ranked list measures.

Visual support: How well the visual separation of documents judged relevant and other documents is catered for. For this we use the newly developed SRP and SRE measures.

Diversity of access How different concepts are participating in the query answering. This is done using descriptive summary statistics, counting concepts that participate in query answering, summarizing them using curves. This is done using simple distribution charts.

8.2.2 Procedure

We have experimented with two versions of the Cranfield collection, both indexed by extraction: For one version (below referred to as the automatically indexed version) the abstracts are indexed automatically and for the other version the terms have been extracted from the full texts of the documents and post processed manually (see Subsection 5.4.3). The main line of experimentation uses the automatically indexed version, and some results are also presented for the manually indexed version.

We start by decomposing the term-document matrix in 4 different numbers of dimensions[1]. We rotate each of the decompositions both obliquely and orthogonally, giving us three organizations, resp. non-rotated, orthogonally rotated (varimax) and obliquely rotated (promax), for each number of dimensions. Coordinate values of terms and documents for each of these organizations are stored in a coordinate database, allowing us to manipulate and interact with them: querying spatial locations of terms and documents, calculating centroids and other operations.

Simulation scenarios 1 and 2 (see Subsections 8.2.2.1 and 8.2.2.2 below) are implemented by scripting against the coordinate database. Each run of a scenario, applied to some data organization, basically generates a list of documents as a response to each query. Each document in the list is assigned an RSV based on its coordinate value on the combined axis. Document lists, with RSVs, are stored in the database. Comparison with the recall bases for each query (also stored in the database) allows us to run evaluation programs

[1]The large quantities of data necessary fort each experiment prohibit the use of a larger numbers of dimensions

and calculate the measure scores for the queries. For each measure we use the average over all the queries applied to each organization (organization average) to obtain the summary measure score. We save also the scores for individual queries for the calculation of significance statistics, as described in Section 7.6.

8.2.2.1 Simulation scenario 1

Recalling from Subsection 6.4.2, for each request, simulation scenario 1 uses a single Uexküll group to represent the entire user interaction. The procedure is implemented using direct interaction with the coordinate database. For each query we do the following:

- calculate the centroid vector
- extract the three axes with the highest coordinate values, giving us the Uexküll group
- extract the combined loadings of the documents on the combined axis, retrieving the 40 highest loaded ones
- evaluate this ranked list against the recall base of the query.

8.2.2.2 Simulation scenario 2

Recalling from Subsection 6.4.3, simulation scenario 2 is an attempt at approaching the simulation of a process. It starts with simulation scenario 1, giving us the initial Uexküll group and a list of documents. From there, a sequence of Uexküll groups is generated by iterating through the terms of the request: For each term, the "weakest axis" of the current Uexküll group (the axis on which the term has the 3rd largest coordinate value in this data organization) is substituted by the axis on which the term loads highest, giving us a new Uexküll group, for which a new list of documents is retrieved. The new document list and the current one are merged, leaving the 40 relatively highest loaded documents from both lists, discarding the rest. This last list becomes the current list, and the process is repeated until the query term list is exhausted. The final list is evaluated against the recall base of the query.

8.2.2.3 Centroid emphasis

To calculate the centroid emphasis we run simulation scenario 1 with the following modification: For each query, in addition to the Uexküll group composed of the three highest loaded axes of the centroid vector, we iterate

through projections (Uexküll groups) composed of a number of supposedly inferior combinations of axes. For each of those inferior projections we collect the aggregate number of relevant documents retrieved for all queries, as well as average precision recall data. This gives us the possibility, for each data organization, to compare the performance of those inferior projection to the performance of the OCP, and the possibility to compute the centroid emphasis measure.

8.2.2.4 Distribution of concepts

The number of distinct concepts that participate in responding to all our queries is an additional gauge of usability, indicating the ability of the data organization to provide entry points for users with different information interests. See a discussion in Subsection 7.5.3. We present a distribution curve of counts of occurrences (usage) of the concepts in the queries. Many concepts, uniformly distributed over the queries would signify a good diversity of the organization in question.

8.2.3 Outline of the result presentation

Our presentation of results consists of three components:

- Presentation of suitability results (centroid emphasis), using point plots, as introduced in Figure 7.1.
- Presentation of the results of the ranking properties of the data organizations. Ranking properties are presented using several techniques, listed below:
  - curves of average interpolated precision at recall points
  - curves of precision at a number of DCV points
  - curves displaying (for each organization) the number of requests (out of 225) for which certain values of recall (0%, 20%, 40%, 60%, 80% and 100%) are attained at DCV@40
  - Tables of ranking-quality differences, with indication of significance based on the R-precision measure, using the non-parametric Friedman Test.
- Presentation of the separation properties (visualization support) of the data organizations. Here we use the two measures of separation developed and discussed in Chapter 7. For each of these measures we

conduct a non-parametric significance test, deriving the significance of the differences in performance among the organizations with respect to this measure.

- Presentation of the distribution of concept axes among queries.

All the results are produced and presented for both scenarios, apart from the CE and the concept axis distribution calculated only within simulation scenario 1.

When producing and presenting significance statistics of any summary measure, the following procedure was followed:

- The measure score is calculated for each query in each data organization.
- Based on the measure scores, the rank of each organization for each query is extracted, following the procedure outlined in Section 7.6.
- If the null hypothesis (no significant differences in performance among organizations) is rejected ($F_c > m$)[2], the ranks for every organization are summed across all queries, and the differences between any two rank sums are compared to the critical value (Equation 7.20).
- The organizations are then tabulated for presentation, sorted by their rank sums (see for example Table 8.1 below, where the table entries are differences of rank-sums). The organization names[3] are listed in the row and column headers in descending order of their rank sums (which is not necessarily equal to the order of the average scores themselves, which are given in the second row but are not used directly in the test). Any two organizations are deemed significantly different if their absolute difference exceeds the critical value from Equation 7.20 (entries in bold face mark significant difference). Table 8.2 is a summary of Table 8.1.

8.2.4 Using the versions of the collection

We conduct our experiments and presentations mainly for the automatically indexed version of our test collection, but we also include some results obtained for a manually indexed version of the same collection[4]. Assuming that

[2]For our data (225 queries over 12 organizations), $m \sim 2$

[3]**V** stands for varimax. **P** for promax and **N** for non-rotated

[4]We deliberately refrain from using the term "intellectual indexing" here, as the indexing in this case was done by extraction, but the procedure resulted in something that is

Table 8.1: *An example of presenting results for summary measures: The organizations are displayed sorted by their sum of ranks. Rows and column headers, and absolute differences, appear as table entries. Note that average scores (second row) sort differently than the ranks. This is due to the Friedman Test being a rank test. Significant differences appear in bold face. The test statistics (24,14) and the critical value for pairwise comparison (134,2) appear on the leftmost column, second and third row, respectively. Tables of this kind are displayed in Appendix B*

org	rp1d	V158	V309	P158	V75	P75	V36	P36	P309	N75	N309	N158	N36
24,14	Mean result	0,262	0,245	0,228	0,207	0,186	0,144	0,142	0,148	0,129	0,126	0,123	0,11
> 134,2	Σ ranks	1823	1786	1656	1594	1493	1420	1387	1360	1281	1269	1261	1224
V158	1823	0	37	***167***	***229***	***330***	***403***	***436***	***463***	***542***	***554***	***562***	***599***
V309	1786		0	130	***192***	***293***	***366***	***399***	***426***	***505***	***517***	***525***	***562***
P158	1656			0	61,5	***163***	***236***	***269***	***296***	***375***	***387***	***395***	***432***
V75	1594				0	101,5	***175***	***208***	***234***	***314***	***325***	***333***	***371***
P75	1493					0	73	106	132,5	***212***	***224***	***232***	***269***
V36	1420						0	33	59,5	***139***	***151***	***159***	***196***
P36	1387							0	26,5	106	117,5	125,5	***163***
P309	1360								0	79,5	91	99	***137***
N75	1281									0	11,5	19,5	57
N309	1269										0	8	45,5
N158	1261											0	37,5
N36	1224												0

Table 8.2: *A summary of Table 8.1. The third to last rows specify the significant differences in relation to the entries in the column headers. "∼" designates "no significant difference" and ">" designates "significantly better than ($\alpha = 0,025$)".*

It is a table of this kind that summarizes each scenario for each measure

rp1d	V158	V309	P158	V75	P75	V36	P36	P309	N75	N309	N158	N36
Mean results	0,262	0,245	0,228	0,207	0,186	0,144	0,142	0,148	0,129	0,126	0,123	0,11
V158		~	>	>	>	>	>	>	>	>	>	>
V309			~	>	>	>	>	>	>	>	>	>
P158				~	>	>	>	>	>	>	>	>
V75					~	>	>	>	>	>	>	>
P75						~	~	~	>	>	>	>
V36							~	~	>	>	>	>
P36								~	~	~	~	>
P309									~	~	~	>

the term names of the manually indexed databases would provide better axis names than the normalized vocabulary inherent to automatic indexing, we would like to get an indication of the axis interpretability provided by decomposing this kind of indexing.

8.3 Results for the automatically indexed version

In this section we present results for the automatically indexed version, starting with the centroid emphasis, proceeding with simulation scenario 1 and simulation scenario 2. For each simulation scenario we first present all the ranked list results, followed by the visual support results.

8.3.1 Centroid emphasis

In Table 8.3 we present the aggregate numbers of relevant documents retrieved (among 40 documents for each query) for 225 queries (Cranfield collection) for various projections of different data organizations, resulting from applying different rotations to SVD of different dimensionalities. For every data organization we provide the metrics of the CE for this organization, using Equation 7.1.

The results of the centroid emphasis show two main features:

- higher score for lower dimensionality
- higher score for rotated organizations than for non-rotated

The numbers are not easy to interpret, but it is obvious that for low-dimensional organizations the inferior axes for every query lose more interpretation power relative to the OCP axes, rendering their OCP-figure higher. For the higher-dimensional organizations, the secondary axes for the query preserve some of their power to attract relevant documents. This renders their CE-figure smaller

If we follow the results from left to right, we see that the organizations gain interpretation power as they are rotated, and the low-dimensional gain more than the high-dimensional in relative power, even though the absolute

closer to an intellectually created index, with counts as weights. We use a dichotomously weighted version of this index to try to approach the predicted behavior of an intellectually indexed collection.

Table 8.3: *The aggregate numbers of documents retrieved for the different projections of the data organizations, with the centroid emphasis scores*

sumc		NonRotated	Promax	Varimax
	1,2,3(OCP)	557	770	772
	2,3,4	492	315	310
36	**3,4,5**	352	143	157
	5,6,7	172	79	76
	7,8,9	118	35	40
	CE	0,491	0,814	0,811
	1,2,3	587	854	894
	2,3,4	559	504	428
75	**3,4,5**	432	300	235
	5,6,7	324	144	117
	7,8,9	235	72	72
	CE	0,340	0,701	0,762
	1,2,3	579	932	937
	2,3,4	554	664	570
158	**3,4,5**	437	477	429
	5,6,7	306	231	234
	7,8,9	249	160	165
	CE	0,332	0,589	0,627
	1,2,3	569	644	863
	2,3,4	552	468	623
309	**3,4,5**	432	318	507
	5,6,7	287	221	298
	7,8,9	249	168	234
	CE	0,332	0,544	0,519

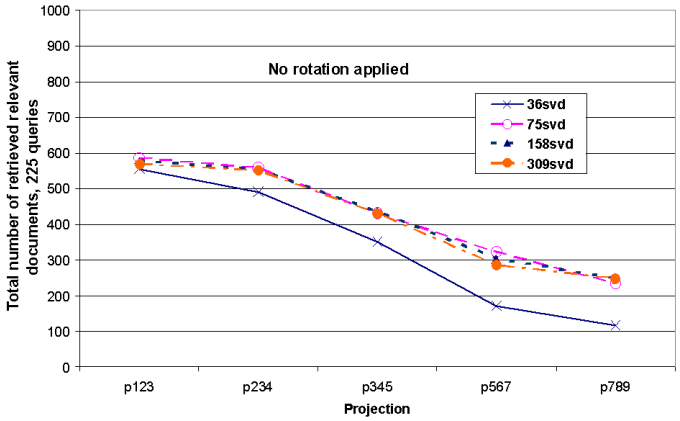


Figure 8.1: *Aggregate numbers of relevant documents retrieved for five projections, four dimensionalities, non-rotated*

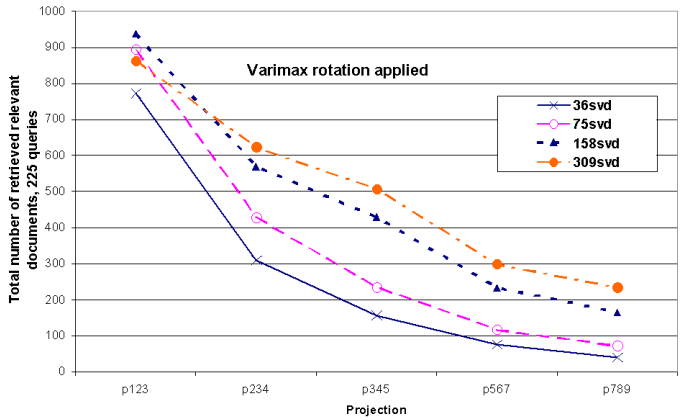


Figure 8.2: *Aggregate numbers of relevant document retrieved for five projections, four dimensionalities, varimax-rotated*

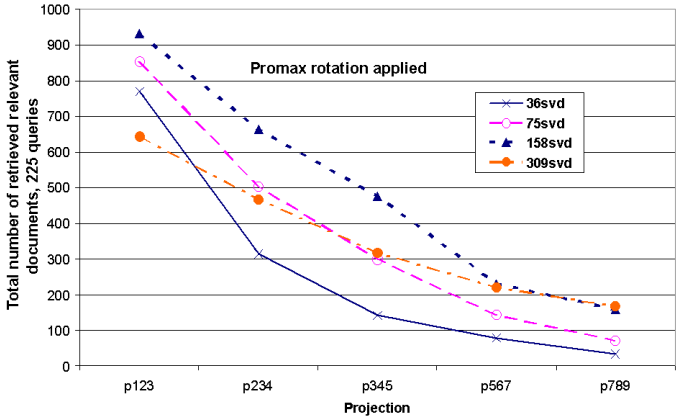
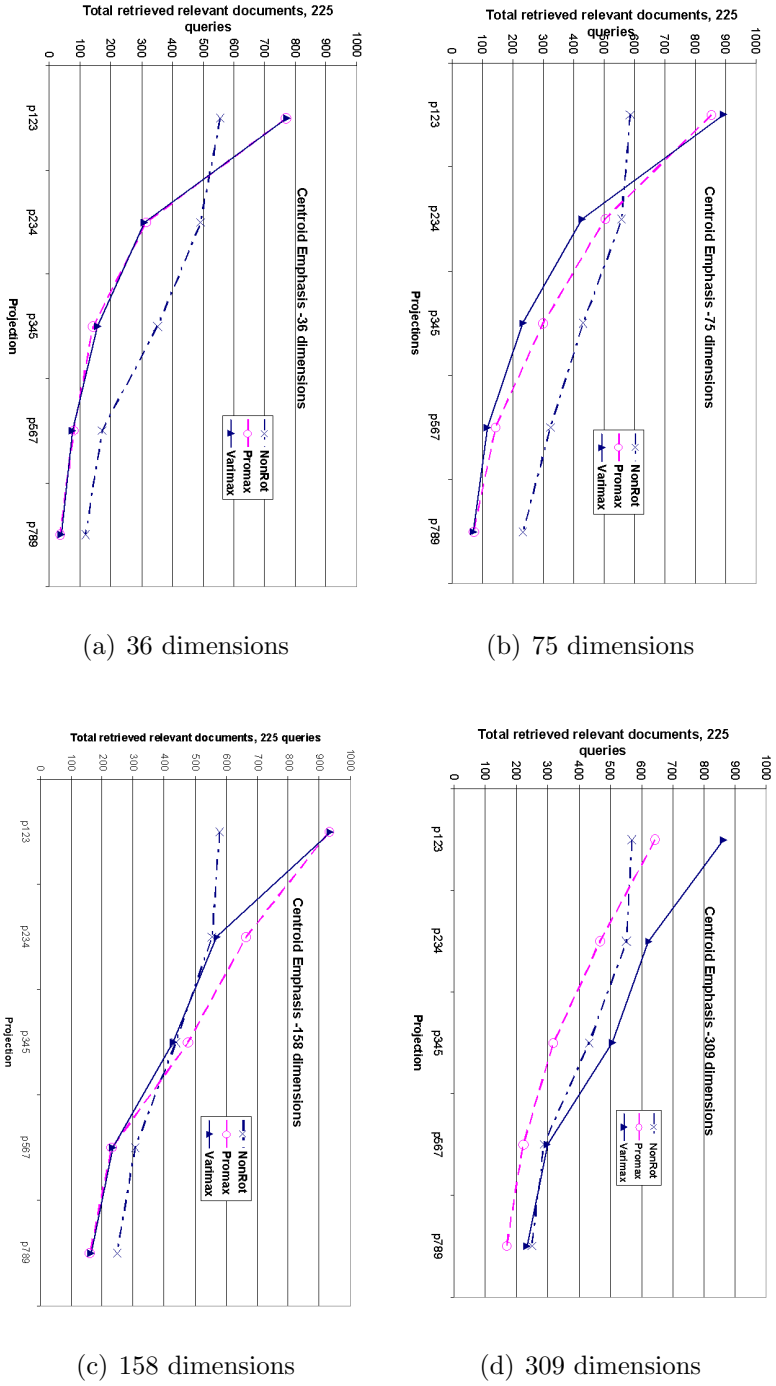


Figure 8.3: *Aggregate numbers of relevant document retrieved for five projections, four dimensionalities, promax-rotated*



(a) 36 dimensions

(b) 75 dimensions

(c) 158 dimensions

(d) 309 dimensions

Figure 8.4: *The curves from Figures 8.1 - 8.3 ordered by dimensionality*

numbers of documents retrieved are higher for the high-dimensional, rotated organizations.

Figures 8.1 through 8.3 and Figures 8.4(a) - 8.4(d) are a graphical presentation of the columns and rows of the subtables of Table 8.3, respectively. They are summary plots, where each of the count entries in Table 8.3 is represented by a point. The lines connecting the (genuinely discrete) points on the curves are only meant for graphical clarity, as the difference between the various rotations manifests itself better this way. The lines do not represent any meaningful interpolation.

Even though the number of relevant documents retrieved is greater for decompositions with higher dimensionality, the lower dimensionality organizations feature the steepest CE curves. The differences among the OCP and the inferior projections, both in steepness and the numbers of relevant documents retrieved, are consistently greater for the rotated than for the non-rotated organizations.

The descent of the promax-rotated organization is consistently more gradual than for the varimax organization. It is not easy to explain this, as the "chain of events" leading to this is rather complex, and involves both the combination of loadings of the 3 axes constituting a projection, as well as the "magnification" of the differences between high and low loadings that promax superimposes on the varimax rotation, and the averaging of scores for a high number of queries. The analysis of this effect may involve the observation of individual queries that score high with varimax and low with promax, respectively, and observing how documents judged relevant and documents judged non-relevant are lined up on the axes constituting the projections. This analysis is considered beyond the scope of this dissertation.

With all due care about generalizing from one set of results, Figures 8.1 - 8.3 seem to render the SVD with the rotations applied principally suitable to be used for an Uexküll - based visualization. As could be expected, both oblique and orthogonal rotations render the centroid emphasis effect stronger than the non-rotated organizations.

Even though we are quantifying the CE, we are not making any statistical test of the significance of the differences here, and therefore no inference. But as the most important role of the centroid emphasis, the presented data organizations pass the initial test of suitability. The CE also supports our strategy of representing queries through the optimal centroid projection of the query terms.

8.3.2 Results for simulation scenario 1: ranked list measures

The following subsections present results for measurements of ranked list measures, representing traditional IR-evaluation. The curves are presented grouped by the type of rotation applied, each chart applying to all tested dimensionalities. The charts feature normal precision/recall curves, as well as average precision at a number of DCV points. In addition we present, for each organization, a summary table of the R-precision results, trying to isolate the statistically significant differences in performance among organizations.

8.3.2.1 Average interpolated precision

In Figures 8.5 - 8.14 results for different decompositions are displayed, with the dimensionality as the immediate comparison criterion. The charts in the first group, Figures 8.5 - 8.7, show average interpolated precision-recall plots for three different rotations. We note that for the non-rotated (or arbitrarily rotated) SVD space there is little gain in performance using more than 36 dimensions, and no gain in performance using more than 75 dimensions. Figure 8.8 shows the same curves, grouped by dimensionality.

Judging by the curves, the obliquely as well as orthogonally rotated versions each provide a noticeable improvement in interpretability over the non-rotated version.

At this point it is worth reminding that the interpolation of the curves towards 100% recall is done on the basis of retrieving a constant number of documents (40) for each query, and are therefore based on very few queries ever reaching 100% recall. To correct for this we also present curves of precision at DCV points, as well as curves for numbers of queries attaining different values of recall.

An interesting trait regarding the promax rotation used here (Figure 8.7) is that it renders the lowest dimensional decomposition roughly the same as the varimax rotation (Figure 8.6) does, but seems to deteriorate the performance of the highest dimensionality (309 dimensions). One effect of the promax rotation is that it amplifies differences. A part of the promax treatment is the raising of loadings of the varimax results to a power of 3 or 4, thereby enhancing the difference between highly loaded and lower loaded documents. We suspect that 309 dimensions is an excessive dimensionality for our collection, and that retrieval noise due to it may be enhanced by the promax algorithm.

mm

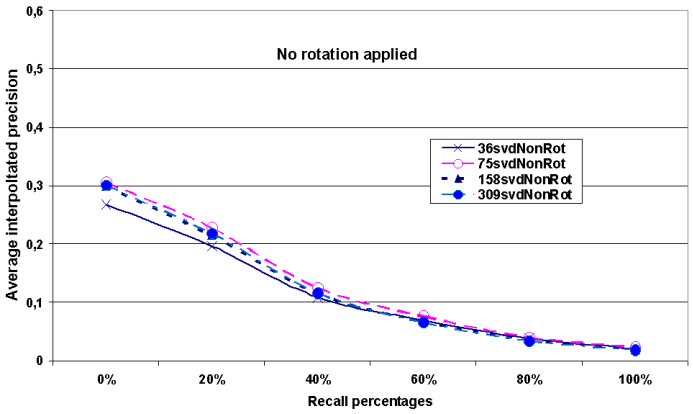


Figure 8.5: *Simulation scenario 1: average interpolated precision at recall percentage points for non-rotated SVD*

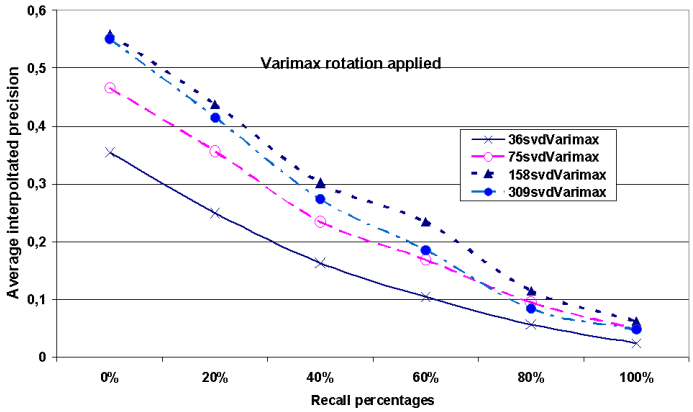


Figure 8.6: *Simulation scenario 1: average interpolated precision at recall percentage points for orthogonally rotated SVD*

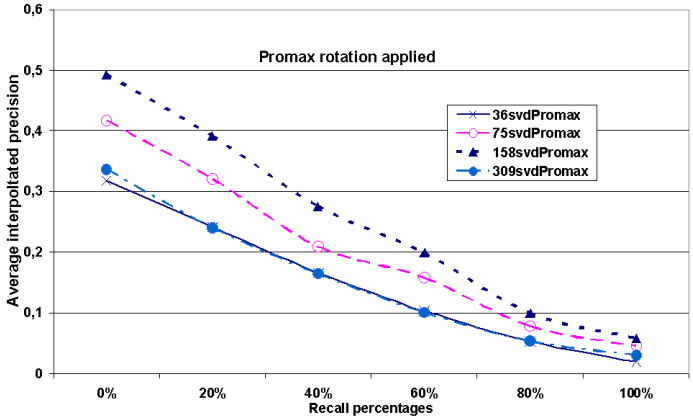
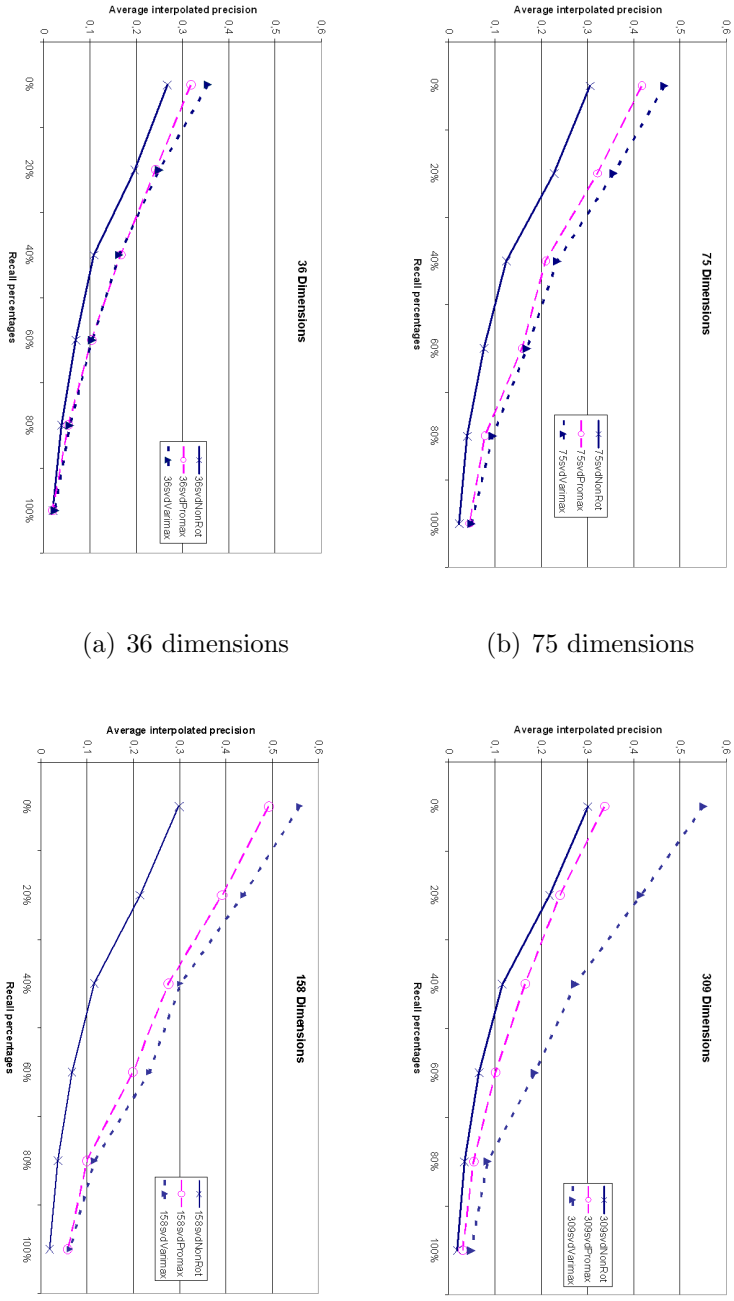


Figure 8.7: *Simulation scenario 1: average interpolated precision at recall percentage points for obliquely rotated SVD*



(a) 36 dimensions

(b) 75 dimensions

(c) 158 dimensions

(d) 309 dimensions

Figure 8.8: *The curves from Figures 8.5 - 8.7 ordered by dimensionality*

8.3.2.2 Precision@DCV

Figures 8.9 - 8.11 show curves of average recall and precision at constant DCVs of 5, 10, 15, 20, 25, 30, 35 and 40 documents. The recall axis is used so that the DCV points are aligned with the average recall the organization reached at the corresponding DCVs. As an example, in Figure 8.10 the organization *158svdVarimax* has reached an average recall of 57%, combined with 10% precision at 40 documents. On the average, when 5 documents have been retrieved for the varimax-rotated 158 dimensional organization, 30% precision combined with 22% recall have been attained. Note that no interpolation procedure (Baeza-Yates & Ribeiro-Neto, 1999, p.77) is used in these figures.

Tendencies from the previous plots are confirmed in these plots. We note that differences in performance are hardly detectable for the non-rotated organizations.

Comparing Figures 8.10 and 8.11, we see that the performance of the obliquely rotated and orthogonally rotated organization is comparable at the high recall end of the plot, while the orthogonally rotated organization performs seemingly better at the low recall end. This may indicate that a fact oriented search (the SRE user, see Subsection 7.4.5) has better chances of coming up with answers for the varimax-rotated organization, whereas for more thorough searches there is nothing to indicate that any of these would perform better.

An interesting difference between the two rotations regards the scores for the (seemingly excessive) dimensionality of 309. It seems that the promax rotation penalizes excessive dimensionality harder than the varimax rotation. This may be explained by the combined effect of the added noise from the dimensionality itself, magnified by the raising of the loadings to a power of 2.

8.3.2.3 Queries attaining predetermined values of recall

As already mentioned in Subsection 7.5.2, in order to limit the cognitive load on the user, we need to limit the number of documents retrieved for any Uexküll scene, a situation that implies loss of recall. In line with this, many of the queries in our simulations never attain full recall. Figures 8.12-8.14 show the number of queries that attain prescribed recall values (> 0, 20%, 40%, 60%, 80% and 100%). The value at point 0 gives the number of queries for which any relevant documents (more than 0) were retrieved. In

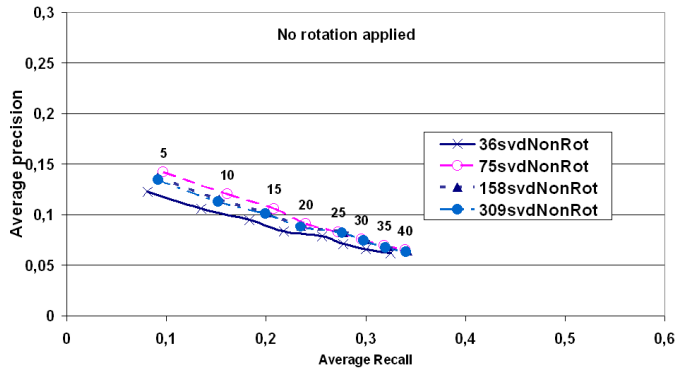


Figure 8.9: *Simulation scenario 1: average recall and precision at DCV points for non-rotated SVD*

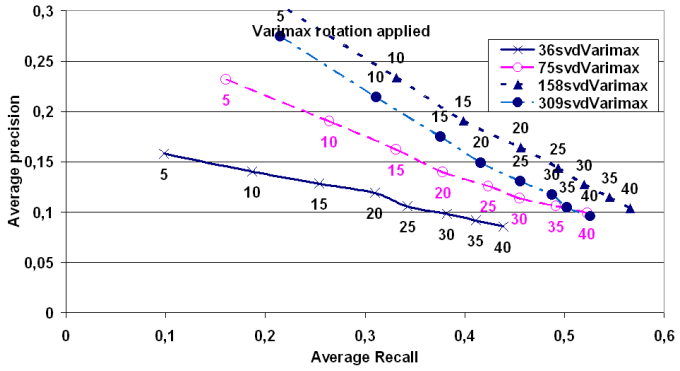


Figure 8.10: *Simulation scenario 1: average recall and precision at DCV points for orthogonally rotated SVD*

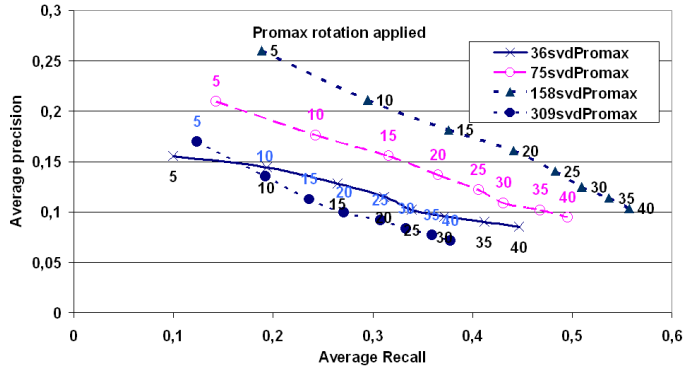


Figure 8.11: *Simulation scenario 1: average recall and precision at DCV points for obliquely rotated SVD*

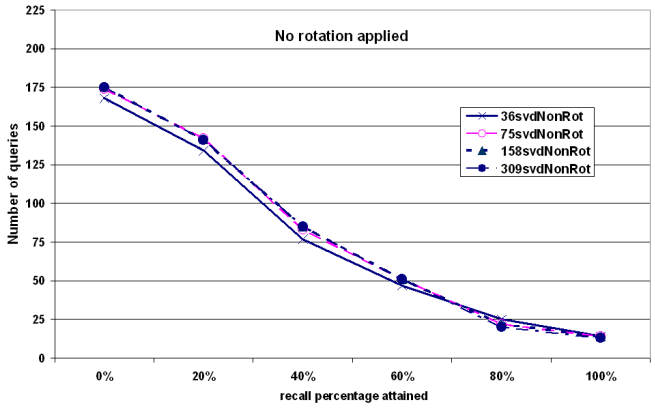


Figure 8.12: *Simulation scenario 1: numbers of queries (out of 225) attaining at least the indicated values of recall: non-rotated SVD, varying dimensionality*

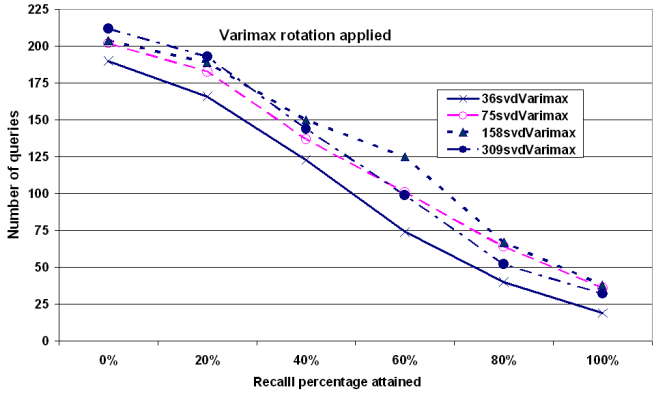


Figure 8.13: *Simulation scenario 1: numbers of queries (out of 225) attaining at least the indicated values of recall: varimax-rotated SVD, varying dimensionality*

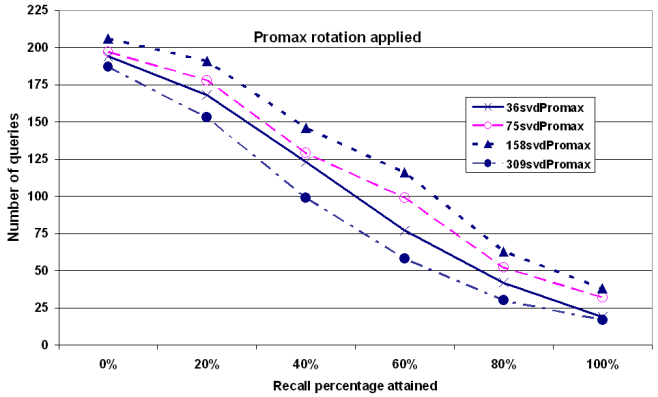


Figure 8.14: *Simulation scenario 1: numbers of queries (out of 225) attaining at least the indicated values of recall: promax-rotated SVD, varying dimensionality*

appendix C enlarged versions of these same figures, and in addition figures that show rotation differences within dimensionalities, are provided.

8.3.2.4 The R-precision measure

In Table 8.4 we present the summary results of the R-precision measure over all organizations. As far as this simulation scenario can detect, there is no gain in using 309 dimensions in relation to 158, and for the best dimensionalities the varimax-rotated organizations outperforms the promax-rotated ones.

Table 8.4: *Simulation scenario 1: summary of the R-precision results for organizations, ordered by the rank sums. Data organizations that differ significantly . See Table B.1 for detailed statistics*

rp1d	V158	V309	P158	V75	P75	V36	P36	P309	N75	N309	N158	N36
Mean results	0,262	0,245	0,228	0,207	0,186	0,144	0,142	0,148	0,129	0,126	0,123	0,11
V158		~	>	>	>	>	>	>	>	>	>	>
V309			~	>	>	>	>	>	>	>	>	>
P158				~	>	>	>	>	>	>	>	>
V75					~	>	>	>	>	>	>	>
P75						~	~	~	>	>	>	>
V36							~	~	>	>	>	>
P36								~	~	~	~	>
P309									~	~	~	>

8.3.3 Results for simulation scenario 1: visualization support measures

Results for the visualization support measures, SRE and SRP, developed in Subsections 7.4.3 and 7.4.4, respectively, are presented in Tables 8.5 and 8.6. The results in Tables 8.5 and 8.6 are similar to the results obtained for the R-precision measure(Table 8.4). Also for the visualization support measures, the superiority of the varimax rotation is apparent. In similarity with the R-precision results, the promax-rotated organization (unlike varimax), lose interpretability at high dimensionalities.

Table 8.5: *Simulation scenario 1: summary of the SRE results, with the significant differences between data organizations at $\alpha = 2.5\%$. See Table B.6 for detailed statistics*

se1d	V309	V158	P158	V75	P75	P309	V36	P36	N309	N158	N75	N36
Mean result	0,27	0,239	0,225	0,178	0,173	0,178	0,133	0,13	0,116	0,114	0,112	0,097
V309		~	>	>	>	>	>	>	>	>	>	>
V158			~	>	>	>	>	>	>	>	>	>
P158				>	>	>	>	>	>	>	>	>
V75					~	~	>	>	>	>	>	>
P75						~	>	>	>	>	>	>
P309							~	>	>	>	>	>
V36								~	>	>	>	>
P36									>	>	>	>

Table 8.6: *Simulation scenario 1: summary of the SRP results, with the significant differences between data organizations at $\alpha = 2.5\%$. See Table B.2 for detailed statistics*

srp1d	V309	V158	P158	V75	P75	P309	V36	P36	N75	N158	N309	N36
Mean result	0,167	0,143	0,123	0,091	0,083	0,102	0,05	0,045	0,045	0,043	0,044	0,034
V309		~	>	>	>	>	>	>	>	>	>	>
V158			>	>	>	>	>	>	>	>	>	>
P158				>	>	>	>	>	>	>	>	>
V75					~	~	>	>	>	>	>	>
P75						~	>	>	>	>	>	>
P309							>	>	>	>	>	>
V36								~	>	>	>	>
P36									>	>	>	>

8.3.4 Summary measures: discussion

After presenting the statistical significance overview of the summary measures, we wish to discuss the results and their implications. An important question to discuss is how the novel measures behave in relation to R-precision, which is an established measure. Another question is how different rotations and differences in dimensionality are rewarded/penalized by the measures.

In Figures 8.15 - 8.17 we present both the measure scores in three configurations. The first configuration shows the ranking of the measure scores sorted in descending order of R-precision scores. The second configuration orders data organizations by dimensionality within each rotation type, and the third shows the data organizations ordered by rotations within each dimensionality.

The two visualization support measures are quite well correlated. Also, both visualization support measures perform quite similar to R-precision. The exception is the 309-dimensional rotated organizations (V309 and P309). Recalling from the curves in Figures 8.5 - 8.14 (average interpolated precision), as well as Figures 8.9 - 8.11 (recall/precision at DCVs), 309 dimensional organizations render poorer overall performance, seemingly due to a large component of excessive-dimensionality noise. This is, of course, still apparent in the plots above. Still, both visualization support measures seem to indicate a better interpretation of the axes of the rotated 309-dimensional organizations than R-precision does. From Figure 8.16 we see that this goes particularly for the orthogonal rotation, but also for the oblique rotation the loss of interpretability is less severe for the visualization support measures than for R-precision. These tendencies need further scrutiny before they are established, but the apparent loss of interpretability for very high dimensional organizations, particularly for the obliquely rotated ones, also seems to be in line with their poor distribution of concept axes, which will limit the *diversity of access*. This is discussed in Subsection 8.3.7.

8.3.5 Results for simulation scenario 2: ranked list measures

In the following we are presenting and discussing plots and tables of results for Simulation scenario 2 for the automatically indexed version.

The plots and tables are parallel in layout to those created from simulation scenario 1. Figures 8.18 - 8.20 show the average interpolated precision for 3 rotations, Figures 8.21 - 8.23 show the average recall at constant DCV

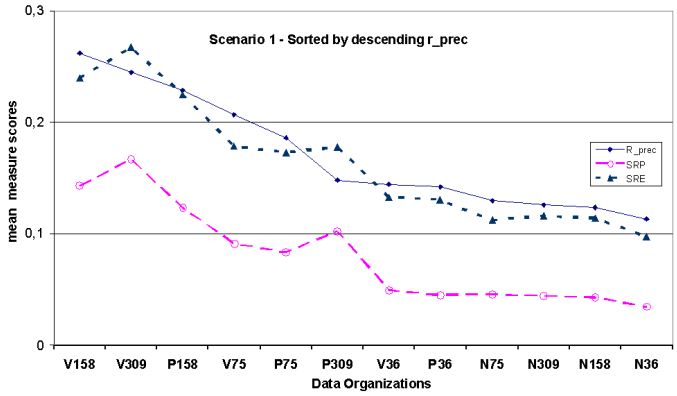


Figure 8.15: *Simulation scenario 1: measure scores ordered by descending R-precision.*

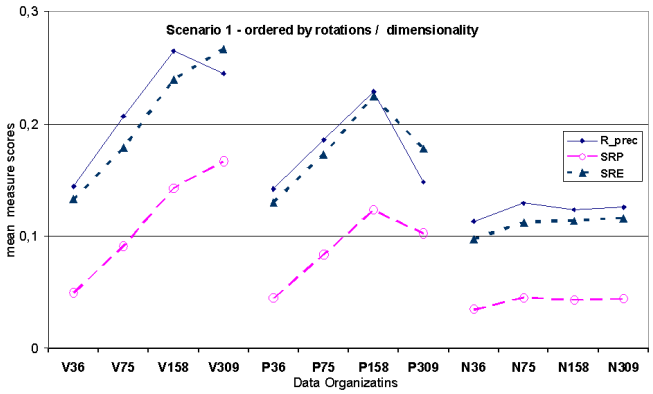


Figure 8.16: *Simulation scenario 1: measure scores ordered by rotations/dimensionalities.*

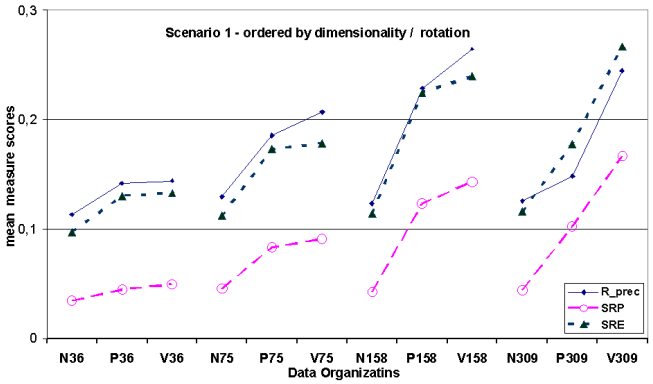


Figure 8.17: *Simulation scenario 1: measure scores ordered by dimensionalities/rotations.*

points, and Figure 8.24 - 8.26 show the numbers of queries (out of 225) that attain some prescribed values of recall.

Simulation scenario 2 shows a slight decline in recall for all organizations, but the tendencies exhibited in Simulation scenario 1 are generally preserved. For the 158-dimensional varimax-rotated organization we detect a slight improvement in precision, but it is hardly significant.

8.3.5.1 Results for simulation scenario 2: R-precision

We conclude the presentation of ranked list measures with Table 8.7. As for Simulation scenario 1, the results are given in their ranked order for the data organizations, specifying the significant differences. Also here, the tendencies are the same as for Table 8.4, but the current scenario exhibits a slightly higher discriminatory power.

Table 8.7: *Simulation scenario 2: summary of the R-precision results, with the significant differences between data organizations at $\alpha = 2.5\%$. See Table D.1 for detailed statistics*

rp2d	V158	V309	P158	V75	P75	P309	P36	V36	N158	N309	N75	N36
mean result	0,204	0,198	0,198	0,172	0,172	0,15	0,121	0,12	0,093	0,079	0,082	0,067
V158		~	~	>	>	>	>	>	>	>	>	>
V309			~	~	>	>	>	>	>	>	>	>
P158				~	~	>	>	>	>	>	>	>
V75					~	>	>	>	>	>	>	>
P75						~	>	>	>	>	>	>
P309							~	~	>	>	>	>
P36								~	>	>	>	>
V36									>	>	>	>

8.3.6 Results for simulation scenario 2: visualization support measures

In Tables 8.8 and 8.9, values for the SRP and the the SRE measures are listed respectively. The tables give the overall average scores of the measure for all 225 queries, and list the significant differences. More detailed data are presented in Tables D.2 and D.3, respectively. Also here, the results reinforce the tendencies shown in the parallel figures for simulation scenario 1, but the results here exhibit a higher discriminatory power.

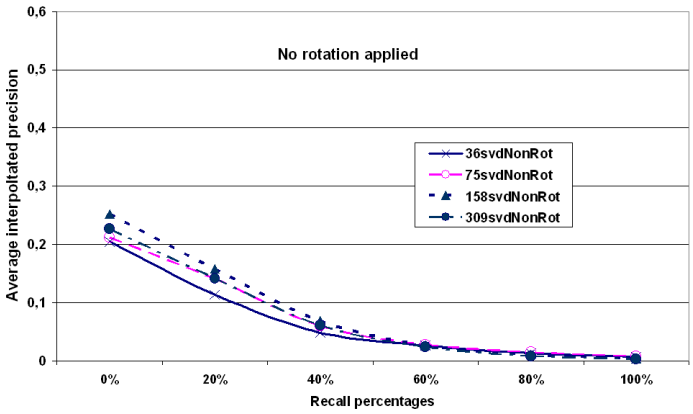


Figure 8.18: *Simulation scenario 2: average interpolated precision at recall percentage points for non-rotated SVD - varying dimensionalities*

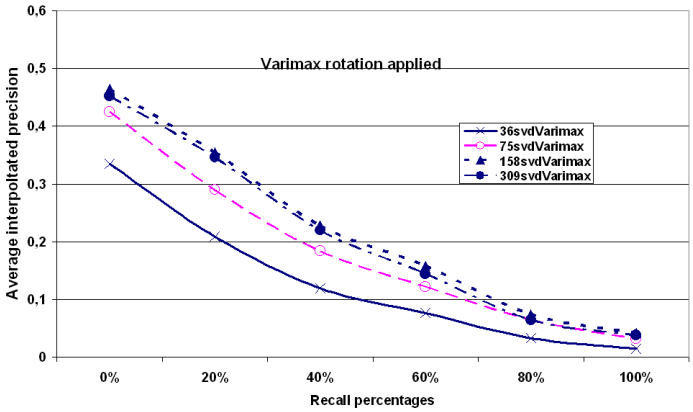


Figure 8.19: *Simulation scenario 2: average interpolated precision at recall percentage points for orthogonally rotated SVD - varying dimensionalities*

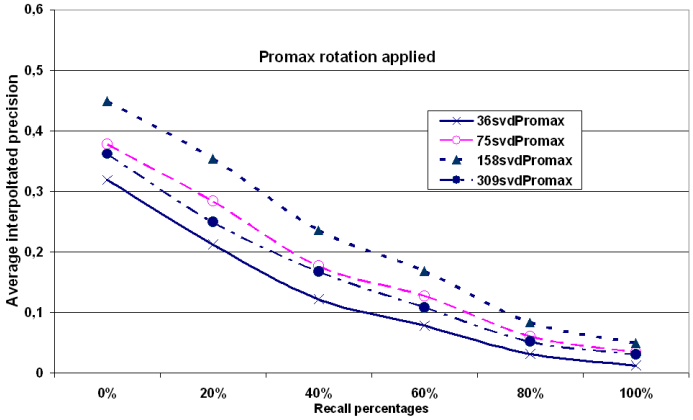


Figure 8.20: *Simulation scenario 2: average interpolated precision at recall percentage points for obliquely rotated SVD - varying dimensionalities*

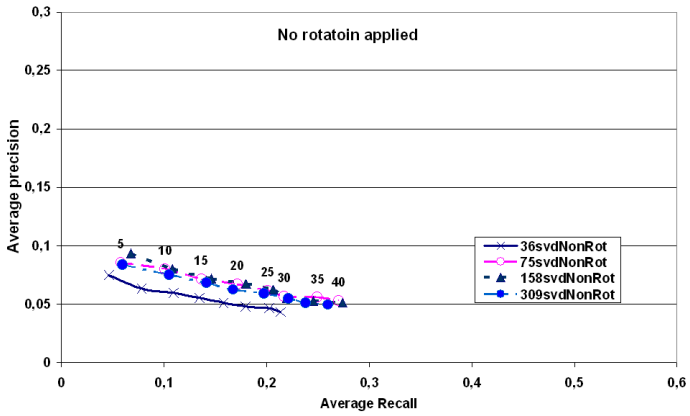


Figure 8.21: *Simulation scenario 2: average precision at DCV points for non-rotated SVD - varying dimensionalities dimensionalities*

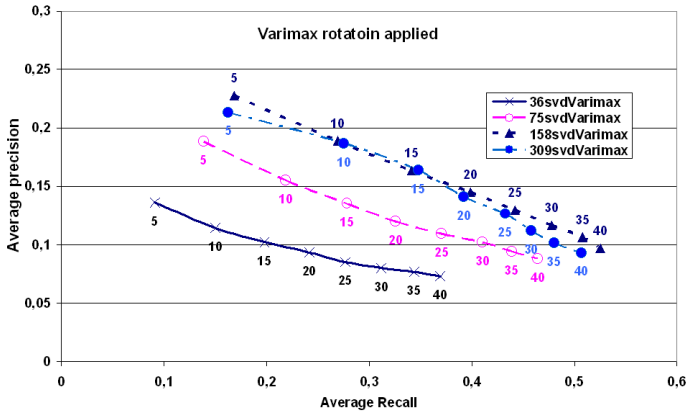


Figure 8.22: *Simulation scenario 2: average recall and precision at DCV points for orthogonally rotated SVD - varying dimensionalities*

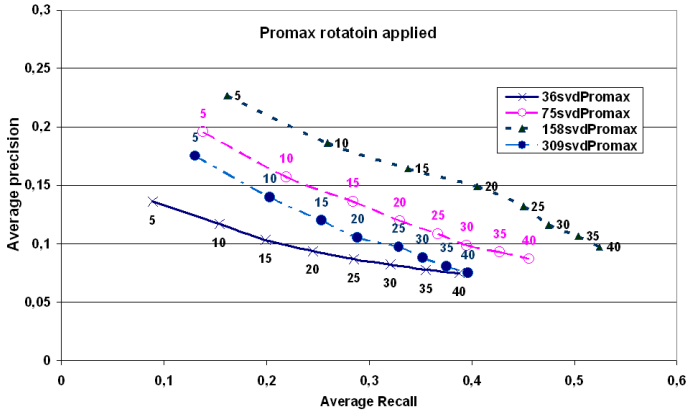


Figure 8.23: *Simulation scenario 2: average recall and precision at DCV points for obliquely rotated SVD - varying dimensionalities*

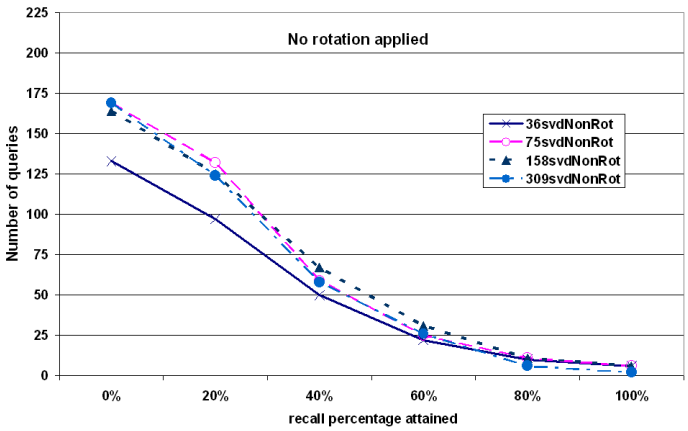


Figure 8.24: *Simulation scenario 2: numbers of queries attaining different values of recall: non-rotated SVD, varying dimensionalities*

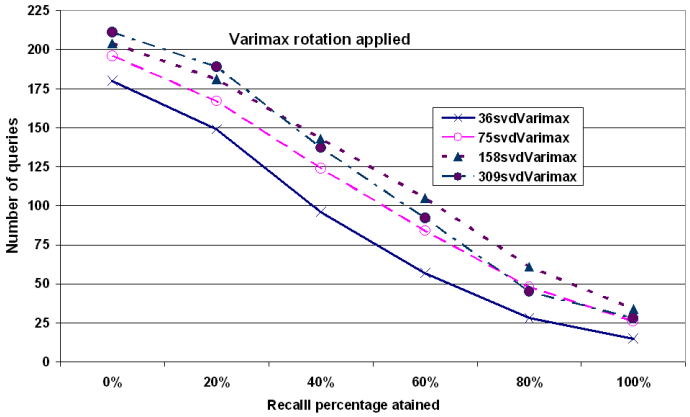


Figure 8.25: *Simulation scenario 2: numbers of queries attaining different values of recall: varimax-rotated SVD, varying dimensionalities*

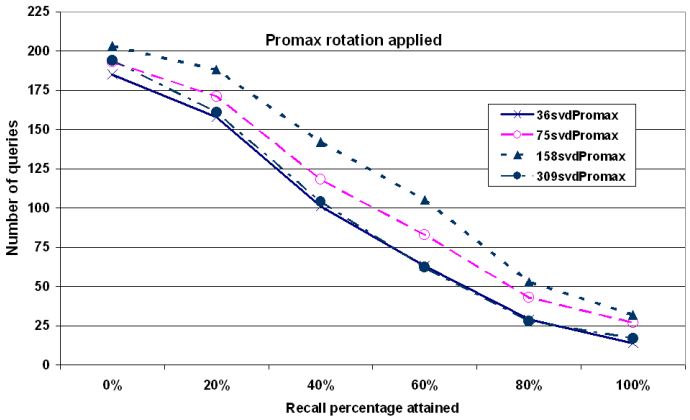


Figure 8.26: *Simulation scenario 2: numbers of queries attaining different values of recall: promax-rotated SVD, varying dimensionalities*

Table 8.8: *Simulation scenario 2: summary of the SRE results, with the significant differences among data organizations at* $\alpha = 2.5\%$

se2d	V309	P158	V158	P309	P75	V75	P36	V36	N158	N309	N75	N36
Mean result	0,22	0,196	0,177	0,187	0,145	0,137	0,099	0,099	0,061	0,058	0,056	0,046
V309		>	>	>	>	>	>	>	>	>	>	>
P158			~	>	>	>	>	>	>	>	>	>
V158				>	>	>	>	>	>	>	>	>
P309					~	>	>	>	>	>	>	>
P75						~	>	>	>	>	>	>
V75							>	>	>	>	>	>
P36								~	>	>	>	>
V36									>	>	>	>
N158										~	~	>

Table 8.9: *Simulation scenario 2: summary of the SRP results, with the significant differences among data organizations at* $\alpha = 2.5\%$

srp2d	V309	V158	P158	V75	P309	P75	P36	V36	N158	N309	N75	N36
Mean result	0,108	0,086	0,074	0,058	0,078	0,047	0,033	0,032	0,032	0,032	0,033	0,028
V309		~	~	>	>	>	>	>	>	>	>	>
P158			~	~	>	>	>	>	>	>	>	>
V158				~	>	>	>	>	>	>	>	>
P309					~	~	>	>	>	>	>	>
V75						~	>	>	>	>	>	>
P75							>	>	>	>	>	>
P36								~	>	>	>	>
V36									>	>	>	>
N158										~	~	>

8.3.7 Diversity of access: distribution of concept axes among queries

We refer to Subsection 7.5.3, where we have discussed the distribution of concept axes among queries as a gauge for the quality of the data organization. In Figures 8.27 - 8.33 we present some curves of this gauge. The data are extracted from the results of simulation scenario 1. The differences between the rotated and the non-rotated organizations are apparent. Whereas for the non-rotated organizations one axis occurs in more than half of the queries, the distribution for the rotated organizations is more uniform. It is also interesting to note that the distribution gets closer to uniform as the number of dimensions increases from 36 to 75 and to 158, but 309 dimen-

sions provide no further gain in uniformity. For the varimax rotation there is little change, whereas for the promax rotation we observe loss of uniformity, which is in line with the changes we observed for the ranked list measures (e.g. Figures 8.5 - 8.7).

8.4 Some results for the manually indexed version

In this section we present some results for the manually indexed version of the collection, discussed in Subsection 5.4.4.3.

8.4.1 Selected precision recall results

Figures 8.34 - 8.36 replicate the results from Figures 8.5 - 8.7, applied to The manually indexed version. The results are worse in absolute terms, but the tendencies are similar, indicating that rotations emphasize relevant documents, as well as improve the gain from extra dimensionality. The performance gain from additional dimensionality under rotation is apparent also in this case, albeit on a smaller scale. The highest number of dimensions we experimented with, 309, shows best relative performance where oblique rotation has been applied, and slightly worse relative performance where the orthogonal rotation or no rotation at all has been applied. This is unlike the case of the automatically indexed collection, where the 309 dimensional organizations entail slightly worse relative performance particularly when promax rotation is applied. The reason for this difference is unclear.

8.4.2 Summary measure statistics

In Figures 8.37 - 8.39 we present charts that show how the different measures express changes in the parameters of the experiments, following the presentation pattern of Figures 8.15 - 8.16. These build on statistics presented in Tables 8.10 - 8.12.

Also these results show the same tendencies as the results from Subsection 8.4.1, where the P309 organizations provide relatively better performance. We repeat, again that we must be cautious in interpreting the results, as even though some of the readings are statistically significant (e.g. the difference between P309 and P158 in Figure 8.39), some other are not. There is no clear pattern here that distinguishes the novel from the established measures, besides a somewhat better discriminatory power. More of

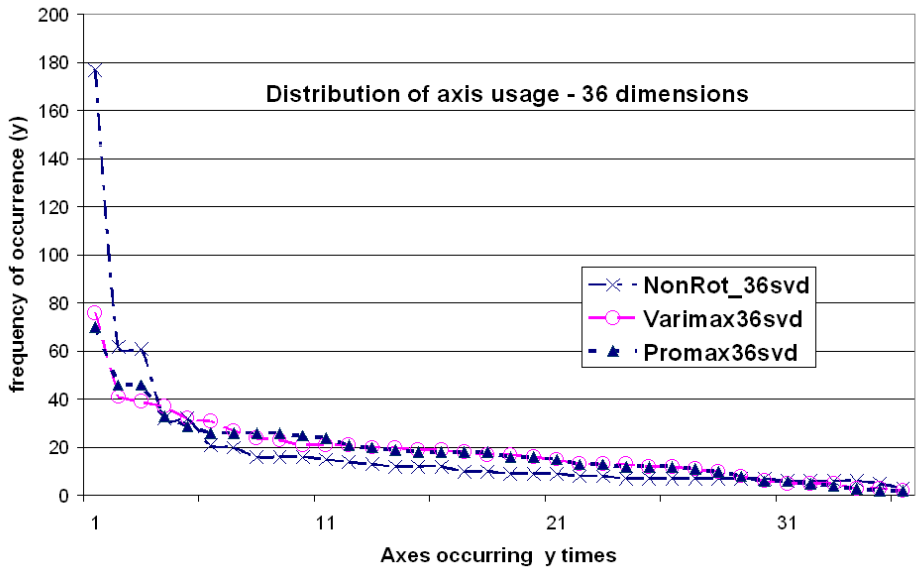


Figure 8.27: *Distribution of concept axes among queries: non-rotated, Varimax rotated and promax-rotated SVD, 36 dimensions*

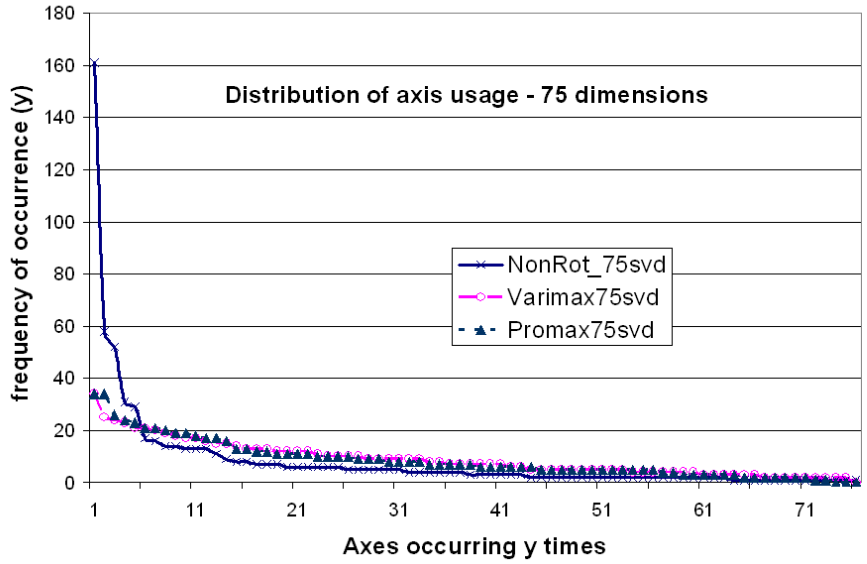


Figure 8.28: *Distribution of concept axes among queries: non-rotated, varimax-rotated and promax-rotated SVD, 75 dimensions. See a detailed view of the lower part of the scale in Figure 8.31*

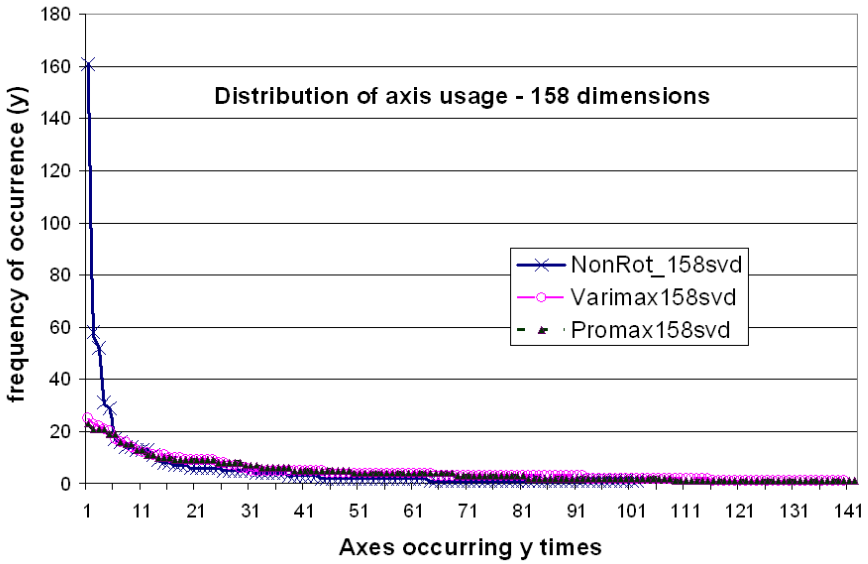


Figure 8.29: *Distribution of concept axes among queries: non-rotated, varimax-rotated and promax-rotated SVD, 158 dimensions. See a detailed view of the lower part of the scale in Figure 8.32*

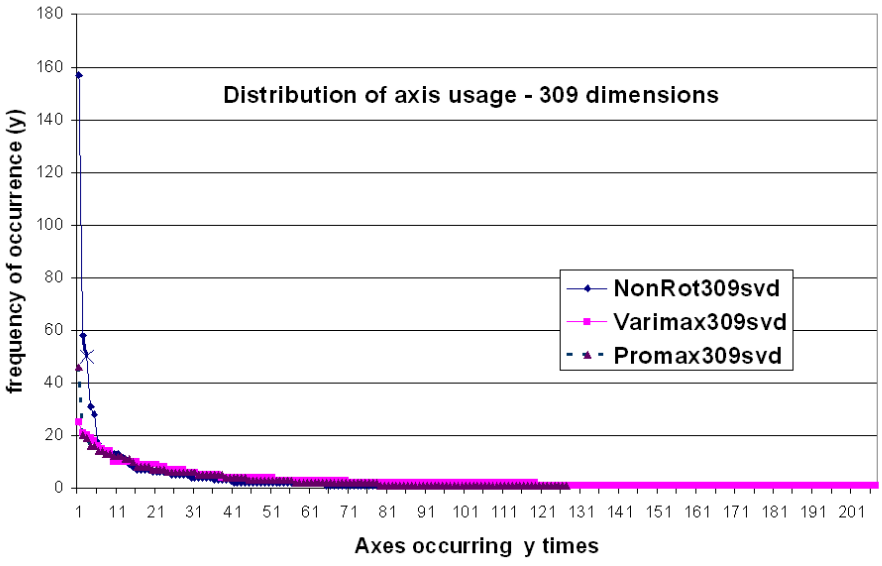


Figure 8.30: *Distribution of concept axes among queries: non-rotated, varimax-rotated and promax-rotated SVD, 309 dimensions. See a detailed view of the lower part of the scale in Figure 8.33*

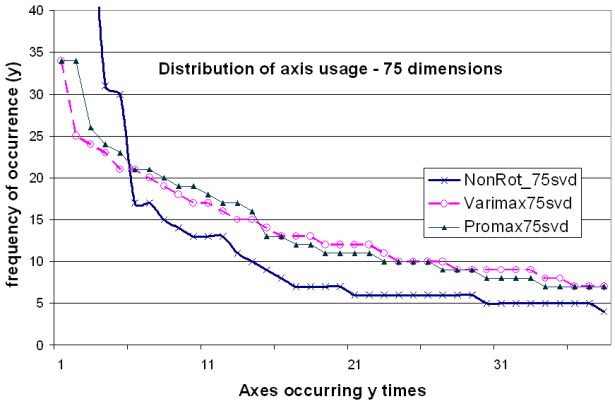


Figure 8.31: *Distribution of concept axes among queries (75 dimensions): a detailed view of the lower part of the scale of Figure 8.28*

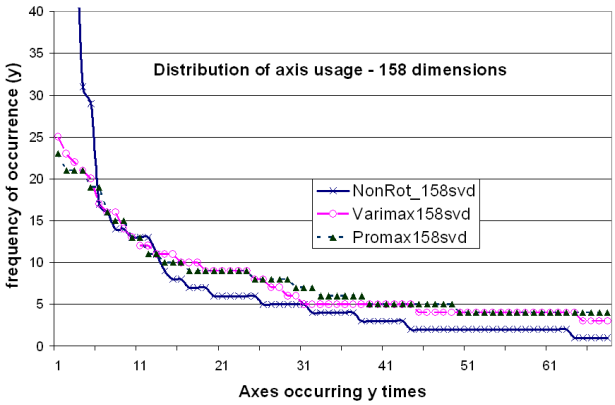


Figure 8.32: *Distribution of concept axes among queries (158 dimensions): a detailed view of the lower part of the scale of Figure 8.29*

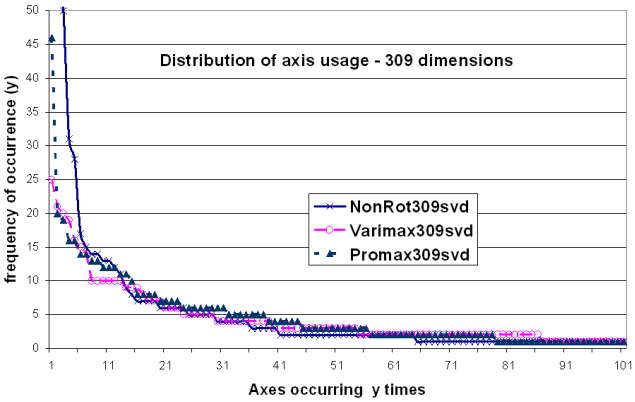


Figure 8.33: *Distribution of concept axes among queries (309 dimensions): a detailed view of the lower part of the scale of Figure 8.30*

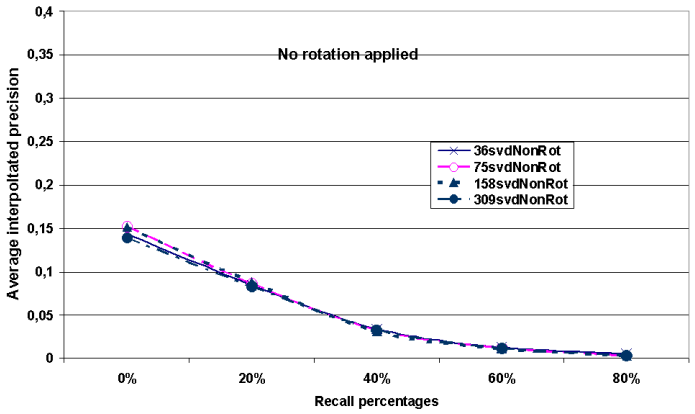


Figure 8.34: *Scenario 1 (manually indexed version): average interpolated precision at recall percentage points for non-rotated SVD*

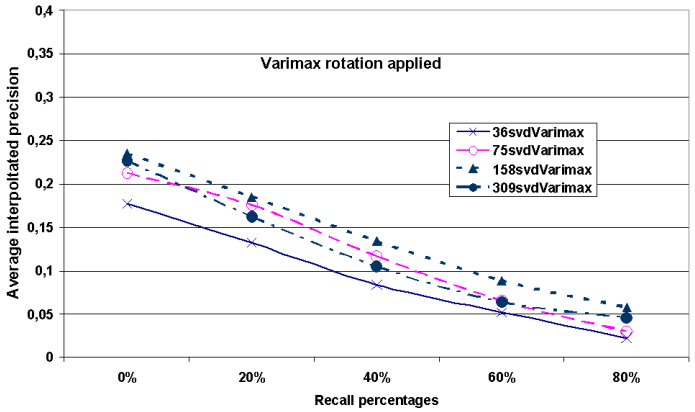


Figure 8.35: *Scenario 1 (manually indexed version): average interpolated precision at recall percentage points for orthogonally rotated SVD*

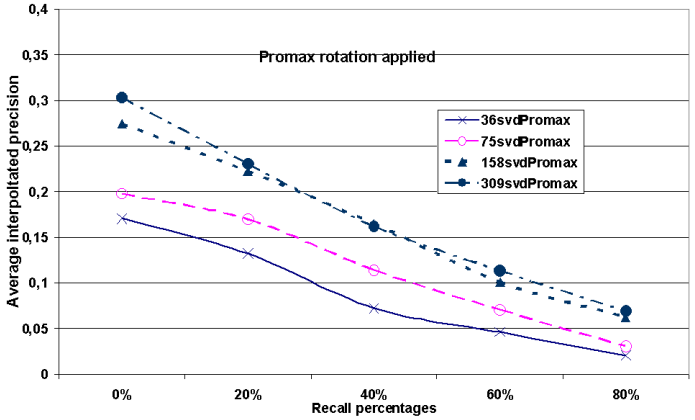


Figure 8.36: *Scenario 1 (manually indexed version): average interpolated precision at recall percentage points for obliquely rotated SVD*

the differences between dimensionalities within rotations are significant for the visualization support measures than for R-precision.

8.5 Different rotations

Our treatment of axis interpretability is highly exploratory, and by no means exhaustive. When choosing rotations to use in our experiments, we had to assume that the implementations of the rotation algorithms are well tested and will not produce unpredictable results due to the large number of factors we experimented with. One way of ensuring this is the use of well tested implementations, such that are available in a package like SPSS. Algorithms like varimax and promax are fast and efficient, and were obvious candidates for experimentation. We have also applied some other rotations (the orthogonal quartimax and the oblique oblimin) to our decompositions, without getting any better results.

8.6 Summary - axis interpretability

In our context, axis interpretability is, as indicated in the beginning of this chapter, an aggregate quality that defines and characterizes the suitability of a data organization to the Uexküll approach. Since this is not a unidimensional characteristic, it is not trivial to rank data organization by interpretability.

The goal of the experiments reported in this chapter was not the characterization of the Cranfield collection as such, but if we do take Cranfield as an example, then our results suggest that for our purposes, the optimal dimensionality of the Cranfield collection lies closer to 158 than to any other of the dimensionalities experimented with, and that rotations increase the interpretability significantly.

The major variables we were allowing to vary, dimensionality and rotation, influence the interpretability of an organization, and will determine the potential of a data organization to support Uexküll-based retrieval for end users. Still, exploiting this potential is highly dependent on a semantic component - contextually sound naming of axes, that aggregates on top of the variables discussed here to enable users to find their way in the material.

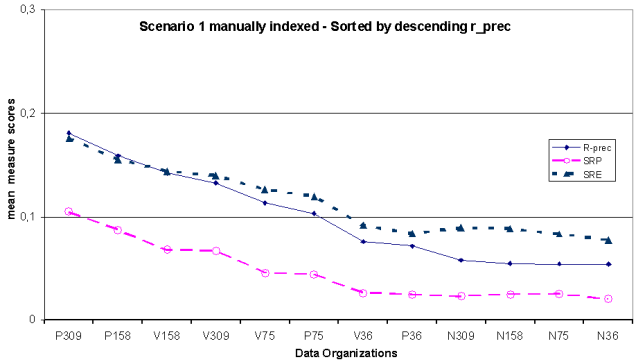


Figure 8.37: *Simulation scenario 1 (manually indexed collection): measure scores ordered by descending R-precision.*

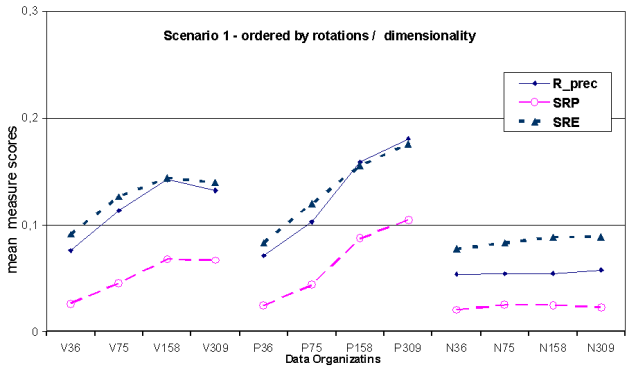


Figure 8.38: *Simulation scenario 1 (manually indexed collection): measure scores ordered by rotations/dimensionalities.*

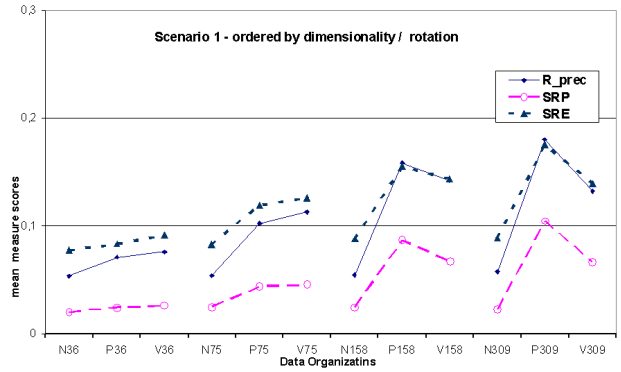


Figure 8.39: *Simulation scenario 1 (manually indexed collection): measure scores ordered by rotations/dimensionalities.*

Table 8.10: *Simulation scenario 1 (manually indexed collection): summary of the R-precision results, with the significant differences among data organizations at $\alpha = 2.5\%$. See Table F.1 for detailed statistics*

rp1d	P309	P158	V309	V158	V75	P75	P36	V36	N309	N158	N75	N36
Mean results	0,181	0,159	0,132	0,142	0,113	0,103	0,071	0,076	0,058	0,055	0,054	0,05
P309		~	>	>	>	>	>	>	>	>	>	>
P158			~	~	>	>	>	>	>	>	>	>
V309				~	~	>	>	>	>	>	>	>
V158					~	>	>	>	>	>	>	>
V75						~	>	>	>	>	>	>
P75							~	~	>	>	>	>

Table 8.11: *Simulation scenario 1 (manually indexed collection): summary of the separation-rewarded exposure results, with the significant differences among data organizations at $\alpha = 2.5\%$. See Table F.3 for detailed statistics*

se1d	P309	P158	V158	V309	V75	P75	V36	N309	N158	N75	P36	N36
Mean result	0,18	0,155	0,144	0,14	0,126	0,12	0,091	0,089	0,088	0,083	0,084	0,078
P309		>	>	>	>	>	>	>	>	>	>	>
P158			~	~	~	>	>	>	>	>	>	>
V158				~	~	>	>	>	>	>	>	>
V309					~	~	>	>	>	>	>	>
V75						~	>	>	>	>	>	>
P75							>	>	>	>	>	>
V36								~	~	~	~	>

Table 8.12: *Simulation scenario 1 (manually indexed collection): summary of the SRP results, with the significant differences among data organizations at $\alpha = 2.5\%$. See Table F.2 for detailed statistics*

srp1d	P309	P158	V158	V309	V75	P75	V36	P36	N158	N309	N75	N36
Mean results	0,108	0,086	0,074	0,058	0,078	0,047	0,033	0,032	0,032	0,032	0,033	0,028
P309		~	>	>	>	>	>	>	>	>	>	>
P158			~	~	>	>	>	>	>	>	>	>
V158				~	~	~	>	>	>	>	>	>
V309					~	~	>	>	>	>	>	>
V75						~	>	>	>	>	>	>
P75							>	>	>	>	>	>
V36								~	~	~	>	>

Part III

Summary and further research

Chapter 9

Discussion

In this chapter we discuss the Uexküll approach and our approach to evaluating salient aspects of it, in light of our results. We assess the results obtained for the simulation experiments and discuss open questions and further research.

In the dissertation we have done the following:

- Developed an apparatus of simulations and measures that makes it possible to algorithmically test and evaluate the suitability of multivariate data organizations to the Uexküll approach, with usability as an important criterion.

- Decomposed a test collection of classical IR, the Cranfield collection (represented by a term-document matrix), using the singular value decomposition (LSI) into a number of data organizations. We were applying three different rotations: arbitrary (non-rotated), varimax (orthogonal rotation) and promax (oblique rotation).

- Applied the apparatus to the organizations, indicating the connection between parameters like dimensionality and rotation and the potential of the data organizations for good retrieval.

In this chapter we follow a reverse path, starting by reviewing the results we have presented in Chapter 8, thereof reflecting on the measures that we have used. Further, we discuss the viability of our simulation apparatus. The last sections address the approach in light of some related work and problem topics, also presenting some possibilities for further research. We summarize with answers to the research questions presented in Chapter 1.

9.1 Measures and results

The experiments reported in Chapter 8 were meant to combine two purposes: attempting to assess the support of the data organizations to the Uexküll approach[1], as well as supporting the assessment of the evaluation approach as such, as a separate research issue (research question 4). The present section and the sections below discuss results in this light.

The support for visual retrieval is dependent on two components: the quality of the ranking and the separation of relevant document from non-relevant ones. If the ranking obtained by matching of queries against documents within a data organization is good, and this organization also supports visual separation of documents, then it will provide potential for good retrieval. The set up of the experiments was meant to give us detailed information about this combined effect. This information was sought from several angles, represented by the different measures that we were using. In the following we discuss some of our results, trying to summarize aspects for which we get answers from our simulations.

Unless otherwise indicated, all results referred to in this section are obtained using the sum model (see Subsection 6.3.2).

9.1.1 Centroid emphasis

The centroid emphasis measure was developed with a graphical and a numeric representation. The purpose of the numeric calculation of the CE was to measure the slope of the superiority of the OCP over the inferior projections. This was meant to give an initial numerical characterization of the organizations. Looking at the results, we note two important points:[2]

- Rotated organizations score higher, meaning that the CE rewards rotation.

- Low-dimensional organizations insistently score higher. As high values of CE indicate a high ratio between the OCP and the inferior projections, this means very rapid loss of interpretability for inferior projections of low dimensional organizations.

[1] The usability of the Uexküll approach itself would have to be determined through user tests.

[2] In this discussion we use CE to denote the numerical measure rather than the graphical expression

Axis interpretability measured exclusively in terms of the centroid emphasis is far too restricted if we wish to predict performance. As a consequence, CE would only be partially instrumental in predicting the differences in performance (the way performance is defined and measured in our treatment) among data organizations and their ranking. With our current knowledge , the value of the CE lies in facilitating the detection of the appropriateness of a data organization for Uexküll based retrieval.

9.1.2 Precision characteristics of data organizations

For the graphical presentation of precision performance, we chose to compare data organizations by average interpolated precision, as well as by average precision at constant DCV points. From the latter measure (Figures 8.9-8.11), we see that our best performing organizations attain average non-interpolated precision (at the lower recall region) of about 30%.

Average interpolated precision results show precision of about 60% at 0 recall, which, being attained with a limited number of retrieved documents, is an acceptable result.

9.1.3 Recall characteristics of Uexküll

In our treatment, we have three sources of information about the recall performance of each organization: plots of average interpolated precision, plots of average precision at DCV points, and plots graphing the numbers of queries attaining prescribed values of recall, for all our data organizations.

Recall at constant DCVs is closely related to precision, because it is the higher the more documents judged relevant are ranked within each DCV. From Figures 8.9 - 8.11 we see that on average, our best organizations attain around 50% recall, while the inferior organizations attain about 30% recall.

Since few queries ever attain full recall at DCV@40, it was also interesting to see how many *queries* were attaining more modest (prescribed) measures of recall, 20%, 40%, 60% and 80%.

In Chapter 8 we have produced some figures (for example Figures 8.12 - 8.14, the purpose of which was to gauge recall performance in these recall points. These figures show dimensionality differences within rotations. In Appendix C we were providing enlarged versions of these same figures, and additionally figures that show rotation differences within dimensionalities.

For the non-rotated configurations there is, also when focusing exclusively on recall, very little gain in increasing the dimensionality.

9.1.4 Summary measures

We have used a number of summary measures, the R-precision representing the legacy ranked list measures, as well as the specially developed separation-rewarded exposure (SRE) and separation-rewarded precision (SRP), both in an attempt at measuring visualization support. The summary measures also provided a possibility for a simple significance test of the results.

The visualization support measures are an important contribution to the approach, and an interesting question is whether the experiments are rendering any difference between them and the R-precision. Such difference can be either in providing different ranking of organizations than the R-precision, or providing a richer result, showing better visual performance for equally performing organizations.

Observing the results of the sum model (Figures 8.15 - 8.17), we see that the visualization support measures do correlate quite well, both deviating slightly from the R-precision measure, notably for rotated organizations with excessive dimensionality. The latter seem to perform better than the lower dimensional ones when it comes to visual separation, while performing worse than the former when it comes to pure ranking. For high-dimensional organizations, the R-precision rewards the promax-rotated organization only marginally in relation to the non-rotated organization, whereas the visual support measures reward it significantly. This may be due to the promax rotations separate documents to a greater extents, which would manifest in the separation but not manifest in the ranking.

In Subsection 7.4.5 we discussed the two visualization support measures, and their significance as gauges for the ability of a data organization to support two kinds of users, respectively the user who is in search of a factual answer (or maybe a known item), and the user who is engaged in a broader search about a topic. We discussed the extent to which the SRP and the SRE would measure the ability of a system implementing the approach to accommodate the needs of these two types of users.

The problem in algorithmic detection of the difference has two main aspects to it:

- Support for the two types of users is not mutually exclusive. A single organization may indeed provide support for both user groups.
- The differences between support for the user types would manifest in different layouts of the results (represented as ranked lists). An "SRE user" would be served sufficiently well by having a few relevant documents prominently located. The "SRP user" would probably suffice

having the same documents detectable, but will *in addition* require the remaining relevant documents to be positioned so he can find them. This distinction may be hard to detect algorithmically within the evaluation paradigm and the test collection we are using.

To gauge this distinction we would need to

- design special groups of requests that would potentially be posed exclusively by either of these kinds of users
- run separate experiments applying these groups of requests to the collection, applying our measures to those.

Such requests would not represent *different topics* as requests ordinarily do within the laboratory model, rather *different expected result layouts* (see Subsection 7.4.6). To do that we would need to classify requests as geared towards these type of users in a mutually exclusive manner. This may require a test collection that is constructed in a different way.

Having said all that, we would like to draw attention to Figures B.1 - B.3 reporting results from the max model (see Subsection 6.3.1). In these figures, particularly Figure B.2, we see a slight change in behavior among the organizations, where the SRP-measure penalizes the 309 dimensional promax organization less than the other measures do. This would give rise to the hypothesis that to the extent that the max model is a good representative for user behavior against an Uexküll system, and to the extent that the result (which is not rendered statistically significant by our experiments) holds, the "SRP user" (user in search of all documents about a one or more topics) is better served by a high dimensional varimax-rotated organization.

9.1.5 Diversity of access

Figures 8.27 - 8.33 are an attempt at gauging the effect of rotations on the distribution of axes among Uexküll groups when automatically choosing prominent axes as representatives for the different queries. These figures show that more axes are used when rotation are employed, and also the dimensionality affects this phenomenon positively, up to a point. This indicates that at least in the configurations we were experimenting with, diversity of access is positively correlated with good retrieval performance, which is encouraging. At this point we only have summary results regarding this phenomenon, but it may, also here, be interesting to try to see a more detailed picture by looking at sample queries and see which concepts are employed in creating the Uexküll groups for these.

9.1.6 Implications of the results

The correspondence between the relative quality of the diversity of access (see Subsection 7.5.3) and that of the results in terms of the legacy measures as well as the newly defined retrieval measures, having the same organizations perform well on both, is a promising result of the current treatment. Further pursuing this result entails controlling variables other than those controlled so far, possibly a change in the evaluation paradigm. One way of achieving this might be a contentual analysis of the queries, choosing queries with properties that contentually span many aspects of the collection, at the same time having more control of the axes, and their names (or their contentual significance, in lack of a proper naming scheme).

All above discussed reservations taken into account, we claim that our results are positive and encouraging, and that there are data organizations that would be able support the Uexküll approach. Referring to results presented earlier for the best organizations we were evaluating, there is evidence that an Uexküll group generated from such an organization may potentially support Uexküll retrieval. And, of course, the way users in practice will access such a system will also determine the extent of benefit they would draw from the better data organizations.

The above derives from the purely technical side of the treatment. Usefulness here is potential, and (as we have repeatedly claimed) contingent on facilitating for semantic access through naming of concepts as well as visual access through design of an adequate graphical user interface.

9.2 The viability of the simulation approach

The simulation approach used in this project is simpler and more direct than the simulations in the efforts discussed in Chapter 6, with the exception of Leouski and Allan (1998). It has some similarity with the approaches of Magennis and van Rijsbergen (1997), R. W. White et al. (2004) and Lin (2007). The simulations made use of a simple model of a 3D scene, called a location model, and accessed the parameters of this model directly. The relatively large numbers of queries provided us with good potential of attaining descriptive statistics.
The simulation approach was considered as a strategy because of the concern that a tedious user test, if failed (meaning that the users were not satisfied) would not provide us with answers as to the reason of the failure. Thus, the simulation should be considered a preliminary, not a substitute, for user test.

Prior to a user test we will not have an estimate of the extent of improvement in *experienced performance* (by a user at a terminal) that would correspond to any improvement in *simulated performance*.

In the following we will review some aspects of the simulations, and on this basis try to discuss the viability of the simulations as an instrument.

9.2.1 Validity of the information provided by the simulations

In practical terms, the simulation provided a tentative performance ranking of the data organizations that were subject to experimentation. As shown in Chapter 8, the ranking was consistent with what could be expected. Rotation promotes interpretability (which in our case improves retrieval), and increase in dimensionality, up to a certain point provides more evidence to support good retrieval. Both charts and significance statistics seem to support this assertion.

9.2.2 Using both scenarios

Comparing the plots of simulation scenario 1 with the corresponding plots of simulation scenario 2, the results are generally unchanged, except for a slight fall in performance level for simulation scenario 2. This may raise the issue of the necessity of simulation scenario 2. Looking at the rankings of the summary measures, simulation scenario 2 provides us with a finer evaluation tool for data organizations used for the Uexküll approach. Taking the R-precision measure first, comparing Tables 8.4 and 8.7, we see that simulation scenario 2 finds 49 differences statistically significant while simulation scenario 1 only detects 44. For the SRP and the SRE, comparing Tables 8.5 and 8.6 with Tables 8.8 and 8.9, respectively, simulation scenario 2 finds an additional organization having significant differences to any other organization. Even differences that may not pass a significance test, may serve as *indications* for variation in performance, which may be worth to pursue.

9.2.3 Viability of the simulations

In Section 9.1 we have summarized the information provided by the evaluation, both in terms of graphical results, and in terms of summary measures. Our approach provided us with a rich set of results that we can evaluate aspects of Uexküll with. Notably, the approach gave us a ranking of different

ways of organizing data for visualization, and a framework in which variations of our organizations (different rotations, different data decompositions) may be tested.

The set up of the simulations was meant to compare and rank data organizations. The fact that we could in advance hypothesize which data organizations would be the best to facilitate the usage of the Uexküll approach as defined, and the fact that the results, at least to some extent, are consistent with these assumption, may serve as an evidence (according to Newell (1969)), that our simulations are a viable way of testing the approach. On the other hand, the simulations are too coarse to accommodate e.g. for the distinction between different user types and the usability of Uexküll for their purposes. This is due to the experiments using the laboratory model, characterized by pre-defined queries, as well as the size of the test collection.

9.3 Further research

This section discusses the following aspects:

- Additional work following the same research path
- Towards a user test
- Other directions of research to pursue.
- Open questions

9.3.1 Further work following the same research path

9.3.1.1 Repeating the investigations with a different collection

In IR there is a tradition of using more than one test collection for retrieval experiments. The storage demands from our experiments made us refrain from this in the current research.

Quite many of our results were indicative, and were not statistically significant. Looking more closely into these would require the use of additional test collections, but it will also be interesting for reviewing the results that are rendered statistically significant with the Cranfield collection.

9.3.1.2 Further investigations into query performance

In Subsection 7.5.2 we discussed the benefits of counting queries that attain certain values of recall in a situation with a limited number of retrieved documents. This measure is taken at a collection level, and, not surprisingly, results show that it correlates well with the precision-recall measures.

At a lower level, with queries attaining high recall levels across all organizations, one could check:

- how early relevant documents are retrieved (precision/recall at lower DCVs),
- what kinds of non-relevant documents the user needs to browse ahead of the relevant ones, and try to analyze the reasons for that.

One could additionally observe more closely queries that retrieve very few, or no relevant documents at all across all organizations, and try to find common traits attributable to these.

Those investigations, at a document/query level may require some contentual knowledge of the test collection, which was considered to be beyond the scope of the present project. They would also benefit highly from tests and comparisons with similar investigations carried out against other collections.

9.3.1.3 Experiments using graded relevance judgments

In Subsection 7.3.2 we presented some measures of retrieval performance, among which we briefly discussed cumulated gain-based measures (Järvelin & Kekäläinen, 2002) and generalized recall and precision (Kekäläinen & Järvelin, 2002b). These measures were designed to exploit test collections having graded relevance judgments, i.e. relevance judgments that go beyond the binary relevant/non-relevant assessment.

Since the Cranfield collection has graded relevance judgments, it would be interesting to see whether the well performing organizations also perform well using these measures.

9.3.1.4 Uexküll and intellectually indexed databases

The experiments that we have conducted in Section 8.4 are ad-hoc, in the sense that they use an indexing that is not genuinely intellectual. Besides, it uses mathematical methods that are not designed for this type of input matrix. Still, the results that we get, though poor, match the results from the automatically indexed version of the collection.

The aim of these experiments was to gauge the technical aspects of facilitating intellectually indexed databases to the Uexküll approach. The present methods that we are using will not be able to detect the effect of the improved intellectual guidance that intellectually assigned concept names would provide when choosing Uexküll groups for navigation. This will, again need to be done with user participation.

Multivariate transformation for facilitating binary indexed databases for Uexküll will also have to be looked into. Term-document matrices of binary indexed databases are of low density (most 0s, few 1s), and methods that are vulnerable to deviation from normality will probably not be appropriate. (Schein, Saul & Ungar, 2003).

9.3.2 User evaluation of the IR interfaces

A very important aspect of the research around Uexküll is the evaluation through possible IR interfaces, which (as we have repeatedly argued) must take place with user participation. This section does not intend to present an evaluation design, rather discuss some challenges that such an evaluation entails. Needless to say, the evaluation presupposes that we have a working prototype of an Uexküll-based system.

Newby (2002) has evaluated a system that presents retrieval results laid out in 3D. The interface as it appears to the user resembles Uexküll, even though there are some important differences in how the visualized representations are brought about. Newby lists some challenges in evaluating IR visualization interfaces of that kind. Newby is particularly aware of the disadvantage of the visual interface when compared to a traditional, text based, IR-system that users, including test subjects, are well acquainted with. This, in the present author's opinion, is an important point to be aware of in the design of such a test. Below we are quoting Newby's list of challenges(p. 7):

- Generating prototype systems suitable for evaluation is time-consuming, and important features may be overlooked.
- While IR systems typically store many thousands or millions of documents, this scale is seldom achieved in prototype visualization systems.
- Human subjects are likely to be familiar with non-visual IR interfaces, but not visual interfaces. This could put the visual interface at a disadvantage, or create a need for extensive training.

- The benefits of human information navigation might not match usual IR performance measures such as recall and precision. For example, an information visualization might provide a better overview of a subject domain than a text based system, but this aspect might not be measured during an evaluation. In particular, visualization may be valuable for educating about an IR domain, yet this role does not fit well with typical relevance-based IR evaluation.

The above points point, particularly the last one, indicate the need of designing careful criteria of evaluation of such an approach, not only in case of success, but also in case of "failure" (meaning that the users are dissatisfied), so that appropriate lessons can be learned. It is not certain that "finding relevant documents" is the only criterion to be taken into account. Questionnaires designed for the purpose of such evaluation must also include questions that aim to uncover the extent to which a user gets a better overview of the material present in the collection through navigation in an Uexküll interface. In this context, the role of axis names, the diversity of access (see Subsection 7.5.3) and other factors would probably be interesting to gauge.

9.3.3 Other research directions to pursue

After discussing research directions that naturally continue the path of the present research, we would like to present some other directions of research that we regard worth following.

9.3.3.1 Elaborate axis naming schemes

Our treatment so far does not introduce a naming scheme for axes, apart from the simple naming scheme that assigns every concept axis the term that loads highest on it, and combined with the second highest as well in case of ties. This is a limited naming scheme, for a number of reasons:

- It does not guarantee us that axes carry names that users conceive as prominent,
- It does not guarantee us a unique name for each axis.

We believe that naming schemes will be a bottleneck on the road to making Uexküll based operational systems, and should be looked into. This is very largely a problem of the ambiguity of human language, and is related to the problem of representing documents by keywords.

9.3.3.2 Better presentation of concept axes

In Chapter 3 we discussed the implementation of the Uexküll approach. We presented a solution to the selection of access through lists that show all axis names. For large databases with many concept axes this may represent a significant increase in the cognitive load on the user.

One way of reducing cognitive load is trying to identify axes that have better interpretability, thereby reducing the available number of axes.

Another way of reducing the cognitive load would be to provide hierarchical access to the concept list. We have experimented with adding one more level of navigation to the Uexküll prototype. In the prototype that has been created of this system, an experiment has been done providing a hierarchic entry to the system, ordering the axes into groups. In this experiment the groups are created running the vectors representing the terms that give their names to concepts, through a self-organizing map (SOM) algorithm (Kohonen, 1995)[3] The regions represent groupings of the concepts. The user may choose a limited number of these groups, thereby including only the concepts he is interested in for choice of Uexküll groups. This method has only been experimented with on the user interface side, in an exploratory manner. We have unfortunately no results that can be used to evaluate the retrieval properties of it.

9.3.4 Open questions

We have so far discussed some information that our treatment seems to provide us with when evaluating the Uexküll approach. We have seen some strengths and some limitations, and proposed some other paths of research that may be fruitful. Upon closing this dissertation, we are left with a number important open questions that belong naturally within the scope of the current research.

1. To what extent are the results that we obtain indicative of the performance of a system implementing the approach?

2. We have not been able to show any significant differences between the behavior of our two measures of visualization support. To what extent is this a genuine attribute (or lack of attribute) of the approach itself, or an artefact of the evaluation approach? Would focus on specially selected queries be able to detect that?

[3]Our gratitude to Andi Rauber from the Technical University of Vienna for providing the Java implementation.

3. (particularly for indicated effects that are not verified statistically:) Would a different test collection render some of the indicated effects statistically significant?

4. What is the potential scalability of the approach? This question has a number of aspects, the most prominent of which are: (a) How large databases will a system implementing the approach be able to handle in the foreseeable future and (b) will such a system be able to provide manageable access to users, both in terms of navigation, and in providing them with the documents most relevant to their information needs?

These questions could partly be answered by the paths of research proposed in the previous sections if conducted, notably the user tests.

9.4 Conclusion

In this section we will try to provide answers to our research questions in light of the most important results and the discussion in the previous sections of this chapter. The following sections take the structure of the research question presented in Chapter 1.

9.4.1 Research question 1: interpretability of the LSI

> Given the definition of the Uexküll approach in page 14, and measured in terms of overall ranking and visual support, how well do traditional methods of data reduction, particularly LSI, support visualization based on the Uexküll approach? How interpretable are the axes of subspaces derived by LSI?

The simulation results presented in Chapter 8 confirm that raw, unprocessed LSI results have limited interpretation power, and therefore limited value for Uexküll-based user interaction. Using a varying number of dimensions, Figures 8.5, 8.9 and 8.12 indicate that higher dimensionality (and thereby a larger, more memory-consuming decomposition) may not pay off in increased support for navigation and retrieval. This is unlike results of LSI used in conventional vector space retrieval (retrieving by cosine similarity (Dumais, 1991; Deerwester et al., 1990)), that show a great impact of increasing the number of dimensions in the region within which we are operating (36-309 dimensions)[4].

[4]The projects referred to used collections comparable in size to ours

9.4.2 Research question 2: axis interpretability, rotations and dimensionality

> Measured in terms of overall ranking and visual separation, how well will rotation of LSI-derived data organizations support retrieval?

This general question was formulated as four sub questions discussed in the following subsections

9.4.2.1 Effect of rotation (2a)

> How do traditional methods of rotating axes influence the axis interpretability, and thereby the support, within the Uexküll approach, for retrieving relevant documents?

Our simulation results indicate that rotations, both orthogonal and oblique, improve performance. Both the orthogonal and the oblique rotations that we were using had a positive effect on retrieval parameters, both precision and recall.

Also when using measures that take into account visual separation, overall performance for rotated data organizations is superior.

The results in Chapter 8, and the summarizing discussion in Section 9.1 indicate that promax-rotated organizations are more prone to retrieval noise. We are not certain about the reason for this, but given that a part of the computation of this rotation is the raising of the loadings to a power of 2 to 4 (Abdi, 2003), it may be hypothesized whether this process changes the ranking of documents, as an artefact of the combined loading algorithm, thereby increasing retrieval noise.

9.4.2.2 Effect of dimensionality (2b)

> What is the effect of the dimensionality of the decompositions on the interpretability of the rotated spaces?

Our simulation results indicate that the dimensionality of the subspace comprising the data organization is an important factor when facilitating collections for Uexküll-based retrieval. For cosine based retrieval, increase in dimensionality is reported to have a positive effect on performance up to a certain point (Dumais, 1991). On the other hand, when using the same subspaces to facilitate Uexküll based retrieval, increase in dimensionality seems to have little or no effect when applied to raw, non-rotated spaces.

It is when rotating the axes that differences in axis interpretability among different dimensionalities become noticeable[5].

The performance improvements, due to rotations, is not equal for all decompositions, but it is consistent. At moderate recall values (40%), precision more than doubles. The same applies to the difference between various numbers of dimensions. At 40% recall, the difference between 158 and 36 dimensions is fourfold for the rotated than for the non-rotated version. Improvement is also noticeable regarding the usage distribution of axes.

9.4.2.3 Diversity of access (2c)

> To what extent will the methods presented here support the diversity of access-entries to presented material, where an access-entry is defined as a named axis? Would the data organizations enable different users to access materials processed and presented in an Uexküll based system through different access points?

From the results of our experiments, rotations increase the simulated usability of a greater number of the axes in constructing Uexküll groups. Whereas the non-rotated decompositions use a low portion of the axes (typically the axis corresponding to the largest singular value participates in more than half the queries), the distribution of the participation of axes as Uexküll group concepts becomes more even as we rotate. This effect may also be associated with other effects that user tests should take into account, as discussed in Subsection 9.3.2.

9.4.2.4 Novel measures and their significance (2d)

> What is the potential of the approach in helping the two types of readers defined by Foskett (1996) and quoted in Subsection 1.2.1

An important aspect of the current research has been the development of and the experiments with the two novel measures of performance, the SRE and the SRP. The SRP and SRE were supposed to facilitate the evaluation of data organization from the point of view of support they provide for visual retrieval. They should take into account the ranking, as well as the separation of documents.

In Subsection 9.1.4 we have discussed the fact that on the whole, our results do not detect significant differences in performance between the visualization

[5]Orthogonal rotations have no effect on retrieval based on vector similarity, and there is no reason to assume that oblique rotations will have any positive effect on such retrieval.

support measures and the ranking measures (represented by R-precision). But we have indicated the seemingly (not statistically confirmed) superiority of both the visualization support measures above R-precision in rewarding high (excessive) dimensional organizations. Further research needs to be done in order to confirm or reject this.

We also need more research regarding the potential of the visualization support measures to detect the superiority of a data organization when serving one of the user types over the other. The indication to the different behavior of the measure discussed in 9.1.4 is not sufficient for drawing conclusions beyond the formulation of a hypothesis.

9.4.3 Research questions 3: intellectually indexed collections

> What is the effectiveness of intellectually indexed data in supporting the retrieval of documents within the Uexküll approach?

Whereas automatically indexed databases are built of index terms extracted from the collections and often confounded and normalized into stems and other representations of the words, terms in intellectually indexed collections may also be assigned from other sources than the texts themselves, and are, at least in some cases, adapted to the context of the database. Those terms might make more meaningful axis names, and therefore better entry point for retrieval in an Uexküll based system on (otherwise) equal terms.

On the other hand, such index terms are more often than not dichotomously assigned. This, and the fact that we often have much fewer terms, means that our indexing carries less evidence as to relevance of documents to queries.

Not having an intellectually (by assignment) indexed version of our collection, we were fortunate to be able to use the original, manual indexing of the Cranfield collection, believing that for our needs it is a good approximation of an intellectually indexed database.

The analysis we performed on this collection was less thorough than on the automatically indexed collection, and was more adapted to the intellectual version. We were only opting at getting some indication as to the possibilities. Results for this version of the collection show the same tendencies as for the automatically indexed collection.

Even though the evidence of performance is not as clear and discriminatory for the manual indexing as it is for the automatic indexing of the collection,

results indicate that there is a potential of supporting Uexküll based retrieval from intellectually indexed databases with non-weighted (binary) indexing.

In the intellectual process of search by navigation along concept axes there is an advantage in having named axes associated with more pregnant terms, as intellectually assigned terms tend to be, but the question whether this advantage could outweigh the inferiority of the dichotomous term-document matrix as input data for multivariate transformations (both due to sparsity and due to the dichotomous indexing scheme), is open.

9.4.4 Research questions 4: the potential of simulations

> What are the advantages and the disadvantages of using simulations of the type performed in this project? To what extent do the simulations capture the properties of the approach, and what is the potential of using simulations in similar approaches?

As explained in Chapter 5, the evaluation approach, employed in this research, has had the purpose of finding data organizations that have a potential of serving the Uexküll approach as defined in page 14, and characterize the extent to which they support the approach.

The approach was meant to be general, meaning that information that was provided with the test collection we were using could be applicable to other data as well. To this end we were employing a test collection with a large number of queries, performing conservative statistical testing. Being strict in this way, placing demands on generality, it is natural that the level of detail in the conclusions we can draw from the approach must be limited. Phenomena that are not statistically significant cannot be verified and cannot be but indicated. And although the size of our test collection is also a factor that may burden the applicability of the results, we believe that our careful attitude counterbalances this burden, rendering our approach viable.

9.4.5 Concluding remarks

The traditional laboratory model has been challenged by newer approaches and models to information retrieval evaluation, and its applicability within the new paradigms of information seeking and retrieval has been questioned.

In this research we have tried to evaluate the potential usability of Uexküll data organizations, and in service of this purpose developed an evaluation

approach that broadens the role of the laboratory model, and makes it helpful also within more usability-oriented efforts.

The development of the Uexküll approach into working systems is still at a starting point, and there are still many unresolved issues. Some of these issues will be easier to resolve with the development of technology, and other issues will need a more direct follow-up.

References

Abdi, H. (2003). Factor rotations. In T. F. M. Lewis-Beck A. Bryman (Ed.), *Encyclopedia for research methods for the social sciences* (pp. 978–982). Thousand Oaks, CA: Sage.

Allan, J., Leouski, A. & Swan, R. (1997). *Interactive cluster visualization for information retrieval* (Tech. Rep. No. IR-116). Amherst, MA: Department of Computer Science, University of Massachusetts.

Ames, A. L. & Nadeau, D. R. (1996). *The vrml sourcebook.* New York: Wiley.

Ando, R. K. (2000). Latent semantic space: Iterative scaling improves inter-document similarity measurement. In N. J. Belkin, P. Ingwersen & M.-K. Leong (Eds.), *Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval* (pp. 216–223). Athens, Greece: ACM.

Ando, R. K. & Lee, L. (2001). Iterative residual rescaling: An analysis and generalization of lsi. In D. J. Croft W. Bruce ABD Harper, D. H. Kraft & J. Zobel (Eds.), *Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 154–162). New Orleans, LA.: ACM.

Anton, H. (1987). *Elementary linear algebra.* New York: John Wiley & sons.

Baeza-Yates, R. & Ribeiro-Neto, B. (1999). *Modern information retrieval.* New York: ACM Press.

Bartholomew, D. & Knott, M. (1999). *Latent variable models and factor analysis. - 2nd ed.* London: Arnold.

Belkin, N. J., Ingwersen, P. & Leong, M.-K. (Eds.). (2000). *Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval.* Athens, Greece: ACM.

Belkin, N. J., Oddy, R. & Brooks, H. (1982). Ask for information retrieval: Part i. background and theory. , *38*(2), 61–71.

Belkin, N. J., Seeger, T. & Wersig, G. (1983). Distributed expert problem treatment as a model of information system analysis and design.

Journal of Information Science, *5*, 153–167.

Benford, S., Snowdon, D., Greenhalgh, C., Ingram, R., Knox, I. & Brown, C. (1995). VR-VIBE: A virtual environment for Co-operative information retrieval. *Computer Graphics Forum*, *14*(3), 349–360.

Blom, K. (1999). *Information retrieval using the singular value decomposition and krylov subspaces*. Unpublished master's thesis, Chalmers University of Technology.

Blom, K. & Ruhe, A. (2001). Information retrieval using very short Krylov sequences. In M. W. Berry (Ed.), *Computational information retrieval* (Vol. 106, p. 41-56). Philadelphia, PA.: SIAM.

Borko, H. & Bernick, M. (1963). Automatic document classification. *Journal of the ACM*, *10*(2), 151–162.

Borlund, P. & Ingwersen, P. (1997). The development of a method for the evaluation of interactive information retrieval systems. , *53*(3), 225–250.

Börner, K., Chen, C. M. & Boyack, K. (2003). Visualizing knowledge domains. *Annual Review of Information Science and Technology*, *37*, 179–255.

Börner, K., Dillon, A. & Dolinsky, M. (2000, July). Lvis - digital library visualizer. *Information Visualisation 2000, Symposium on Digital Libraries*, 77–81.

Buckley, C. & Vorhees, E. M. (2000). Evaluating evaluation measure stability. In N. J. Belkin, P. Ingwersen & M.-K. Leong (Eds.), *Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval* (pp. 33–40). Athens, Greece: ACM.

Buckley, C. & Vorhees, E. M. (2004). Retrieval evaluation with incomplete information. In M. Sanderson, K. Järvelin, J. Allan & P. Bruza (Eds.), *Sigir 2004: Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval, sheffield, uk, july 25-29, 2004* (pp. 25–32). Sheffield, UK.: ACM.

C. J. van Rijsbergen. (1979). *Information retrieval. - 2nd ed.* London: Butterworths.

Card, S. K. (1996). Visualizing retrieved information: A survey. *IEEE Computer Graphics and Applications*, *16*(2), 63–67.

Chalmers, M. & Chitson, P. (1992, June). Bead: Explorations in information visualisation. In N. J. Belkin, P. Ingwersen & A. M. Pejtersen (Eds.), *Proceedings of the 15th annual international ACM SIGIR conference on research and development in information retrieval.* Copenagen, Denmark: ACM.

Cheng, X. & Dunkerton, T. (1995). Orthogonal rotation of spatial patterns

derived from singular value decomposition analysis. *Journal of Climate*, *8*, 2631-2643.

Cleverdon, C. W. (1972). On the inverse relationship of recall and precision. , *28*(3), 195–201.

Cleverdon, C. W., Mills, J. & Keen, M. (1966a). *Factors determining the performances of indexing systems: Volume i. design, part 1. appendices.* Cranfield, UK.: Aslib.

Cleverdon, C. W., Mills, J. & Keen, M. (1966b). *Factors determining the performances of indexing systems: Volume i. design, part 1. text.* Cranfield, UK.: Aslib.

Conover, W. J. (1980). *Practical nonparametric statistics. - 2nd ed.* New York: Wiley.

Cooper, W. S. (1968). Expected search length: A single measure of retrieval effectiveness based on weak ordering action of retrieval systems. *American Documentation*, *19*, 30–41.

Darányi, S., Zawiasa, R. & Hajnal, Z. (1996). Conceptual mapping of a database in the humanities: First results of an experiment with sophia. , *52*(1), 86–99.

Deerwester, S. et al. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407.

Dumais, S. T. (1991). Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments and Computers*, *23*(2), 229–236.

Dupret, G. (2003). Latent concepts and the number [of] orthogonal factors in latent semantic analysis. In C. Clarke, G. Cormack, J. Callan, D. Hawking & A. Smeaton (Eds.), *Proceedings of the 26th annual international ACM SIGIR conference on research and development in informaion retrieval* (pp. 221–226). Toronto, Canada: ACM.

Efron, M. (2005, July). Eigenvalue-based model selection during latent semantic indexing. *Journal of the American Society for Information Science and Technology*, *56*(9), 969–988.

Ellis, D. (1996a). The dilemma of measurement in information retrieval research. *Journal of the American Society for Information Science*, *47*(1), 23–36.

Ellis, D. (1996b). *Progress and problems in information retrieval. - 2nd ed.* London: Library Association.

Foskett, A. C. (1996). *The subject approach to information.* London: Library Association Publishing.

Fuhr, N. (2006). *Initiative for the evaluation of xml retrieval.* Retrieved 9 February, 2007 from the World Wide Web: `http://inex.is.informatik.uni-duisburg.de/`.

Furnas, G. W., Deerwester, S., Dumais, S. T., Landauer, T. K., Harshman, R., Streeter, L. A. et al. (1988). Information retrieval using a singular value decomposition model of latent semantic structure. In Y. Chiaramella (Ed.), *Sigir'88, proceedings of the 11th annual international ACM SIGIR conference on research and development in information retrieval, grenoble, france, june 13-15, 1988* (pp. 465–480). Grenoble, France: ACM.

Golub, G. H. & Van Loan, C. F. (1983). *Matrix computations.* Oxford: North Oxford Academic.

Gövert, N., Kazai, G., Fuhr, N. & Lalmas, M. (2003). *Evaluating the effectiveness of content-oriented xml retrieval* (Tech. Rep.). Dortmund: University of Dortmund, Department of Computer Science. Retrieved 23 March, 2004 from the World Wide Web: `http://www.is.informatik.uni-duisburg.de/bib/fulltext/ir/Goevert{$_$}etal:03a.pdf`.

Harter, S. P. (1996). Variations in relevance assessments and the measurement of retrieval effectiveness. *Journal of the American Society for Information Science*, *47*(1), 37–49.

He, X., Cai, D., Liu, H. & Ma, W.-Y. (2004). Locality preserving indexing for document representation. In M. Sanderson, K. Järvelin, J. Allan & P. Bruza (Eds.), *Sigir '04: Proceedings of the 27th annual international conference on research and development in information retrieval* (pp. 96–103). Sheffield, UK.: ACM Press.

Heine, M. D. (1981). Simulation, and simulation experiments. In K. Spärck Jones (Ed.), *Information retrieval experiment* (pp. 179–198). London: Butterworth.

Hoenkamp, E. (2003). Unitary operators on the document space. *Journal of the American Society for Information Science and Technology*, *54*(4), 314–320.

Hoffman, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval* (pp. 50–57). Berkeley, CA.: ACM.

Hull, D. (1993, June). Using statistical testing in the evaluation of retrieval experiments. In R. Korfhage, E. M. Rasmussen & P. Willett (Eds.), (pp. 329–338). Pittsburgh, PA.: ACM.

Hull, D. (1994, July). Improving text retrieval for the routing problem using latent semantic indexing. In W. B. Croft & C. J. van Rijsbergen (Eds.), *Proceedings of the 17th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 282–291). Dublin, Ireland: Springer-Verlag New York, Inc.

Ingwersen, P. (1992). *Information retrieval interaction.* London: Taylor

Graham Publishing.

Ingwersen, P. (1999). Cognitive information retrieval. *ANNUAL REVIEW OF INFORMATION SCIENCE AND TECHNOLOGY*, *34*, 3–52.

Ingwersen, P. & Järvelin, K. (2005). *The turn: integration of information seeking and retrieval in context*. Dordrecht: Springer.

International Organization for Standardization. (1997). *Information technology – computer graphics and image processing – the virtual reality modeling language (vrml) – part 1: Functional specification and utf-8 encoding.* (Tech. Rep. No. ISO/IEC 14772-1:1997). Geneva, Switzerland: International Organization for Standardization.

Jackson, J. E. (1991). *A users guide to principal components*. New York: Wiley.

Järvelin, K. & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, *20*(4), 422–446.

Kantor, P. B. (1994). Information retrieval techniques. *Annual Review of Information Science and Technology*, *29*, 53–90.

Keen, E. M. (1992). Presenting results of experimental retrieval comparisons. *Information Processing and Management*, *28*(4), 491–502.

Kekäläinen, J. (1999). *The effects of query complexity, expansion and structure on retrieval performance in probabilistic text retrieval*. Unpublished doctoral dissertation, University of Tampere.

Kekäläinen, J. & Järvelin, K. (2002a). Evaluating information retrieval systems under the challenges of interaction and multidimensional dynamic relevance. In H. Bruce, R. Fidel, P. Ingwersen & P. Vakkari (Eds.), *Proceedings of 4th CoLIS conference* (pp. 253–270). Greenwood Village, CO.: Libraries Unlimited.

Kekäläinen, J. & Järvelin, K. (2002b). Using graded relevance assessments in ir evaluation. *Journal of the American Society for Information Science and Technology*, *53*(13), 1120–1129. Available from `http://dx.doi.org/10.1002/asi.10137`

Kohonen, T. (1995). *Self-organizing maps*. Berlin: Springer.

Konig, R. (2002). On the rotation of non-linear principal components analysis (princals) solutions: Description of a procedure. *ZUMA-Nachrichten*, *26*(50), 114–120.

Lancaster, F. W. (1998). *Indexing and abstracting in theory and practice. - 2nd ed.* London: Library Association.

Landauer, T. K. & Dumais, S. T. (1997). Solution to plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, *104*(2), 211-240.

Landauer, T. K., Laham, D. & Foltz, P. (1997). Learning human-like knowl-

edge by singular value decomposition: a progress report. In *Nips '97: Proceedings of the 1997 conference on advances in neural information processing systems 10* (pp. 45–51). Cambridge, MA: MIT Press.

Landauer, T. K. et al. (2004). From paragraph to graph: Latent semantic analysis for information visualization. In *Proceedings of the national academy of sciences of the united states of america* (Vol. 101(Suppl. 1), p. 5214-5219). Washington, DC: National Academy of Sciences of the United States of America.

Leouski, A. & Allan, J. (1998). Visual interactions with a multidimensional ranked list. In W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson & J. Zobel (Eds.), *Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval* (pp. 353–354). Melbourne, Australia: ACM.

Lin, J. (2007). User simulations for evaluating answers to question series. *Information Processing and Management*, *43*(3), 717–729.

Losee, R. M. (1997). A discipline independent definition of information. *Journal of the American Society of Information Science*, *48*(3), 254-269.

Losee, R. M. (1998). *Text retrieval and filtering: Analytic models of performance*. Boston: Kluwer Academic Publishers.

Magennis, M. & van Rijsbergen, C. J. (1997). The potential and actual effectiveness of interactive query expansion. In N. J. Belkin, A. D. Narasimhalu & P. Willett (Eds.), *Sigir '97: Proceedings of the 20th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 324–332). Philadelphia, PA.: ACM Press.

Marcus, R. S. & Reintjes, J. F. (1981). A translating computer interface for end- user operation of heterogeneous retrieval systems; part i: Design. *Journal of the American Society for Information Science*, *32*(4), 287–303.

Mardia, K. V., Kent, J. T. & Bibby, J. (1979). *Multivariate analysis*. London: Academic Press.

Maron, M. E. & Kuhns, J. L. (1960). On relevance, probabilistic indexing and information retrieval. *J. ACM*, *7*(3), 216-244.

McGill, M. J. (1975). Projections within knowledge spaces: the implications for information storage and retrieval systems. In R. L. T. Charles W. Husbands (Ed.), *Proceedings of the 38th asis annual meeting* (pp. 138–139). Washington, D.C: Publications Division, American Society for Information Science.

McGill, M. J. (1976). Knowledge and information spaces: implications for retrieval systems. *Journal of the American Society for Information*

Science, *27*(4).

Meincke, P. P. M. & Atherton, P. (1976). Knowledge space: a conceptual basis for the organization of knowledge. *Journal of the American Society for Information Science*, *2*(1), 18–24.

Mizzaro, S. (1997). Relevance: The whole history. *Journal of the American Society of Information Science*, *48*(9), 810-832. Available from `citeseer.ist.psu.edu/mizzaro96relevance.html`

Muthen, B. (1989). Dichotomous factor analysis of symptom data. *Sociological Methods and Research*, *18*(1), 19–65.

Newby, G. B. (2002). Empirical study of 3d visualization information retrieval tasks. *Journal of Intelligent Information Systems*, *18*(1), 31–53.

Newell, A. (1969). Heuristic programming: ill-structured problems. In J. Aronofsky (Ed.), *Progress in operations research, vol. iii* (pp. 360–414). New York: John Wiley.

Nordlie, R. (2000). *User revealment*. Oslo, Norway: H⊘gskolen i Oslo, Avd. for journalistikk, bibliotek- og informasjonsfag.

Nuchprayoon, A. & Korfhage, R. (1994). a visual tool for retrieving documents. In *Proceedings of 1994 ieee symposium on visual languages*. St. Louis, MO.: IEEE Computer Society Press.

Oddy, R. N. (1977). Information retrieval through man-machine dialogue. , *33*(1), 1–14.

Olsen, K. A., Korfhage, R., Sochats, K. M., Spring, M. B. & Williams, J. (1993). Visualization of a document collection: The VIBE system. *Information Processing and Management*, *29*(1), 69-81.

Ossorio, P. G. (1966). Classification space: A multivariate procedure for automatic document indexing and retrieval. *Multivariate Behavioral Research*, 479–524.

Over, P. (1998). TREC-6 interactive track report. In *Proceedings of the 1998 text retrieval conference* (pp. 73–83). Gaithersburg, MD.: Department of Commerce, National Institute of Standards and Technology.

ParallelGraphics. (2004). Retrieved 29 March, 2004 from the World Wide Web: `http://www.parallelgraphics.com/developer/products/cortona/help`.

Parnas, S. (1994). *Latent semantisk indeksering*. Unpublished master's thesis, Oslo University College.

Pharo, N. (2002). *The sst method schema*. Unpublished doctoral dissertation, University of Tampere.

Pollitt, A. S. (1984). A front-end system: An expert system as an online search intermediary. In *Aslib proceedings* (Vol. 36, pp. 229–234). London: Aslib Institute of Information Scientists.

Preminger, M. & Darányi, S. (2000). Uexküll (demonstration session): an

interactive visual user interface for document retrieval in vector space. In N. J. Belkin, P. Ingwersen & M.-K. Leong (Eds.), *Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval.* Athens, Greece: ACM.

Raghavan, V. V. & Wong, S. K. M. (1986). A critical analysis of the vector space model for information retrieval. *Journal of the American Society for Information Science*, *37*(5), 279–287.

Rees, A. M. (1967). Evaluation of information systems and services. *Annual Review of Information Science and Technology*, *2*, 63–86.

Robertson, S. E. & Hancock-Beaulieu, M. M. (1992). On the evaluation of ir systems. *Information Processing and Management*, *28*(4), 457–466.

Rorvig, M. (1996). Minutes for participants in the ACM SIGIR '96 workshop "foundations of advanced information visualization for visual information (retrieval) systems". *Journal of the American Society for Information Science*, *51*(13), 1205–1210.

Sadeh, T. (2004). The challenge of metasearching. *New Library World*, *105*(1198/1199), 104–112.

Salton, G. & McGill, M. J. (1983). *Introduction to modern information retrieval.* New York: McGraw-Hill.

Sanderson, M., Järvelin, K., Allan, J. & Bruza, P. (Eds.). (2004). *Sigir 2004: Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval, sheffield, uk, july 25-29, 2004.* Sheffield, UK.: ACM.

Saracevic, T. & Kantor, P. (1988). A study of information seeking and retrieving. iii. searchers, searches and overlap. *Journal of the American Society for Information Science*, *39*(3), 197–216.

Schein, A. I., Saul, L. K. & Ungar, L. H. (2003). A generalized linear model for principal component analysis of binary data. In C. M. Bishop & B. J. Frey (Eds.), *Proceedings of the ninth international workshop on artificial intelligence and statistics.* Key West, FL.: Society for Artificial Intelligence and Statistics.

Shneiderman, B. (1998). *Designing the user interface. - 3rd ed.* Reading, MA.: Addison-Wesley.

Sormunen, E. (2000). *A method for measuring wide range performance of boolean queries in full-text databases.* Unpublished doctoral dissertation, University of Tampere.

Spärck Jones, K. (1976). Progress in documentation. , *32*(1), 59–75.

Spärck Jones, K. (1981). The cranfield tests. In K. Spärck Jones (Ed.), *Information retrieval experiment* (pp. 256–284). London: Butterworth.

Spärck Jones, K. (2000). Further reflections on TREC. *Information Processing and Management*, *36*(1), 37–85.

Spink, A., Greisdorf, H. & Bateman, J. (1998). From highly relevant to not relevant: Examining different regions of relevance. *Information Processing and Management*, *35*(5), 599–622.

Su, L. T. (1992). Evaluation measures for interactive information retrieval. *Information Processing and Management*, *28*(4), 503–516.

Tague-Sutcliffe, J. (1992). The pragmatics of information retrieval experimentation, revisited. *Information Processing and Management*, *28*(4), 467–490.

Tai, X., Ren, F. & Kita, K. (2002). An information retrieval model based on vector space method by supervised learning. *Information Processing and Management*, *38*(6), 749–764.

Taylor, R. (1968). Question negotiation and information seeking in libraries. *College & research libraries*, *29*, 178–194.

TREC. (2001). *Text retrieval conference.* Retrieved 20 December, 2001 from the World Wide Web: `http://trec.nist.gov/overview.html`.

Trefethen, L. N. & Bau, D. (1997). *Numerical linear algebra.* Philadelphia, Penn.: Society for Industrial and Applied Mathematics.

Vickery, B. C. (1966). *Cleverdon,c. - factors determining performance of indexing systems, volume 1, design* (Vol. 22) (Book Review No. 3).

Wersig, G. (1971). Der informationsbegriff. In *Information - kommunikation - dokumentation* (pp. 25–41). Munchen - Pullach - Berlin: Verlag Dokumentation.

White, H. D. & McCain, K. W. (1997). Visualization of literatures. *Annual Review of Information Science and Technolog*, *32*, 99–168.

White, R. W., Jose, J. M. & Ruthven, I. (2004). An implicit feedback approach for interactive information retrieval. *Information Processing and Management*, *42*, 166–190.

Williams, J. G., Sochats, K. M. & Morse, E. (1995). Visualization. *Annual Review of Information Science and Technology*, *30*, 161–207.

Wise, J. (1999). The ecological approach to text visualization. *Journal of the American Society for Information Science*, *50*(13), 1224–1233.

Zhang, J. (2001). Tofir: A tool of facilitating information retrieval - introduce a visual retrieval model. *Information Processing and Management*, *37*(4), 639–657.

Appendices

Appendix A

Behavior test for the SRP and SRE mesures

In Figures A.1 - A.9 we show the simulated scenes used to test the sanity of the SRP and SRE measures (see Subsections 7.4.3 and 7.4.4). In the captions, R and N denote the numbers of relevant and non-relevant documents in the scene, respectively. The values in the tables are all normalized loadings.

Table A.1: *Moderate layout of relevant and non-relevant documents for the test of the SRP and SRE measures, $R \sim N$*

moderate_good		moderate_med		moderate_bad	
Relevant loadings	irrelevant loadings	Relevant loadings	irrelevant loadings	Relevant loadings	irrelevant loadings
0,9	0,8	0,9	0,89	0,8	0,9
0,89	0,79	0,88	0,87	0,79	0,89
0,88	0,78	0,86	0,85	0,78	0,88
0,87	0,77	0,84	0,83	0,77	0,87
0,86	0,76	0,82	0,81	0,76	0,86
0,85	0,75	0,8	0,79	0,75	0,85
0,84	0,74	0,78	0,77	0,74	0,84
0,83	0,73	0,76	0,75	0,73	0,83
0,82	0,72	0,74	0,73	0,72	0,82
0,81		0,72			0,81

Table A.2: *Gappy layout of relevant and non-relevant documents for the test of the SRP and SRE measures, $R \sim N$*

gaps_good		gaps_med		gaps_bad	
Relevant loadings	irrelevant loadings	Relevant loadings	irrelevant loadings	Relevant loadings	irrelevant loadings
0,9	0,4	0,9	0,88	0,4	0,9
0,88	0,38	0,86	0,84	0,38	0,88
0,86	0,36	0,82	0,8	0,36	0,86
0,84	0,34	0,78	0,76	0,34	0,84
0,82	0,32	0,74	0,72	0,32	0,82
0,8	0,3	0,4	0,38	0,3	0,8
0,78	0,28	0,36	0,34	0,28	0,78
0,76	0,26	0,32	0,3	0,26	0,76
0,74	0,24	0,28	0,26	0,24	0,74
0,72		0,24			0,72

Table A.3: *Steep layout of relevant and non-relevant documents for the test of the SRP and SRE measures, $R \sim N$.*

steep_good		steep_med		steep_bad	
Relevant loadings	irrelevant loadings	Relevant loadings	irrelevant loadings	Relevant loadings	irrelevant loadings
0,9	0,4	0,9	0,85	0,4	0,9
0,85	0,35	0,8	0,75	0,35	0,85
0,8	0,3	0,7	0,65	0,3	0,8
0,75	0,25	0,6	0,55	0,25	0,75
0,7	0,2	0,5	0,45	0,2	0,7
0,65	0,15	0,4	0,35	0,15	0,65
0,6	0,11	0,3	0,25	0,1	0,6
0,55	0,06	0,2	0,15	0,06	0,55
0,5		0,1	0,06		0,5
0,45		0			0,45

Table A.4: *Moderate layout of relevant and non-relevant documents for the test of the SRP and SRE measures, $R >> N$*

moderate_good		moderate_med		moderate_bad	
Relevant loadings	nonrelevant loadings	Relevant loadings	nonrelevant loadings	Relevant loadings	nonrelevant loadings
0,9	0,8	0,9	0,89	0,88	0,9
0,89	0,79	0,88	0,81	0,87	0,89
0,88		0,87		0,86	
0,87		0,86		0,85	
0,86		0,85		0,84	
0,85		0,84		0,83	
0,84		0,83		0,82	
0,83		0,82		0,81	
0,82		0,8		0,8	
0,81		0,79		0,79	

Table A.5: *Gappy layout of relevant and non-relevant documents for the test of the SRP and SRE measures. $R >> N$*

gaps_good		gaps_med		gaps_bad	
Relevant loadings	nonrelevant loadings	Relevant loadings	nonrelevant loadings	Relevant loadings	nonrelevant loadings
0,9	0,4	0,9	0,88	0,86	0,9
0,88	0,24	0,86	0,72	0,84	0,88
0,86		0,84		0,82	
0,84		0,8		0,8	
0,82		0,82		0,78	
0,8		0,78		0,76	
0,78		0,76		0,74	
0,76		0,74		0,72	
0,74		0,4		0,4	
0,72		0,24		0,24	

Table A.6: *Steep layout of relevant and non-relevant documents for the test of the SRP and SRE measures, $R >> N$*

steep_good		steep_med		steep_bad	
Relevant loadings	nonrelevant loadings	Relevant loadings	nonrelevant loadings	Relevant loadings	nonrelevant loadings
0,9	0,4	0,9	0,85	0,8	0,9
0,85	0,35	0,8	0,45	0,75	0,85
0,8		0,75		0,7	
0,75		0,7		0,65	
0,7		0,65		0,6	
0,65		0,6		0,55	
0,6		0,55		0,5	
0,55		0,5		0,45	
0,5		0,4		0,4	
0,45		0,35		0,35	

Table A.7: *Moderate layout of relevant and non-relevant documents for the test of the SRP and SRE measures, $N >> R$*

moderate_good		moderate_med		moderate_bad	
Relevant loadings	irrelevant loadings	Relevant loadings	irrelevant loadings	Relevant loadings	irrelevant loadings
0,9	0,88	0,9	0,89	0,81	0,9
0,89	0,87	0,8	0,88	0,8	0,89
	0,86		0,87		0,88
	0,85		0,86		0,87
	0,84		0,85		0,86
	0,83		0,84		0,85
	0,82		0,83		0,84
	0,81		0,82		0,83
	0,8		0,81		0,82

Table A.8: *Gappy layout of relevant and non-relevant documents for the test of the SRP and SRE measures, $N >> R$*

gaps_good		gaps_med		gaps_bad	
Relevant loadings	irrelevant loadings	Relevant loadings	irrelevant loadings	Relevant loadings	irrelevant loadings
0,9	0,4	0,9	0,72	0,26	0,9
0,72	0,38	0,24	0,4	0,24	0,72
	0,36		0,38		0,4
	0,34		0,36		0,38
	0,32		0,34		0,36
	0,3		0,32		0,34
	0,28		0,3		0,32
	0,26		0,28		0,3
	0,24		0,26		0,28

Table A.9: *Steep layout of relevant and non-relevant documents for the test of the SRP and SRE measures, $N >> R$*

steep_good		steep_med		steep_bad	
Relevant loadings	irrelevant loadings	Relevant loadings	irrelevant loadings	Relevant loadings	irrelevant loadings
0,9	0,4	0,9	0,45	0,11	0,9
0,45	0,35	0,06	0,4	0,06	0,45
	0,3		0,35		0,4
	0,25		0,3		0,35
	0,2		0,25		0,3
	0,15		0,2		0,25
	0,11		0,15		0,2
	0,06		0,11		0,15

Appendix B

Result details for summary measures

Average results and Friedman pairwise test statistics of significance (as described in Section 7.6). The result are presented by scenarios. The row and column headings follow the format $< rotation >< dimensionality >$, where for rotation V=varimax, P=promax and N=non-rotated.

B.1 Simulation scenario 1

B.1.1 Significance statistics for summary measures: sum model

Table B.1: *Summary of differences between rank-sums of 12 organizations, R-precision measure (sum model). The bold-faced entries represent significant difference (critical value 134.2). The organizations are sorted from best to poorest.*

org	rp1d	V158	V309	P158	V75	P75	V36	P36	P309	N75	N309	N158	N36
24,14	Mean result	0,262	0,245	0,228	0,207	0,186	0,144	0,142	0,148	0,129	0,126	0,123	0,11
> 134,2	Σ ranks	1823	1786	1656	1594	1493	1420	1387	1360	1281	1269	1261	1224
V158	1823	0	37	***167***	***229***	***330***	***403***	***436***	***463***	***542***	***554***	***562***	***599***
V309	1786		0	130	***192***	***293***	***366***	***399***	***426***	***505***	***517***	***525***	***562***
P158	1656			0	61,5	***163***	***236***	***269***	***296***	***375***	***387***	***395***	***432***
V75	1594				0	101,5	***175***	***208***	***234***	***314***	***325***	***333***	***371***
P75	1493					0	73	106	132,5	***212***	***224***	***232***	***269***
V36	1420						0	33	59,5	***139***	***151***	***159***	***196***
P36	1387							0	26,5	106	117,5	125,5	***163***
P309	1360								0	79,5	91	99	***137***
N75	1281									0	11,5	19,5	57
N309	1269										0	8	45,5
N158	1261											0	37,5
N36	1224												0

Table B.2: *Summary of differences between rank-sums of 12 organizations, SRP measure (sum model). The bold-faced entries represent significant differences (critical value 149.2)*

org	srp1d	V309	V158	P158	V75	P75	P309	V36	P36	N75	N158	N309	N36
54,29	Mean result	0,167	0,143	0,123	0,091	0,083	0,102	0,05	0,045	0,045	0,043	0,044	0,034
> 149,2	Σ ranks	2020	1984	1823	1629	1543	1514	1354	1281	1129	1126	1118	1031
V309	2020	0	35,5	***197***	***390,5***	***477***	***505,5***	***665,5***	***738,5***	***890,5***	***894***	***901,5***	***988,5***
V158	1984		0	***161,5***	***355***	***441,5***	***470***	***630***	***703***	***855***	***858,5***	***866***	***953***
P158	1823			0	***193,5***	***280***	***308,5***	***468,5***	***541,5***	***693,5***	***697***	***704,5***	***791,5***
V75	1629				0	86,5	115	***275***	***348***	***500***	***503,5***	***511***	***598***
P75	1543					0	28,5	***188,5***	***261,5***	***413,5***	***417***	***424,5***	***511,5***
P309	1514						0	***160***	***233***	***385***	***388,5***	***396***	***483***
V36	1354							0	73	***225***	***228,5***	***236***	***323***
P36	1281								0	***152***	***155,5***	***163***	***250***
N75	1129									0	3,5	11	98
N158	1126										0	7,5	94,5
N309	1118											0	87
N36	1031												0

Table B.3: *Summary of differences between rank-sums of 12 organizations, SRE measure (sum model). The bold-faced entries represent significant differences (Critical value 148.0)*

org	se1d	V309	V158	P158	V75	P75	P309	V36	P36	N309	N158	N75	N36
58,44	Mean result	0,27	0,239	0,225	0,178	0,173	0,178	0,133	0,13	0,116	0,114	0,112	0,097
> 148,049756844334	Σ ranks	2069	1992	1852	1610	1524	1466	1336	1303	1134	1126	1125	1016
V309	2069	0	77	***217***	***459***	***545***	***603***	***733***	***766***	***935***	***943***	***944***	***1053***
V158	1992		0	139,5	***382***	***468***	***526***	***656***	***689***	***858***	***866***	***867***	***976***
P158	1852			0	***243***	***328***	***387***	***516***	***549***	***719***	***727***	***727***	***836***
V75	1610				0	85,5	144	***274***	***307***	***476***	***484***	***485***	***594***
P75	1524					0	58,5	***188***	***221***	***391***	***399***	***399***	***508***
P309	1466						0	129,5	***163***	***332***	***340***	***341***	***450***
V36	1336							0	33	***203***	***211***	***211***	***320***
P36	1303								0	***170***	***178***	***178***	***287***
N309	1134									0	8	8,5	117,5
N158	1126										0	0,5	109,5
N75	1125											0	109
N36	1016												0

B.1.2 Significance statistics for summary measures: max model

Table B.4: *Summary of differences between rank-sums of 12 organizations, R-precision measure (max model). The bold-faced entries represent significant difference (Critical value 124.6). The organizations are sorted from best to poorest.*

org	rp1d	V309	V158	P158	V75	P309	P75	V36	P36	N75	N158	N309	N36
21,23	Mean results	0,178	0,176	0,149	0,129	0,11	0,113	0,089	0,088	0,066	0,066	0,062	0,059563
> 124,6	Σ ranks	1778	1759	1623	1551	1467	1463	1398	1380	1295	1294	1275	1269,5
V309	1778	0	19,5	***155***	***228***	***312***	***315***	***381***	***399***	***483***	***484***	***503***	***508,5***
V158	1759		0	***136***	***208***	***292***	***296***	***361***	***379***	***464***	***465***	***484***	***489***
P158	1623			0	72,5	***157***	***160***	***226***	***244***	***328***	***329***	***348***	***353,5***
V75	1551				0	84	87,5	***153***	***171***	***256***	***257***	***276***	***281***
P309	1467					0	3,5	69	87	***172***	***173***	***192***	***197***
P75	1463						0	65,5	83,5	***168***	***169***	***188***	***193,5***
V36	1398							0	18	102,5	103,5	122,5	***128***
P36	1380								0	84,5	85,5	104,5	110
N75	1295									0	1	20	25,5
N158	1294										0	19	24,5
N309	1275											0	5,5
N36	1270												0

Table B.5: *Summary of differences between rank-sums of 12 organizations, SRP measure (max model). The bold-faced entries represent significant differences (Critical value 148.6)*

org	srp1d	V158	V309	P158	V75	P309	P75	P36	V36	N75	N158	N309	N36
49,91	Mean result	0,084	0,107	0,072	0,055	0,073	0,046	0,032	0,031	0,019	0,019	0,019	0,019
> 148,6	Σ ranks	1959	1958	1756	1686	1548	1532	1381	1366	1104	1103	1085	1076
V158	1959	0	1	***203***	***272,5***	***411***	***426,5***	***577,5***	***593***	***855***	***856***	***874***	***882,5***
V309	1958		0	***202***	***271,5***	***410***	***425,5***	***576,5***	***592***	***854***	***855***	***873***	***881,5***
P158	1756			0	69,5	***208***	***223,5***	***374,5***	***390***	***652***	***653***	***671***	***679,5***
V75	1686				0	138,5	***154***	***305***	***320,5***	***582,5***	***583,5***	***601,5***	***610***
P309	1548					0	15,5	***166,5***	***182***	***444***	***445***	***463***	***471,5***
P75	1532						0	***151***	***166,5***	***428,5***	***429,5***	***447,5***	***456***
P36	1381							0	15,5	***277,5***	***278,5***	***296,5***	***305***
V36	1366								0	***262***	***263***	***281***	***289,5***
N75	1104									0	1	19	27,5
N158	1103										0	18	26,5
N309	1085											0	8,5
N36	1076												0

Table B.6: *Summary of differences between rank-sums of 12 organizations, SRE measure (max model). The bold-faced entries represent significant differences (Critical value 148.16)*

org	se1d	V158	V309	P158	V75	P75	P309	V36	P36	N75	N158	N309	N36
51,08	Mean result	0,187	0,208	0,171	0,138	0,134	0,136	0,095	0,097	0,077	0,078	0,078	0,072
> 148,16	Σ ranks	1985	1984	1806	1610	1575	1459	1354	1349	1125	1124	1110	1073
V158	1985	0	0,5	***179***	***375***	***410***	***526***	***631***	***636***	***860***	***861***	***875***	***912***
V309	1984		0	***179***	***375***	***410***	***526***	***630***	***636***	***859***	***861***	***875***	***911***
P158	1806			0	***196***	***231***	***347***	***452***	***457***	***681***	***682***	***696***	***733***
V75	1610				0	35	***151***	***256***	***261***	***485***	***486***	***500***	***537***
P75	1575					0	116	***221***	***226***	***450***	***451***	***465***	***502***
P309	1459						0	104,5	110	***334***	***335***	***349***	***386***
V36	1354							0	5,5	***229***	***231***	***245***	***281***
P36	1349								0	***224***	***225***	***239***	***276***
N75	1125									0	1,5	15,5	52
N158	1124										0	14	50,5
N309	1110											0	36,5
N36	1073												0

B.1.3 Behavior curves for summary measures: max model

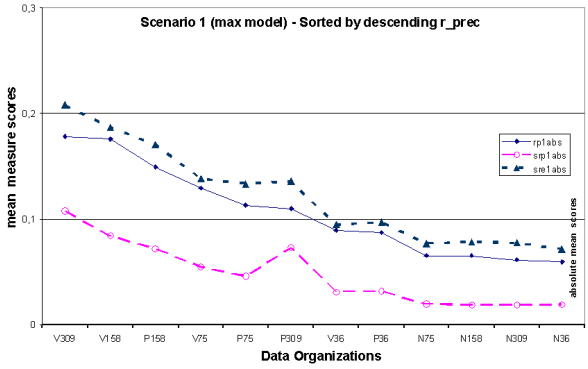


Figure B.1: *Simulation scenario 1 (max model): absolute scores and rank sums ordered by descending R-precision.*

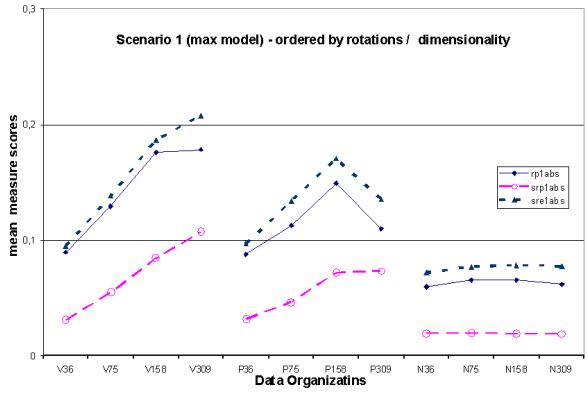


Figure B.2: *Simulation scenario 1 (max model): absolute scores and rank sums ordered by rotations/dimensionalities.*

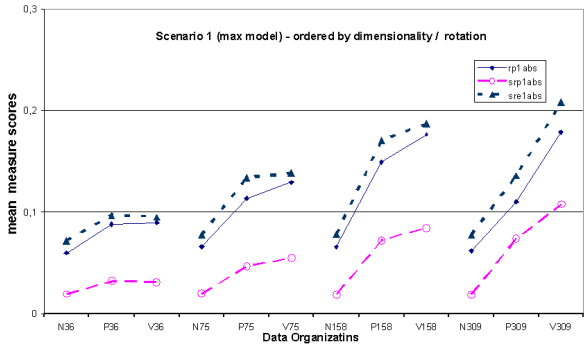


Figure B.3: *Simulation scenario 1 (max model): absolute scores and rank sums ordered by rotations/dimensionalities.*

Appendix C

Recall properties

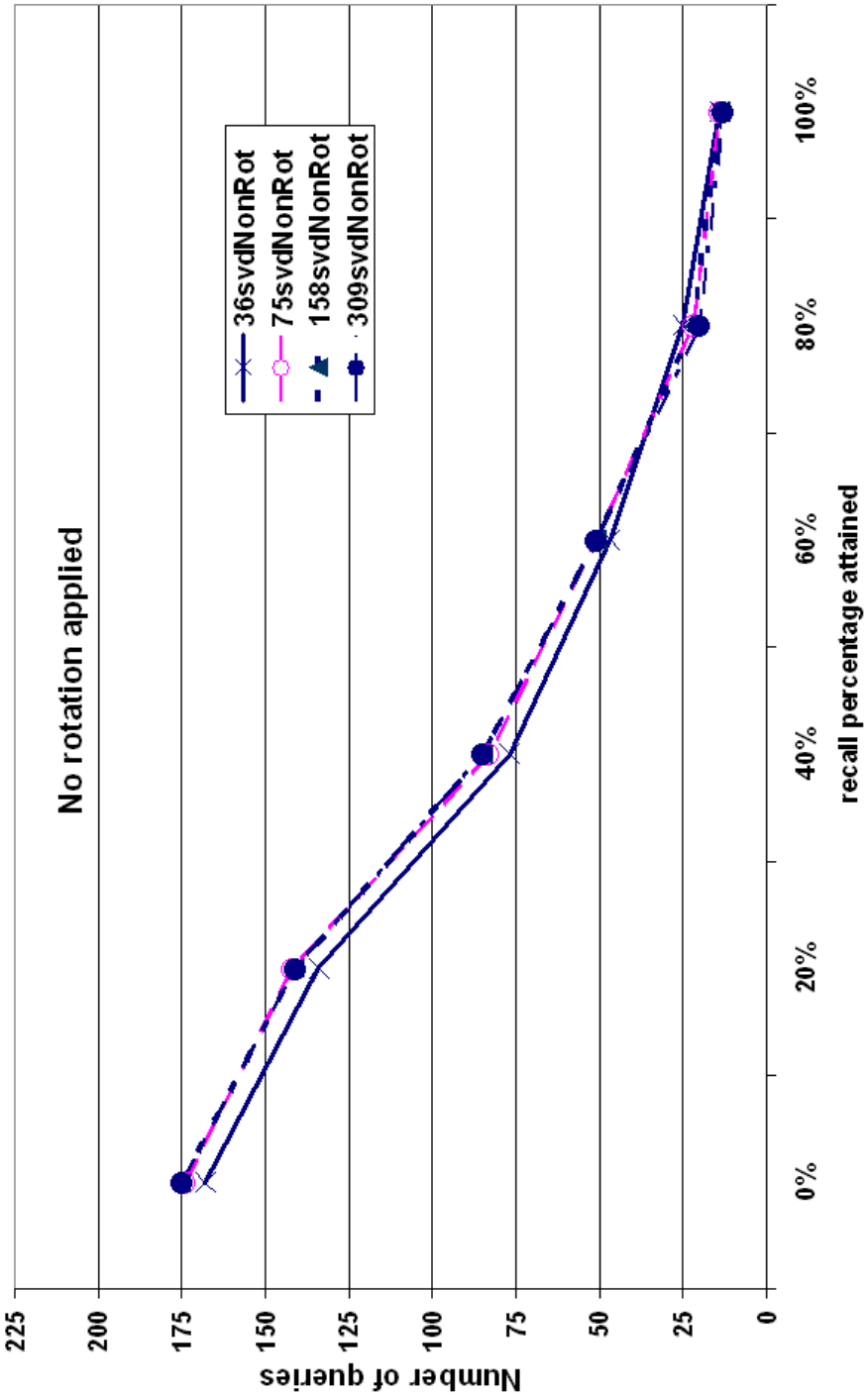


Figure C.1: *Simulation scenario 1: numbers of queries attaining different values of recall: non-rotated SVD, varying dimensionality.*

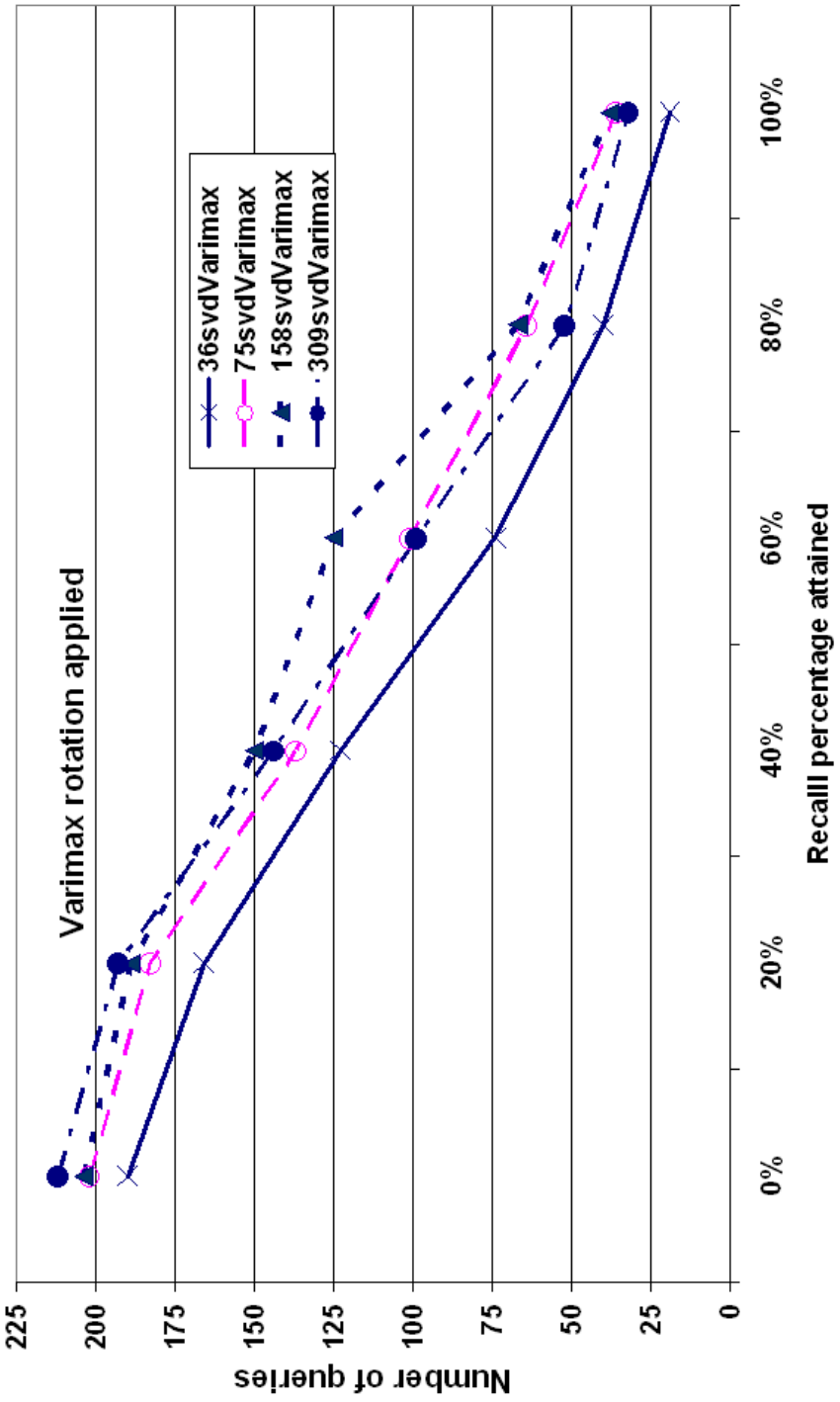


Figure C.2: *Simulation scenario 1: numbers of queries attaining different values of recall: varimax-rotated SVD, varying dimensionality.*

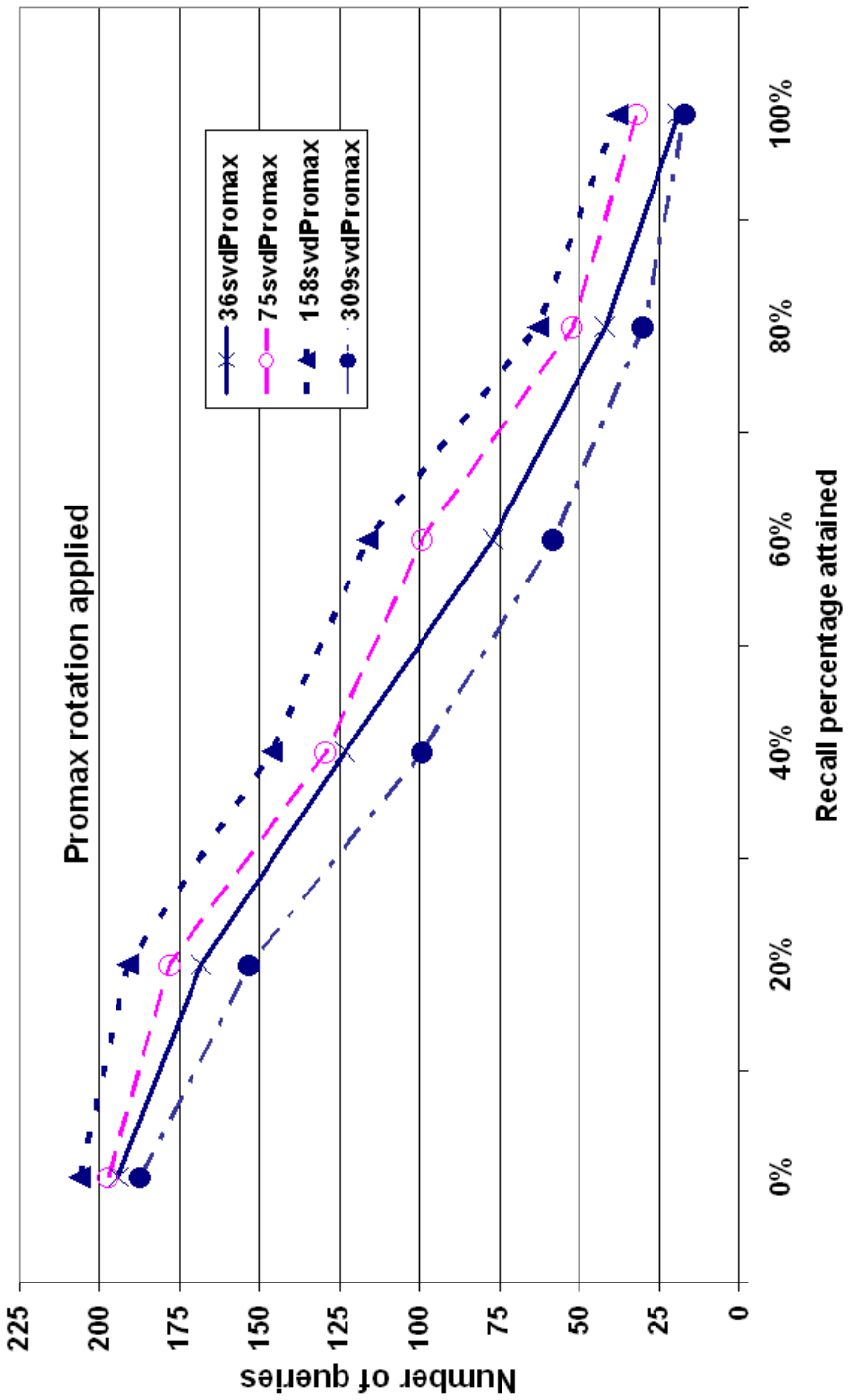


Figure C.3: *Simulation scenario 1: numbers of queries attaining different values of recall: promax-rotated SVD, varying dimensionality.*

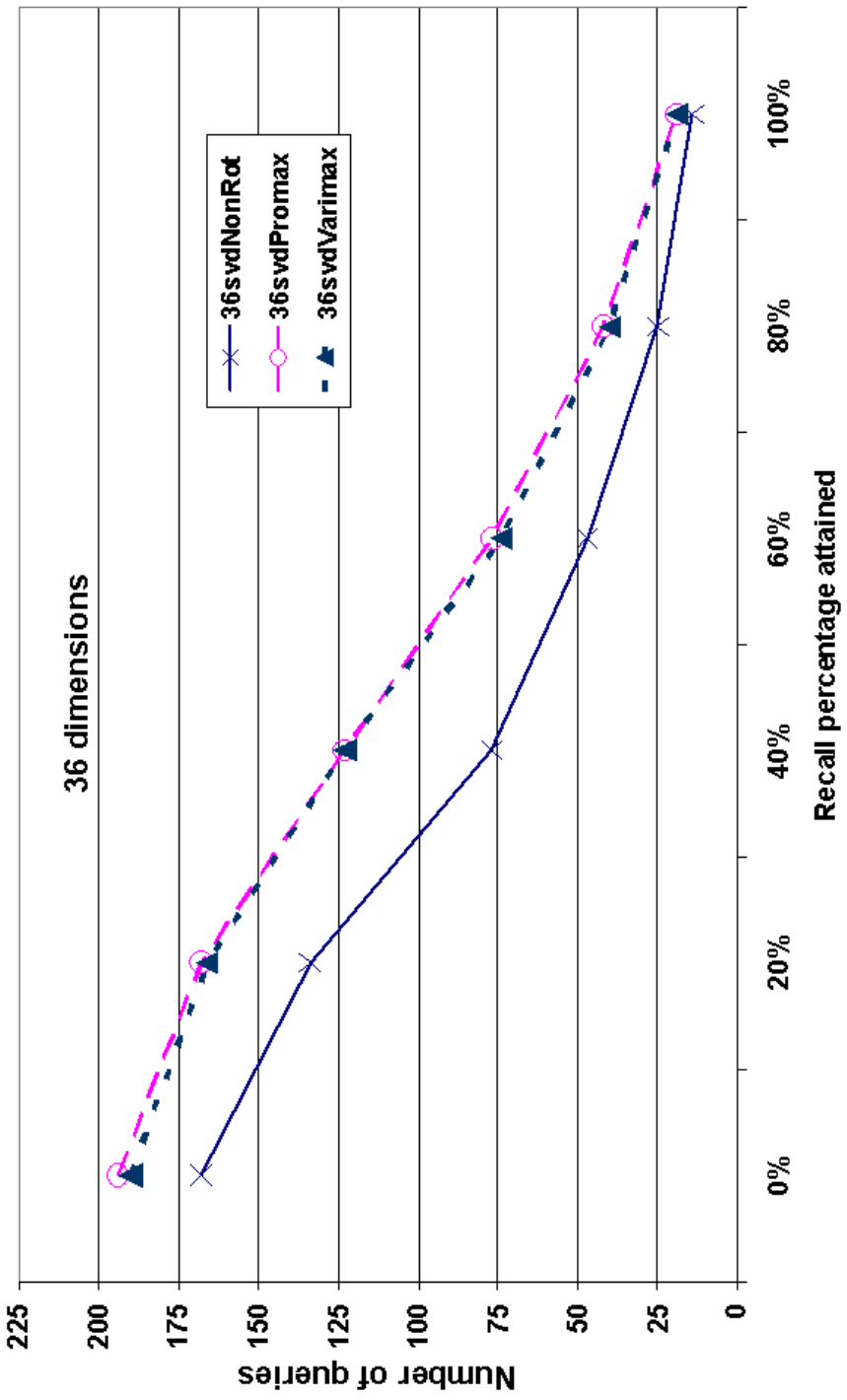


Figure C.4: *Numbers of queries (out of 225) attaining prescribed measures of recall for all the 36 dimensional organizations*

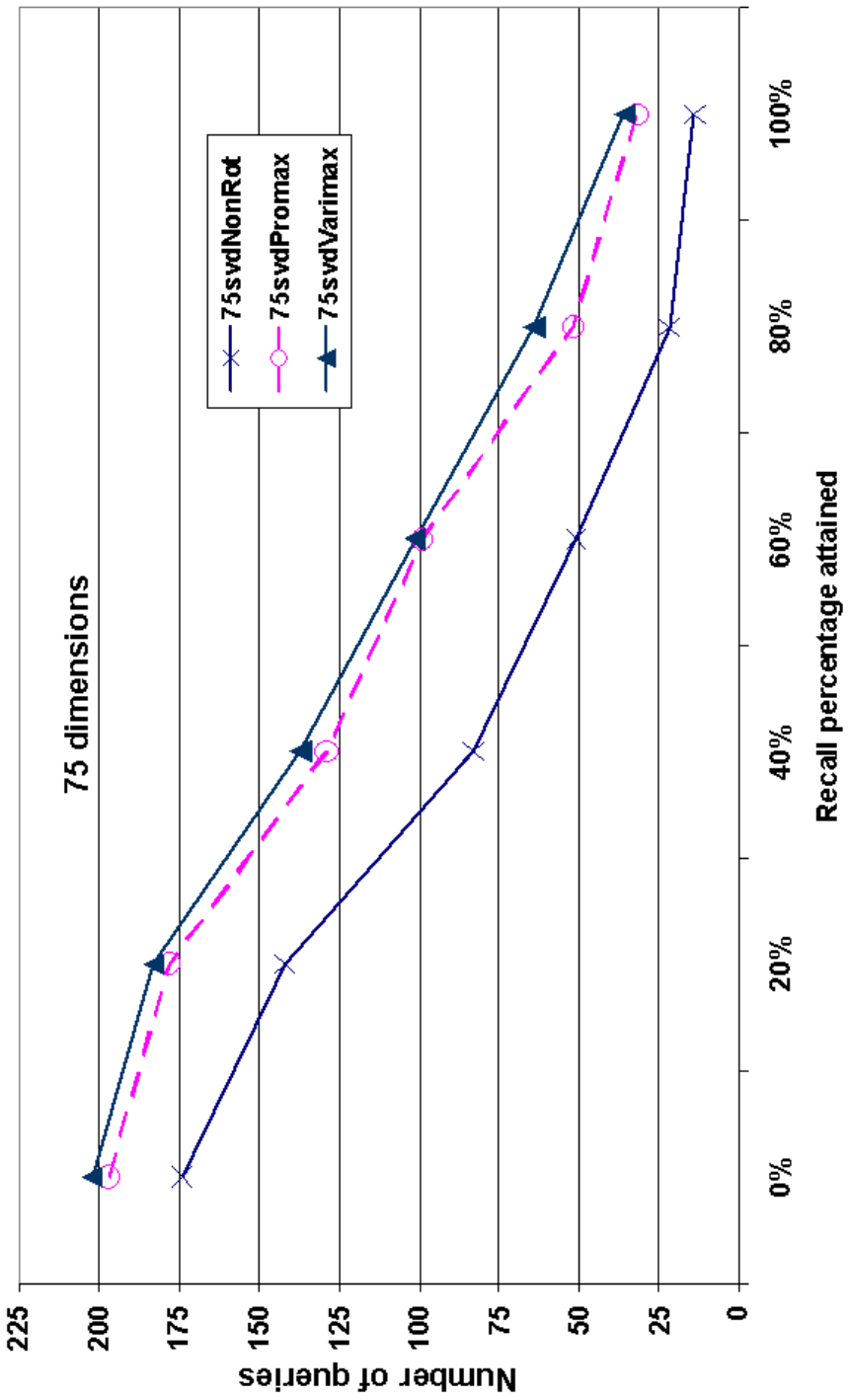


Figure C.5: *Numbers of queries (out of 225) attaining prescribed measures of recall for all the 75 dimensional organizations*

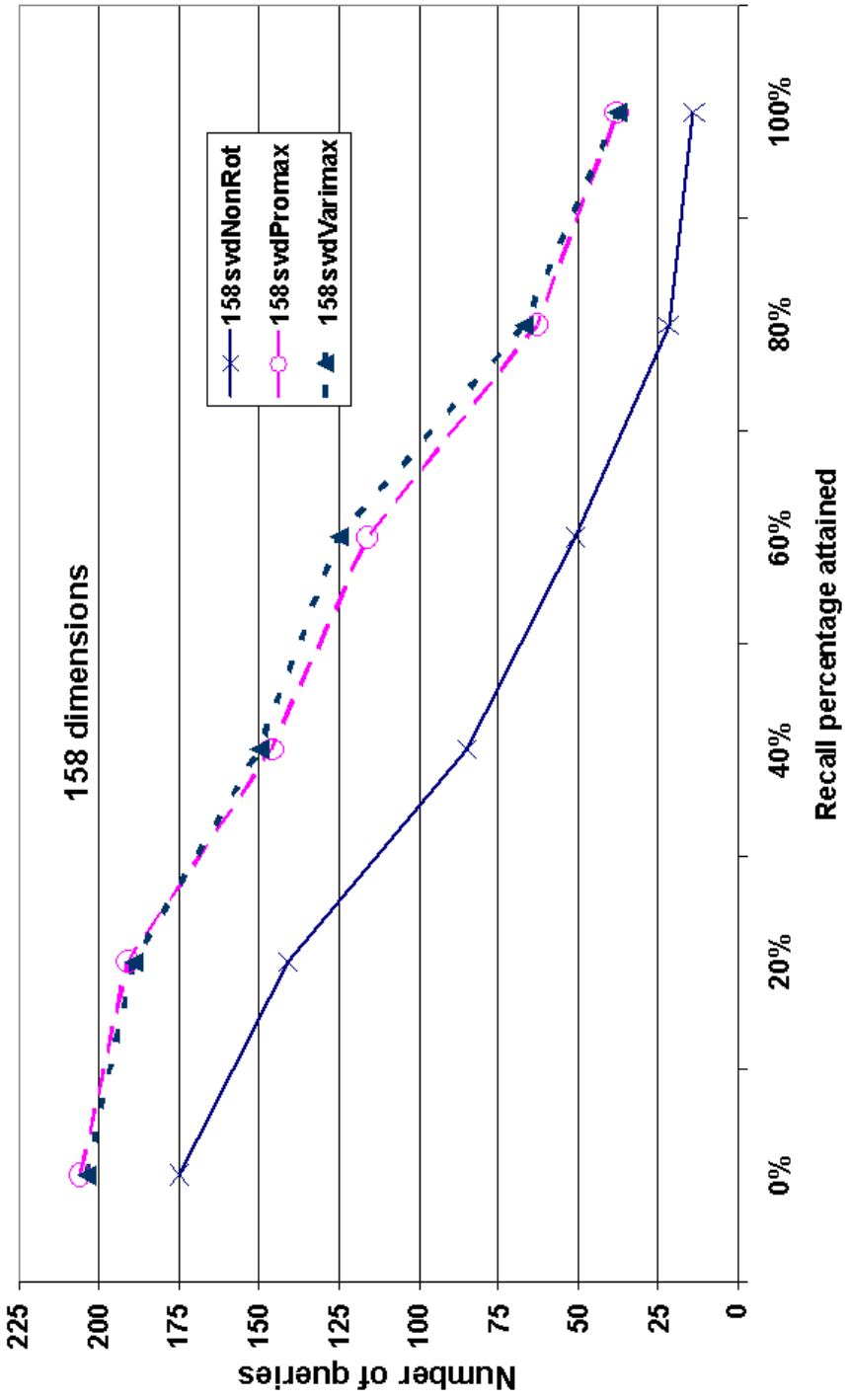


Figure C.6: *Numbers of queries (out of 225) attaining prescribed measures of recall for all the 158 dimensional organizations*

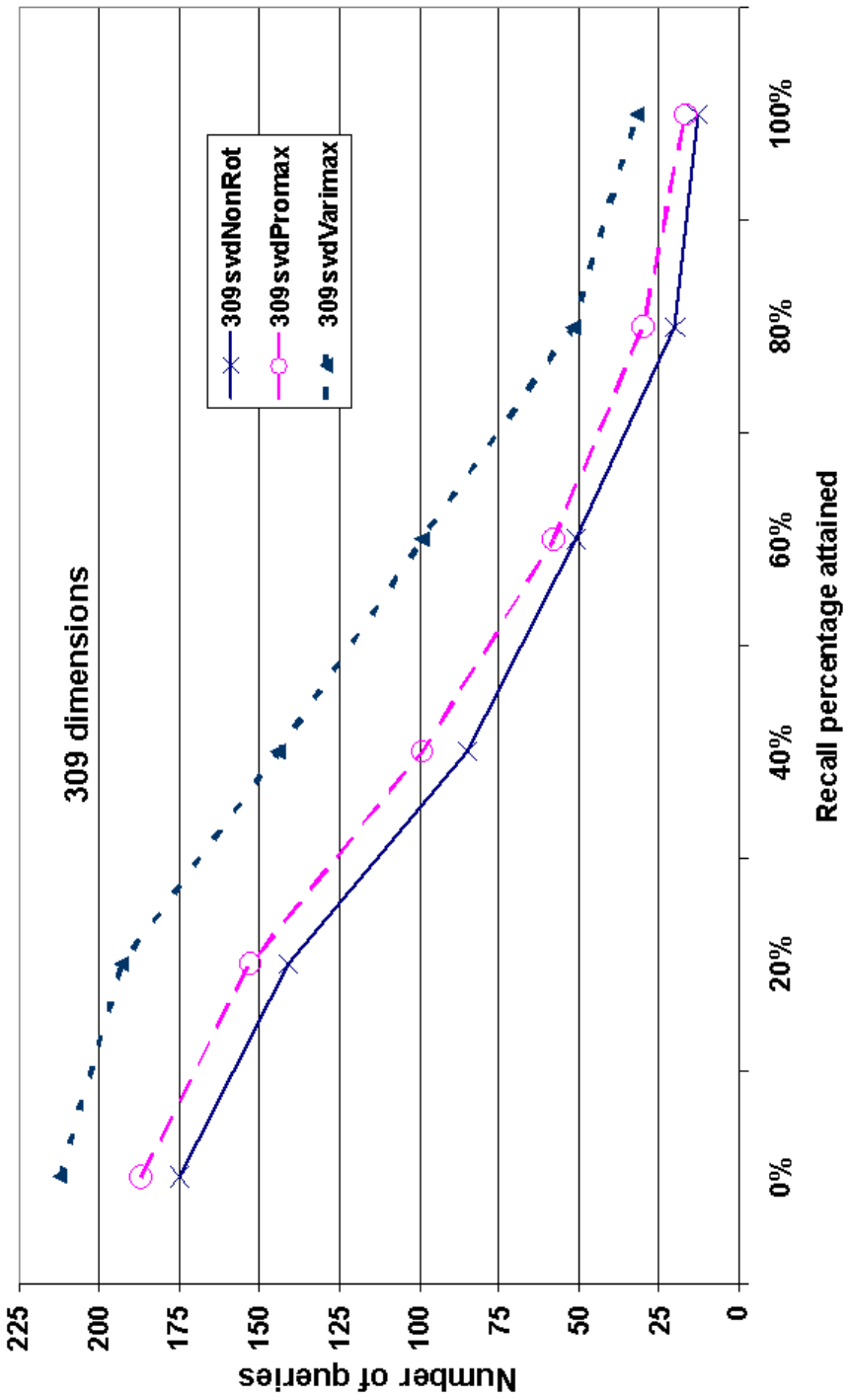


Figure C.7: *Numbers of queries (out of 225) attaining prescribed measures of recall for all the 309 dimensional organizations*

Appendix D

Simulation scenario 2

Table D.1: *Simulation scenario 2: summary of differences between rank-sums of 12 organizations, R-precision measure. The bold-faced entries represent significant difference (Critical value 134.9). The organizations are sorted from best to poorest.*

org	rp2d	V158	V309	P158	V75	P75	P309	P36	V36	N158	N309	N75	N36
29,08	Mean result	0,204	0,198	0,198	0,172	0,172	0,15	0,121	0,12	0,093	0,079	0,082	0,067
> 134,9	Σ ranks	1779	1745	1698	1632	1590	1483	1433	1414	1259	1200	1190	1130
V158	1779	0	33,5	81	***147***	***189***	***296***	***346***	***365***	***520***	***579***	***589***	***648,5***
V309	1745		0	47,5	113,5	***155***	***263***	***313***	***331***	***486***	***546***	***555***	***615***
P158	1698			0	66	107,5	***215***	***265***	***284***	***439***	***498***	***508***	***567,5***
V75	1632				0	41,5	***149***	***199***	***218***	***373***	***432***	***442***	***501,5***
P75	1590					0	107,5	***158***	***176***	***331***	***391***	***400***	***460***
P309	1483						0	50	68,5	***224***	***283***	***293***	***352,5***
P36	1433							0	18,5	***174***	***233***	***243***	***302,5***
V36	1414								0	***155***	***215***	***224***	***284***
N158	1259									0	59,5	69	129
N309	1200										0	9,5	69,5
N75	1190											0	60
N36	1130												0

Table D.2: *Simulation scenario 2: summary of differences between rank-sums of 12 organizations, SRP measure. The bold-faced entries represent significant differences (Critical value 142.0)*

org	srp2d	V309	P158	V158	P309	V75	P75	P36	V36	N158	N309	N75	N36
85,89	Mean result	0,104	0,101	0,086	0,101	0,062	0,065	0,036	0,037	0,019	0,017	0,017	0,015
> 142,0	Σ ranks	2033	1907	1903	1779	1677	1641	1363	1351	1080	991	979	849,5
V309	2033	0	125,5	130	***254***	***356***	***392***	***670***	***682***	***953***	***1042***	***1054***	***1183***
P158	1907		0	4,5	128,5	***231***	***266***	***545***	***557***	***828***	***916***	***928***	***1058***
V158	1903			0	124	***226***	***262***	***540***	***552***	***823***	***912***	***924***	***1053***
P309	1779				0	102	137,5	***416***	***428***	***699***	***788***	***800***	***929***
V75	1677					0	35,5	***314***	***326***	***597***	***686***	***698***	***827***
P75	1641						0	***279***	***291***	***562***	***650***	***662***	***792***
P36	1363							0	12	***283***	***372***	***384***	***513***
V36	1351								0	***271***	***360***	***372***	***501***
N158	1080									0	88,5	100,5	***230***
N309	991										0	12	141,5
N75	979											0	129,5
N36	849,5												0

Table D.3: *Simulation scenario 2: summary of differences between rank-sums of 12 organizations, SRE measure. The bold-faced entries represent significant differences (Critical value 134.4)*

org	se2d	V309	P158	V158	P309	P75	V75	P36	V36	N158	N309	N75	N36
121,3	Mean result	0,22	0,196	0,177	0,187	0,145	0,137	0,099	0,099	0,061	0,058	0,056	0,046
> 134,4 54475 23726 9	Σ ranks	2137	1979	1932	1792	1673	1657	1398	1354	991,5	925	910	804
V309	2137	0	***158***	***205***	***345***	***464***	***480***	***739***	***783***	***1145***	***1212***	***1227***	***1333***
P158	1979		0	46,5	***187***	***306***	***322***	***581***	***625***	***987***	***1054***	***1069***	***1175***
V158	1932			0	***140***	***260***	***276***	***535***	***578***	***941***	***1007***	***1022***	***1128***
P309	1792				0	119,5	***136***	***395***	***438***	***801***	***867***	***882***	***988***
P75	1673					0	16	***275***	***319***	***681***	***748***	***763***	***869***
V75	1657						0	***259***	***303***	***665***	***732***	***747***	***853***
P36	1398							0	43,5	***406***	***473***	***488***	***594***
V36	1354								0	***363***	***429***	***444***	***550***
N158	991,5									0	66,5	81,5	***188***
N309	925										0	15	121
N75	910											0	106
N36	804												0

Appendix E

The relevance judgements of the Cranfield collection

E.1 The README-file accompanying the relevance judgements for the cranfield collection

Here you will find two files containing relevance judgements. CRAN.REL was taken from Ed Fox's Virginia Disc 1 CD-ROM The other came courtesy of Donna Harman (who is a star).

Ed's file contains query-doc id pairs

Donna's file contains the same query doc-id pairs AND includes degrees of relevance which are discussed below.

For some strange reason the files are identical for all but three lines. I leave it to you to figure out the difference.

I am attaching my copy of the qrels for cranfield 1400 (cranqrel.txt), including the codes for relevancy scale, which were added here. The qrels are in three columns: the first is the query number, the second is the relevant document number, and the third is the relevancy code. The codes are defined by Cleverdon as follows:

"1. References which are a complete answer to the question.

2. References of a high degree of relevance, the lack of which either would have made the research impracticable or would have resulted in a considerable amount of extra work.

3. References which were useful, either as general background to the work or as suggesting methods of tackling certain aspects of the work.

4. References of minimum interest, for example, those that have been included from an historical viewpoint.

5. References of no interest."

Obviously no 5's are included in the qrels.

E.2 Recall bases for different dichotomization levels

The relevance grades discussed in Section E.1 allows for four levels of relevance judgement dichotomization (determining the limit between documents judged relevant and non-relevant to a query), based on four different limits (between level 1 and 2, between level 2 and 3, between level 3 and 4 and between level 4 and 5, respectively). Table E.1 provides data about the recall bases that different choices of dichotomization level limit would provide.

Table E.1: *Recall bases at different dichotomization level limits.*

Dichotomization Limit rel <->nonrel	Largest recall base single query	Smallest recall base single query	Total of recall bases for all queries	Average recall base per query
4 <-> 5	41	2	1837	8,16
3 <-> 4	37	1	1474	6,55
2 <-> 3	23	1	740	3,29
1 <-> 2	7	1	353	1,57

Appendix F

Some results for the manually indexed collection

Table F.1: *Simulation scenario 1 (manually indexed collection), R-precision measure: summary of differences between rank-sums of 12 organizations. The bold-faced entries represent significant differences (Critical value 125.7)*

org	rp1 d	P309	P158	V309	V158	V75	P75	P36	V36	N309	N158	N75	N36
19,65	Mean results	0,181	0,159	0,132	0,142	0,113	0,103	0,071	0,076	0,058	0,055	0,054	0,05
> 125,74	Σ ranks	1780	1688	1615	1588	1552	1448	1366	1360	1304	1293	1288	1271
P309	1780	0	92	***165***	***192***	***228***	***332***	***414***	***420***	***476***	***487***	***492***	***509***
P158	1688		0	72,5	99,5	***136***	***240***	***322***	***328***	***384***	***395***	***400***	***417***
V309	1615			0	27	63	***168***	***249***	***255***	***312***	***323***	***328***	***344***
V158	1588				0	36	***141***	***222***	***228***	***285***	***296***	***301***	***317***
V75	1552					0	104,5	***186***	***192***	***249***	***260***	***265***	***281***
P75	1448						0	81,5	87,5	***144***	***155***	***160***	***177***
P36	1366							0	6	62,5	73,5	78,5	95
V36	1360								0	56,5	67,5	72,5	89
N309	1304									0	11	16	32,5
N158	1293										0	5	21,5
N75	1288											0	16,5
N36	1271												0

Table F.2: *Simulation scenario 1 (manually indexed collection): summary of differences between rank-sums of 12 organizations, SRP measure. The bold-faced entries represent significant differences (Critical value 152.7)*

org	srp1d	P309	P158	V158	V309	V75	P75	V36	P36	N158	N309	N75	N36
30,85	Mean results	0,105	0,087	0,068	0,067	0,046	0,044	0,026	0,025	0,025	0,023	0,025	0,02
> 152,74	Σ ranks	1925	1781	1697	1646	1594	1546	1352	1256	1221	1220	1176	1139
P309	1925	0	144,5	***228***	***279,5***	***331***	***379,5***	***573,5***	***669***	***704,5***	***705***	***749***	***786,5***
P158	1781		0	83,5	135	***186,5***	***235***	***429***	***524,5***	***560***	***560,5***	***604,5***	***642***
V158	1697			0	51,5	103	151,5	***345,5***	***441***	***476,5***	***477***	***521***	***558,5***
V309	1646				0	51,5	100	***294***	***389,5***	***425***	***425,5***	***469,5***	***507***
V75	1594					0	48,5	***242,5***	***338***	***373,5***	***374***	***418***	***455,5***
P75	1546						0	***194***	***289,5***	***325***	***325,5***	***369,5***	***407***
V36	1352							0	95,5	131	131,5	***175,5***	***213***
P36	1256								0	35,5	36	80	117,5
N158	1221									0	0,5	44,5	82
N309	1220										0	44	81,5
N75	1176											0	37,5
N36	1139												0

Table F.3: *Simulation scenario 1 (manually indexed collection): summary of differences between rank-sums of 12 organizations, SRE measure. The bold-faced entries represent significant differences (Critical value 154,8)*

org	se1d	P309	P158	V158	V309	V75	P75	V36	N309	N158	N75	P36	N36
23,3	Mean result	0,18	0,155	0,144	0,14	0,126	0,12	0,091	0,089	0,088	0,083	0,084	0,078
> 154,8 76819 53384 7	Σ ranks	1889	1709	1670	1640	1601	1506	1330	1296	1289	1233	1217	1173
P309	1889	0	***180***	***220***	***249***	***288***	***384***	***559***	***593***	***601***	***657***	***673***	***717***
P158	1709		0	39,5	69	108	***204***	***379***	***413***	***421***	***477***	***493***	***537***
V158	1670			0	29,5	68,5	***164***	***340***	***374***	***381***	***437***	***453***	***497***
V309	1640				0	39	134,5	***310***	***344***	***352***	***408***	***424***	***468***
V75	1601					0	95,5	***271***	***305***	***313***	***369***	***385***	***429***
P75	1506						0	***176***	***210***	***217***	***273***	***289***	***333***
V36	1330							0	34	41,5	97,5	113,5	***158***
N309	1296								0	7,5	63,5	79,5	123,5
N158	1289									0	56	72	116
N75	1233										0	16	60
P36	1217											0	44
N36	1173												0