Eero Sormunen, Kai Halttunen and Heikki Keskustalo

Query Performance Analyser - a tool for bridging information retrieval research and instruction

UNIVERSITY OF TAMPERE
DEPARTMENT OF INFORMATION STUDIES
RESEARCH NOTES RN • 2002 • 1

Query Performance Analyser - a tool for bridging information retrieval research and instruction

Eero Sormunen eero.sormunen@uta.fi
Kai Halttunen kai.halttunen@uta.fi
Heikki Keskustalo heikki.keskustalo@uta.fi
Department of Information Studies, University of Tampere, Finland

Abstract

Information retrieval experiments usually measure the average effectiveness of IR methods developed. The analysis of individual queries is neglected although test results may contain individual test topics where general findings do not hold. The paper argues that, for the real user of an IR system, the study of variation in results is even more important than averages. The Interactive Query Performance Analyser (QPA) for information retrieval systems is a tool for analysing and comparing the performance of individual queries. On top of a standard test collection, it gives an instant visualisation of the performance achieved in a given search topic by any user-generated query. In addition to experimental IR research, QPA can be used in user training to demonstrate the characteristics of and compare differences between IR systems and searching strategies. The experiences in applying the tool both in IR experiments and in IR instruction are reported. The need for bridging research and instruction is underlined.

Contents

1 Introduction

The mainstream of research on information retrieval systems is based on the use of standard test collections. A test collection consists of a database, a large set of search topics and relevance assessments linking a relatively small subset of database documents to each search topic. In this setting, the performance of a search system or a retrieval method is examined by conducting a query (or a series of queries) for each search topic, by measuring effectiveness at standard points of operation (e.g., fixed recall or document cut-off values), and by averaging across all search topics. A typical goal is to show that some IR system/technique A is better than some other IR system/technique B in terms of *average* performance. The large number of search topics helps to verify the findings at an appropriate level of statistical significance.

Hull (1996) demonstrated how an analysis of individual queries may reveal notable differences between retrieval methods, even though the methods perform equally well on the average. Kekäläinen (1999, 119-122) and Kekäläinen and Järvelin (2000) made a similar observation. The improvement of average performance through query expansion was shown to be statistically significant but for some individual search topics the results were contradictory. A radically diverging behaviour of the system in some sporadic cases raises critical questions: What has caused the anomaly? Is there an unknown variable affecting the performance of a query? Is this anomaly an instance of a class of cases where the general findings are not valid?

The analysis of individual queries and documents provides clues for improving the systems, and is an essential part of creating new ideas for developing systems (Tague-Sutcliffe 1992). Studying of average performance, only, is risky since it emphasises hypotheses testing instead of innovative experimentation. The test collections and tools developed for the experimental research in IR encourage to the "production line" approach. Hull (1996) makes the problems of this approach clear and concrete:

> "*Since all the experimental data necessary to produce evaluation results have been compiled in advance, these experiments can be successfully run without the researcher reading a single word of text in either a query or a document! One sometimes wonders whether new retrieval methods have been accepted or rejected based solely on an observed difference in the average value of precision and recall*."

One consequence of neglecting the analysis of individual queries is that real searchers face difficulties in applying the findings of experimental research. In a typical experiment, the test is designed to verify that one specified factor (e.g. a modified weighting formula) improves the average performance of an IR system. The user of an IR system is solving a particular search problem that is exposed to a set of potential performance factors. The user should be able to identify which factors are most likely to affect performance in that particular situation. To serve the user, experimental research should characterise the "conditions" when the factor studied is likely to have (a) the usual effect, (b) no effect, or (c) the contradictory effect. The "conditions" may relate to the characteristics of search topics, queries formulated, weighting and ranking schemes applied, documents stored in the database, etc.

Most phenomena associated with IR systems have a stochastic nature. It is unrealistic to expect that experimental research could ever construct a solid and integrated performance model for IR systems representing the effects and co-effects of divergent factors. Past research on professional searchers revealed that expert searchers have adopted a finite set of heuristics to discover appropriate query formulations in different search situations (see e.g. Harter and Peters 1985, Mark Pejtersen 1989, Fidel 1991). Heuristics are rules of thumb that suggest potential tactics for formulating queries in a specified search situation. Searching

heuristics are a kind of hypotheses, the validity of which is tested in the trial-and-error process of searching. The advantage of heuristics is that the user is able to reduce substantially the problem space at hand, and the steps needed to achieve the goals of a search. In IR instruction, searching heuristics are routinely taught to novice searchers in "hands-on-keyboard" classroom exercises on operational IR systems.

At this point, it is easy to see the similarities between the analysis of individual queries in experimental research, and in "hands-on-keyboard" training designed for novice searchers. The experimenter attempts to discover and test new hypotheses (heuristics) of factors affecting retrieval performance while the novice searcher tries to learn the known heuristics through practical training. Obviously, the link between IR experiments and IR instruction is essential in supplying research results into practice. Further, both groups would benefit if effective tools were available for demonstrating retrieval phenomena. This topic is focused on in the paper.

Direct use of an operational retrieval system, both in IR research and instruction, has a shortcoming that the user is not given feedback about the effectiveness of queries. The user has difficulties in perceiving the actual effectiveness of a query formulated, or the effect of changes between queries. In principle, the searcher may judge each retrieved document and estimate the mutual effectiveness of queries but this process is too awkward and time consuming in the time frame of online training/testing. Effective learning and testing requires that the user be given more instant feedback about the effects of query modifications made.

In this paper, the potential solutions to the above problems are discussed by describing the prototype and applications of the Interactive Query Performance Analyser for Information Retrieval Systems[1] (Query Performance Analyser - QPA for short). QPA is a tool developed for the performance analysis and visualisation of individual queries. In experimental research, QPA helps detailed analysis of test results - the variation of performance in queries formulated and reformulated for a given topic. It has been applied also to optimise queries in retrospective IR experiments. In the classroom, QPA is a tool for showing how alternative query formulation strategies affect the performance achieved. The tool can also be used to build learning environments where students may perform exercises and learn IR techniques at their own pace.

The idea of QPA was introduced in earlier papers each focusing on a different application area. The use of QPA in a cross-lingual IR experiment was discussed in Sormunen et al. (1998), in interactive optimisation of queries in Sormunen (2000a), and in learning environments in Halttunen & Sormunen (2000). The goal of this article is to collect the separate application lines together and integrate the knowledge gained from various experiments. The authors aim to promote interaction between experimental IR research and pragmatically oriented IR instruction.

The article is organised in the following way. In Section 2, we will outline the basic ideas behind the Query Performance Analyser and present a description of the implementation. In Section 3, the uses of QPA in experimental IR research are introduced. Section 4 will present how QPA has been used and can be used in IR instruction. Finally, the challenges of developing the Query Performance Analyser and test collections are discussed.

[1] Earlier called Information Retrieval Game, see Sormunen et al. (1998), and Halttunen and Sormunen (2000).

2 Query Performance Analyser

2.1 Structure and Implementation

The Query Performance Analyser was developed at the Department of Information Studies, University of Tampere, to serve as a tool for rapid query performance analysis, comparison and visualisation (Sormunen et al. 1998). The major components of the tool are: (1) a set of pre-defined search topics for a given textual or image database, (2) relevance judgements explicating which documents match the relevance criteria of a search topic, (3) a module supporting query formulation, (4) a front-end system for query execution in selected IR systems, and (5) a module for measuring and visualising the performance of user generated queries (Figure 1). Databases, search topics and relevance data can be similar to or derived from the standard test collections.

At present, QPA (version 3.5) is interfaced to, and utilises, the following external resources:

(1) Retrieval Software

a) The Boolean retrieval system TRIP

b) The probabilistic retrieval system InQuery

(2) IR Test Databases

a) A Finnish research database (54,000 newspaper articles with 35 test topics and 17,000 relevance judgements).

b) A Finnish instructional database (51,500 newspaper articles with sample search topics and relevance judgements).

c) An English research database (514,000 news documents - subset of TREC with corresponding TREC test topics and relevance judgements).

d) An instructional image database (about 24,000 newspaper photographs containing textual metadata with sample search topics and relevance judgements).

(3) Query Manipulation Resources

a) Morphological programs for normalisation of word forms

b) Bilingual electronic dictionaries for word-by-word translations (with English, Finnish, German, and Swedish as languages)

(4) Applications for calculating recall-precision information for a query result set.

Resources used by the tool, e.g., retrieval software, databases and automatic query translation, stemming, expansion or restructuring methods can be replaced with other corresponding components in order to modify the functionality of the application. The test databases, translation dictionaries, and database index types utilised may be changed.

QPA combines the database and software components of the IR environment listed above and is accessible through any standard WWW browser. The implementation is based on using HTML pages as a web interface for executing CGI programs. The CGI programs are used for running applications in the Unix environment which perform the needed operations, like translating query words into another language by using bilingual electronic dictionaries, or calculating the recall precision information for a given search result. A general flow of operations between the WWW interface (WWW browser side), CGI programs, and Unix programs (WWW server side) is given in Figure 2.

2.2 Configuration of the tool

The tool can be used in two modes: in direct mode and in exercise mode. In direct mode the user of the tool defines his working environment by himself by selecting the target database, the search engine, the search task or topic, and the performance feedback types preferred. The selections are done by simply enforcing the appropriate buttons from a HTML page. In the exercise mode, which is utilised mainly in IR instruction, the working environment is pre-defined by the tutor of the course. The tutee enters directly to the query input page by activating a link on the exercise page. He is then served by the set of operations and feedback features selected by the tutor for that exercise. For instance, performance feedback may be shown, shown delayed or hidden completely.

2.3 Query formulation

Query formulation page contains only the description of the search topic, an empty field for query input, and links to the query language help texts. It is up to the user to formulate the query, and check that it is in the syntax of the query language used (Figure 3, upper text field). If some of the bilingual translation dictionaries have been switched on, query terms are automatically translated into the target language. The translated query produced by the tool can be seen in the lower text field. Different operators may be applied in translated queries depending on the selection made at the configuration stage. For example, in structured English-to-Finnish translation, each English query term is replaced automatically with a set of corresponding Finnish terms connected with a synonym operator of the InQuery retrieval engine.

The user is free to edit the translated query for instance to test the effects of removing ambiguous or incorrect translations. When the user clicks the "Submit query" button, the selected search system is logged in, the query is processed, and the set of retrieved documents is downloaded (access numbers and titles). Automatic translation facilities were designed for studying cross-lingual IR and are typically switched off in other uses. In those uses, the user is shown only the upper query entry field.

2.4 Performance feedback

The front-end functions manage the process of running the queries and downloading documents. QPA downloads automatically up to 400 matching documents in short form (identification numbers and titles). The list of retrieved documents is compared to the list of relevant documents to identify the ranks of relevant documents in the query result. This information is used to mark up relevant documents in the list of document titles displayed to the user, and to create a visualisation of the document list called *relevance bar* (Figure 4). The same data is used to compute precision at standard recall levels and presented in the form of a P/R graph (Figure 5). The performance feedback is presented immediately and automatically for any query entered into the system. The effects of query modifications can be easily perceived.

The relevance bar representation was especially designed for novice users. Colour coding in the bar is used for expressing the relevant documents (green) or the non-relevant documents (grey/white). The user may also click the bar to select the set of ten document titles displayed (ranks 1-10 in the picture, scrollable up to 400 titles). The full texts of a document can be displayed by clicking the title. Precision-recall graph is a traditional way to represent averaged performance data. At present, QPA automatically displays the P/R graph of the most recent query ("…your query") and the best query of all earlier sessions ("The best user query" in Figure 5). The P/R graphs of any queries can be attached to the co-ordinates as references.

The relevance bar and the precision-recall graph emphasise different viewpoints on query effectiveness. The relevance bar illustrates the content of the result set as seen by the user top-down. The user may estimate the browsing effort needed to reach a desired number of relevant documents. On the other hand, the precision-recall graph illustrates better the overall performance of the queries: on the basis of the graph (or a single precision-recall value point) it is possible to say which query performed best at a given recall level (a system-oriented view).

The best queries executed so far by any of the users can be saved onto the "Hall of Fame" list. The "Hall of Fame" contains the best queries, user names and precision/recall levels achieved. The list is accessible for all users in direct mode but can be concealed in the exercise mode.

3 The use of QPA in IR research

The Query Performance Analyser is an appropriate tool for different tasks in experimental IR research. As suggested in the introduction, the averaged results of an experiment can be deepened by analysing individual queries. The aim of comparing individual queries is to understand the performance variation and the reasons of variation from one search topic or from one query to another. Instant performance feedback has also been used to optimise queries in retrospective experiments (see Sormunen 2000a and 2000b).

3.1 The analysis of individual queries prior to and after an experiment

The findings based on averaged performance data can be elaborated by analysing the variation of performance in individual search topics. For instance, in a cross-lingual IR experiment, an exceptionally low precision may be a consequence of various factors: a syntactic error in the query, translation error, or a 'noisy' translation-based search term. Usually the potential explanations for the badly performing queries are easy to identify. The Query Performance Analyser is an efficient tool to conveniently verify or reject hypothesised explanations, and to estimate the performance effect of any query 'corrections'.

QPA has been used to analyse the variation of performance in a query expansion experiment by Kekäläinen (1999), and in a dictionary-based CLIR experiment by Pirkola (1998, discussed in Sormunen et al. 1998). The average effect of full expansion when using structured queries was positive and this result was shown to be statistically significant. However, the effect of expansion was clearly negative or negligible in some search requests.

Typically, expanded queries contained several search keys that were potential reasons for the poor performance. For instance, expansion may bring in common or ambiguous search keys or the relaxation of phrases may induce negative effects. The interactive analysis by the Query Performance Analyser showed that actually only some of the problematic search keys added had a real negative effect. Some of the unjustifiable expansions were an aesthetic problem, only. QPA supports a kind of sensitivity analysis: to identify the productive, non-productive, and counter-productive search keys (or other reasons for good or poor performance). Without knowing the real source of poor performance, it is not possible to collect, classify and analyse error data appropriately.

The use of the Query Performance Analyser introduced above takes place after actual query tests. Queries where the measured performance deviates considerably from the average are analysed in detail. Similar analysis of individual queries can be applied in advance of the experimental stage. At this point, the goal of analysing individual queries is to develop research ideas. QPA supports the construction of experimental IR environments by connecting/disconnecting external resource available for query formulation (e.g. translation

dictionaries), and by selecting appropriate combinations of matching algorithms and databases. Newborn ideas can be pre-tested to see if they are promising enough and worthy of a full-scale evaluation. Similarly, the designs of large experiments can be pre-tested to eliminate most potential faults in the procedure.

Puolamäki (1999) used QPA as a tool to develop research ideas for his Master's thesis on CLIR from English to Finnish. He formulated sample CLIR queries with QPA and studied the problems and their fixes at query level. QPA enabled him to characterise the magnitude of each problem and the potential of the proposed fixes. Structured vs. unstructured queries, and fixing proper name transliteration / inflection problems were studied by the tool. The findings of the study are reported in (Puolamäki, Pirkola and Järvelin 2001).

One may argue that the development of research ideas or the detailed analysis of experimental results could be made by designing small scale batch-mode test or by using interactive IR systems directly. Is the investment on the tools justified if one may complete small tests a bit more handily? A counterargument is that an investment on efficiency leads sometimes to a breakthrough in effectiveness or productivity. If some analysis cannot be performed conveniently, it is not made at all. In much of research, speedy processing of queries is an overriding factor (see Hull 1996). If Hull's criticism is to be met constructively, fast and convenient tools for query level study are a necessity.

3.2 Query optimisation in retrospective experiments

The Query Performance Analyser is an appropriate tool for conducting IR experiments by applying the retrospective evaluation method. In this method, queries are optimised for each setting compared on basis of complete relevance data (Robertson 1996). The idea of optimising queries is to find out the maximum capability of a retrieval system under given conditions. The approach has been shown effective, for instance, in comparing the performance of optimal Boolean queries in databases of different sizes or when different conjunctive operators, e.g. Boolean AND vs. proximity operators are exploited (Sormunen 2000a, 2000b).

Sormunen (2000a) developed a heuristic algorithm for optimising Boolean queries, and used QPA to test how credible the queries automatically generated by the optimisation algorithm were. A group of test searchers were given query plans, the same that the optimisation algorithm had used. These query plans were originally formulated by an independent search analyst. A query plan consists of an *ordered* set of facets (an exclusive aspect/concept of a search topic identified by the search expert, e.g. *[South America] AND [debt] AND [crisis]*). All searchable facets identified were included. For each facet, the search analyst had attempted to find all reasonable search keys. In both optimisation cases, the goal was to formulate queries (to select one or more facets and one or more query terms for each facet selected) maximising precision at a particular recall level. For instance, an optimal query could be: *(Argentina OR Brazil) AND (foreign debt OR emergency loan).*

Strict guidelines were designed for the test searchers to guarantee a consistent treatment of different search topics. The search for the optimal Boolean queries was started by first finding out the optimally performing disjunction of query terms for the first facet. Next, the searcher looked for the optimal query for the conjunction of the first and the second facet and so on. By increasing the number of facets exploited gradually, the searcher could learn the most productive query term combinations for each facet, avoid repeating trials with obviously ineffective or harmful query terms, and reduce time required to find good candidates for optimal queries. (Sormunen 2000a.)

The use of the Query Performance Analyser in optimising queries turned out to be a resource consuming but not an unrealistic method for optimising queries. Replacing the automatic algorithm by the real user operating QPA lead to slightly lower precision levels but more realistic query formulations.

The major advantage of the tool is that it offers a convenient manner to make comparisons between Boolean and best-match systems. Only few experiments of this type have been conducted (see Salton 1972, Turtle 1994, Paris and Tibbo 1998). An obvious reason for the lack of experiments is that designing a fair comparison between different matching models is difficult. Queries having an identical content and structure cannot be used. Tests based on autonomously working searchers lead to query statements that are difficult to compare. Using beforehand designed query plans and interactive optimisation of queries by QPA, these problems can be avoided.

A retrospective experiment applying the Query Performance Analyser was conducted to study the maximum performance achieved in Boolean and best-match queries (Sormunen and Kekäläinen 2001). Because the earlier study (Sormunen 2000a, 2000b) raised a methodological question of overfitting[2], the optimisation process was designed in a new way. Two separate sets of relevant documents were used in the optimisation of queries and in the measurement of their performance (a training set and a test set, respectively). This time three independently working test searchers were used to ensure the representativeness of results.

The main contribution of the last experiment was that the differences in the mechanism of constructing effective queries could be revealed. The results were based on a large set of conceptually justified query statements for each matching model. This is an advantage that is difficult to achieve without a tool like QPA. For details, see Sormunen and Kekäläinen (2001).

4 The use of QPA in IR instruction

Information retrieval instruction covers four main areas focusing on presenting (1) the context of information retrieval as a part of information seeking activities, (2) basic principles of information retrieval systems, (3) general search strategies applicable in all ordinary retrieval settings, and (4) specific search strategies for particular retrieval settings and information sources. The main goal of instruction is to develop learners' practical capability to conduct searches and understand the heuristic nature of IR techniques.

4.1 QPA as a component of a learning environment

At the present stage of development, QPA is a novel prototype tool for constructing IR learning environments. As a component of a learning environment, QPA represents phenomenaria, i.e., an area for presenting, observing and manipulating phenomena of IR (see Perkins 1991). The tool can also be seen as a construction kit for query modelling and analysis. It offers possibilities to simulate different kinds of settings of search topics, databases and retrieval systems.

QPA can be used for different purposes: *For a tutor in a classroom*, QPA is a tool to show the overall effectiveness of any query. It is easy to demonstrate how any reformulation of a query, or any change in the retrieval system reflects on query performance. *For a designer of a learning environment*, QPA is a tool for creating web-based searching exercises on which students may work at their own pace. *For an advanced student*, the analyser is an

[2] For examples of the overfitting problem, see Robertson (1996)

environment for learning by doing, for example the query formulation tactics in Boolean or best-match retrieval systems.

In Boolean queries, visualisation of search result with colour coding is an efficient tool to demonstrate the size and the content of the result set. For instance, in a database of news articles it is easy to illustrate how relevant articles occur in clusters in the chronologically ordered result set, sometimes quite far from the top. The observation that so many relevant documents are not retrieved although a query was carefully designed can be a shocking experience for a user who has worked only with traditional searching environments.

In best-match queries, visualisation of search result is very useful when demonstrating the changes in the relevance ranking of documents from one query to another. It is much more difficult for the user to gain control over the search results in best-match systems than in the (exact-match) Boolean systems. Nor is there an established corpus of expertise in everyday best-match searching comparable to that of Boolean searching. Thus, QPA is also an excellent instrument for teachers (and for researchers as well) to learn, demonstrate, and develop searching strategies for best-match systems.

All teachers of IR at the Department have developed their personal ways to demonstrate IR phenomena by the Query Performance Analyser. Typical examples of phenomena that have been clarified by our teachers are:

1) Basic concepts and search techniques
   a) Showing the differences in performance of individual query terms (query terms are not equally good).
   b) Demonstrating the effect of term truncation on recall (especially important in Finnish).
   c) Emphasising the use of database specific search elements: the advantages of non-topical search keys – (field searching for setting limitations by time, document type, database section, etc.).
   d) Demonstrating the concepts of precision and recall.
2) Boolean searching:
   a) Justifying query expansions: higher recall can be achieved with slight loss of precision.
   b) Warning of unjustified conjunctive structures: loss of recall in case of short documents (e.g. caption texts in image databases).
   c) Promoting the use of proximity operators instead of the Boolean AND operator in full-text databases to improve precision.
3) Best-match searching
   a) Showing how relevance ranking works.
   b) Giving examples of cases when ranking works or does not work effectively.
   c) Demonstrating how important the facet structure is in expanding queries for high recall.
4) Comparison of Boolean and best-match searching
   a) Explicating the basic differences between Boolean and best-match IR systems.
   b) Giving examples of situations where Boolean queries or best-match queries work better than the other.

c) Showing that queries have to be formulated differently for different types of IR systems in order to achieve high performance.

The list of examples emphasises that the Query Performance Analyser is appropriate for demonstrating both general IR phenomena and specific, database or search engine related phenomena.

4.2 Evaluation of the instructional use

Since 1998, the Query Performance Analyser has been used as a routine tool on several IR courses at the department. During the autumn term of the year 1999, the first systematic field evaluation of the tool was conducted. The major focus was on how students experienced the learning situation and the capabilities of the tool. The main results of the evaluation are summarised below, for details see (Halttunen and Sormunen 2000).

The basic function of the system – the query effectiveness feedback – was naturally seen to promote learning significantly. Feedback concerning the performance of one's own query, the chance to freely reformulate the query and to further evaluate the effect of changes on performance, was seen as a highly motivating and illuminating advancement. Furthermore, performance feedback allowed students to pay attention to the analysis and evaluation of query formulation and search keys. This was contrasted with the heavy browsing and evaluation of search results, which is typical when operational databases are used for educational purposes.

On the negative side, feedback mechanism could also fix students attention on precision-recall estimates and some of them tried to improve on their previous results mechanically, without analysis and reflection of their preceding queries and results. In this respect, the feedback mechanism tempts searchers to pay attention to the performance measures achieved, not to the analysis of the search task situation. QPA offers the opportunity to see the best query formulations and the achieved results of other searchers. Presenting other searchers' success creates a subtle competition and desire to improve one's own search results.

The search tasks in the test collections are well specified. This exactness can be seen as an obstacle for learning as the linguistic expressions in the requests may be too predefined and artificial. One solution to this problem could be the integration of search requests within the framework of simulated real-life activities or simulated work-tasks, which are proposed in evaluation of interactive information retrieval systems (Borlund 2000). This kind of approach is similar to anchored instruction, an instructional approach developed by the Cognition and Technology Group at Vanderbilt (1990, 1991, 1992). Anchored instruction is strongly associated with situated learning and constructive learning environments. The major goal of anchored instruction is to overcome the problem of inert knowledge by teaching problem solving skills and independent thinking. Anchored instruction in a learning environment is intended to permit sustained exploration by students and teachers and it enables them to understand the kinds of problems and opportunities that experts encounter and the knowledge that experts use as tools.

Collaborative learning, discussion and articulation of problems and solutions at hand are vital components of a successful learning situation. Although the feedback mechanism of search success is the key function of the tool it can also lead to surface level coping and gaming in order to reach high precision - recall scores. The solution to this problem could be found in the articulation and social negotiation of search statements and the learner's own theories on them. Collaboration challenges the "playing the game" approach where the game-like features attract the learners to a groundless pursuit of high query performance based on performance feedback without articulating the goals and strategies. The rotation of exercises where

performance feedback is given, not given or given delayed to the user could also constrain the adoption of surface level learning behaviour.

Students' use of QPA and attitudes toward it in instruction may be a result of an individual learner's prior experience of IR or, for example, the learner's personal learning style. This is also a matter of the instructional design of exercises, and the learning situation as a whole. QPA leaves many options to the tutor to design a suitable learning environment. The tutor can modify the feedback mechanism, game-like features, help functions, and decisions concerning presentation and articulation of learning outcomes.

5 Discussion and conclusions

The present version 3.5 of the Query Performance Analyser is a prototype. However, the encouraging experiences both in research and in instruction convince us about the feasibility of the QPA concept. In experimental IR research, the interactive tool can be used to augment innovative development of research ideas, to analyse more thoroughly the experimental results, and expand the scope of studies by retrospective evaluations. As an instructional instrument, QPA has been a success in its closest educational community. All teachers involved in IR instruction at the University of Tampere have adopted the tool into routine use.

Active use of the prototype has shown also its limitations and development needs. The expandability and maintainability of the present version is limited. For instance, the idea of plug-in modules - the possibility of installing easily new external resources such as search engines and databases when needed - is not yet fully implemented. Thus, the tool has served well in experimental designs based on the present setting only: the study of mono- and cross-lingual Boolean and probabilistic queries. However, this has not been a heavy restriction since TRIP and INQUERY retrieval systems, and the group of test collection available, provide a versatile environment for designing experiments.

The lessons in instruction, and especially in web-based exercises, have shown that the Query Performance Analyser is an invaluable tool but, nevertheless, making a high-quality learning environment for IR is a complex didactic enterprise. Our experience shows that especially at the introductory level most students need basic lectures, carefully designed exercises and personal guidance. Advanced students have a solid knowledge background and motivation to work with QPA even in direct mode. Advanced students and researchers are, basically, quite similar as learners in this context.

So far, we have mainly worked with standard test collections. The obvious limitations of these collections restrict the capabilities of the Query Performance Analyser. When using typical search topics designed for a standard test collection the hands of an instructional designer are tied. It is not possible to simulate a realistic search situation since relevance data is associated only with one aspect of information, topicality (Borlund 2000). The non-topical user requirements for the documents searched for are as important as topical ones (time, language, geographical, style and level, etc. constraints). These user requirements raise also the need of exploiting more fully the metadata and structures of documents in searching - not only the text words of documents.

Our experiences with QPA seem to challenge also the conventional way of focusing laboratory-based IR research. The sole use of average performance measures hides the variation phenomena important to real users. The question is about the goal of research. Large test sets and averaging are useful in comparing the performance of algorithms to be implemented. However, if the experimenter wants to develop better guidelines (heuristics) for searching by the tools developed, the focus should be more on individual queries - at least on

learning about the characteristics of individual requests and queries, and the effects of their variation.

In the past, research has concentrated on developing automated best-match IR systems. The user has been given a role of an eternal novice. The role includes entering the initial natural language query and giving relevance feedback but not reformulating queries directly. The understatement of user learning has lead to a situation where most best-match systems do not give appropriate tools for formulating queries and taking a more direct control over the IR system. However, recent studies have shown that methods based heavily on user control, e.g. Boolean type structured queries, help to achieve higher performance in best-match IR systems (Kekäläinen 1999, Kekäläinen and Järvelin 2000, Pirkola 1998, Sormunen et al. 2001). These findings underline again the need for opening the research line of individual queries and searching heuristics.

Development of a new version of the tool (QPA version 5.0) is currently in progress. The new version is intended to offer more versatile and convenient tools for researchers recording, managing, and analysing data from different users in interactive IR experiments. The new version will also support the use of richer relevance data. The first step is to visualise multilevel relevance values in the relevance bar giving the user an impression about the variation of relevance between documents.

Acknowledgements:

InQuery (TM) SOFTWARE Modifications Copyright (c) 1998-2000 by the Center for Intelligent Information Retrieval (CIIR) at the University of Massachusetts at Amherst. All rights reserved. InQuery (TM) Copyright (c) 1996-2000 by Dataware Technologies, Inc., Hadley, Massachusetts, U.S.A. (413-587-2222; http://www.dataware.com). All rights reserved. The InQuery (TM) software was developed in part at the Center for Intelligent Information Retrieval (CIIR) at the University of Massachusetts at Amherst (For more information, contact 413-545-0463 or http://ciir.cs.umass.edu). InQuery (TM) is registered trademark of Dataware Technologies, Inc.

The development of the Query Performance Analyser was mainly funded by the University of Tampere and the Academy of Finland (Research Project 37078).

References

Borlund P (2000) Experimental components for the evaluation of interactive information retrieval systems. Journal of Documentation 56: 71-90.

Cognition and Technology Group at Vanderbilt (1990) Anchored instruction and its relationship to situated cognition. Educational Researcher 19: 2-10.

Cognition and Technology Group at Vanderbilt (1991) Technology and the design of generative learning environments. Educational Technology 31: 34-40.

Cognition and Technology Group at Vanderbilt (1992) The Jasper experiment : an exploration of issues in learning and instructional design. Educational Technology Research and Development 40: 65-80.

Fidel R (1991) Searcher's Selection of Search Keys: I. The Selection Routine. II. Controlled Vocabulary of Free-Text Searching. III. Searching Styles. Journal of the American Society of Information Science 42: 490-500, 501-514, 515-527.

Halttunen K and Sormunen E (2000) Learning Information Retrieval through an Educational Game. Is Gaming Sufficient for Learning? Education for information 18: 289-311.

Harter SP and Peters AR (1985) Heuristics for online information retrieval: a typology and preliminary listing. Online Review 9: 407-424.

Hull D (1996) Stemming Algorithms: a case study for detailed evaluation. Journal of the American Society for Information Science 47: 70-84.

Kekäläinen J (1999) The Effects of Query Complexity, Expansion and Structure on Retrieval Performance in Probabilistic Text Retrieval. Doctoral Thesis. University of Tampere, Tampere. (Acta Universitatis Tamperensis 678).

Kekäläinen J and Järvelin K (2000) The Co-Effects of Query Structure and Expansion on Retrieval Performance in Probabilistic Text Retrieval. Information Retrieval 1: 329-344.

Mark Pejtersen A (1989) The Book House. Modelling User's Needs and Search Strategies as a Basis for System Design. Riso National Laboratory, Roskilde (Riso-M-2794).

Paris LAH and Tibbo HR (1998) Freestyle vs. Boolean: A comparison of partial and exact match retrieval systems. Information Processing and Management 34: 175-190.

Perkins D N (1991), Technology meets constructivism: do they make a marriage? Educational Technology 31: 18-23.

Pirkola A (1998) The Effects of Query Structure and Dictionary Setups in Dictionary-Based Cross-Language Information Retrieval. In: Croft WB, Moffat A, et al. Eds., Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, New York, 1998. pp. 55-63.

Puolamäki D (1999) Kielten välinen tiedonhaku: Käännöskyselyjen evaluointi englanti-suomi. [Cross-lingual information retrieval: An evaluation of queries translated from English to Finnish.] Master's Thesis. University of Tampere, Department of Information Studies, Tampere.

Puolamäki D, Pirkola A and Järvelin K (2001) Applying query structuring in cross-language retrieval. Submitted to Information Processing and Management.

Robertson SE (1996) Letter to the Editor. Information Processing and Management 32: 635-636.

Salton G (1972) A new comparison between conventional indexing (MEDLARS) and automatic text processing (SMART). Journal of the American Society for Information Science 23: 75-84.

Sormunen E (2000a) A Method for measuring Wide Range Performance of Boolean Queries in Full-Text Databases. Doctoral Thesis. University of Tampere, Tampere. Acta Electronica Universitatis Tamperensis, URL: http://acta.uta.fi/pdf/951-44-4732-8.pdf.

Sormunen E (2000b) A Novel Method for the Evaluation of Boolean Query Effectiveness across a Wide Operational Range. In: Belkin NJ, Ingwersen P and Leong M-K SIGIR 2000, Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Athens July 24-28, 2000. Special Issue of the SIGIR Forum 34:25-32.

Sormunen E and Kekäläinen J (2001) Retrospective evaluation of Boolean and probabilistic queries applying an interactive query performance analyser. A manuscript submitted for publication. (URL: http://www.info.uta.fi/tutkimus/fire/archive/BoolProb.pdf).

Sormunen E, Kekäläinen J, Koivisto J and Järvelin K (2001) Document text characteristics affect the ranking of most relevant documents by expanded structured queries. Journal of Documentation 57: 358-374.

Sormunen E, Laaksonen J, Keskustalo H, Kekäläinen J, Kemppainen H, Laitinen H, Pirkola, A and Järvelin K (1998) The IR Game - A Tool for Rapid Query Analysis in Cross-Language IR Experiments. PRICAI '98 Workshop on Cross Language Issues in Artificial Intelligence. Singapore, Nov 22-24, 1998, p. 22-32.

Tague-Sutcliffe J (1992) The pragmatics of information retrieval experimentation, revisited. Information Processing and Management 28: 467-490.

Turtle H (1994) Natural Language vs. Boolean Query Evaluation: A Comparison of Retrieval Performance. In: Croft WB and van Rijsbergen CJ (Eds) Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. London: Springer-Verlag. pp. 212-220.

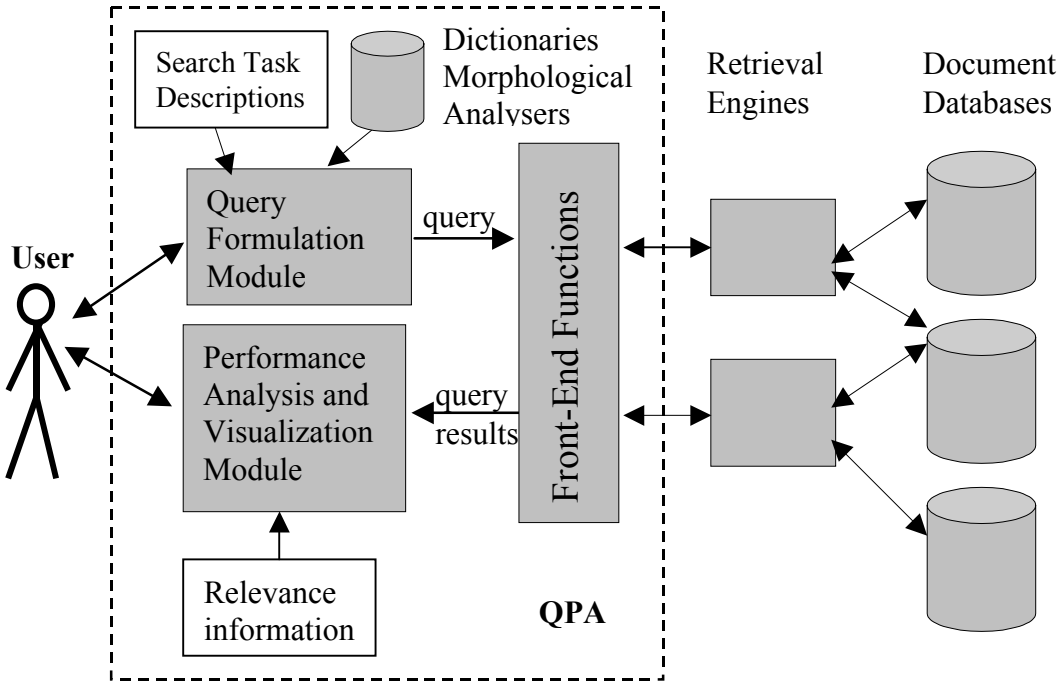

Figure 1. The basic functional components of the Query Performance Analyser.

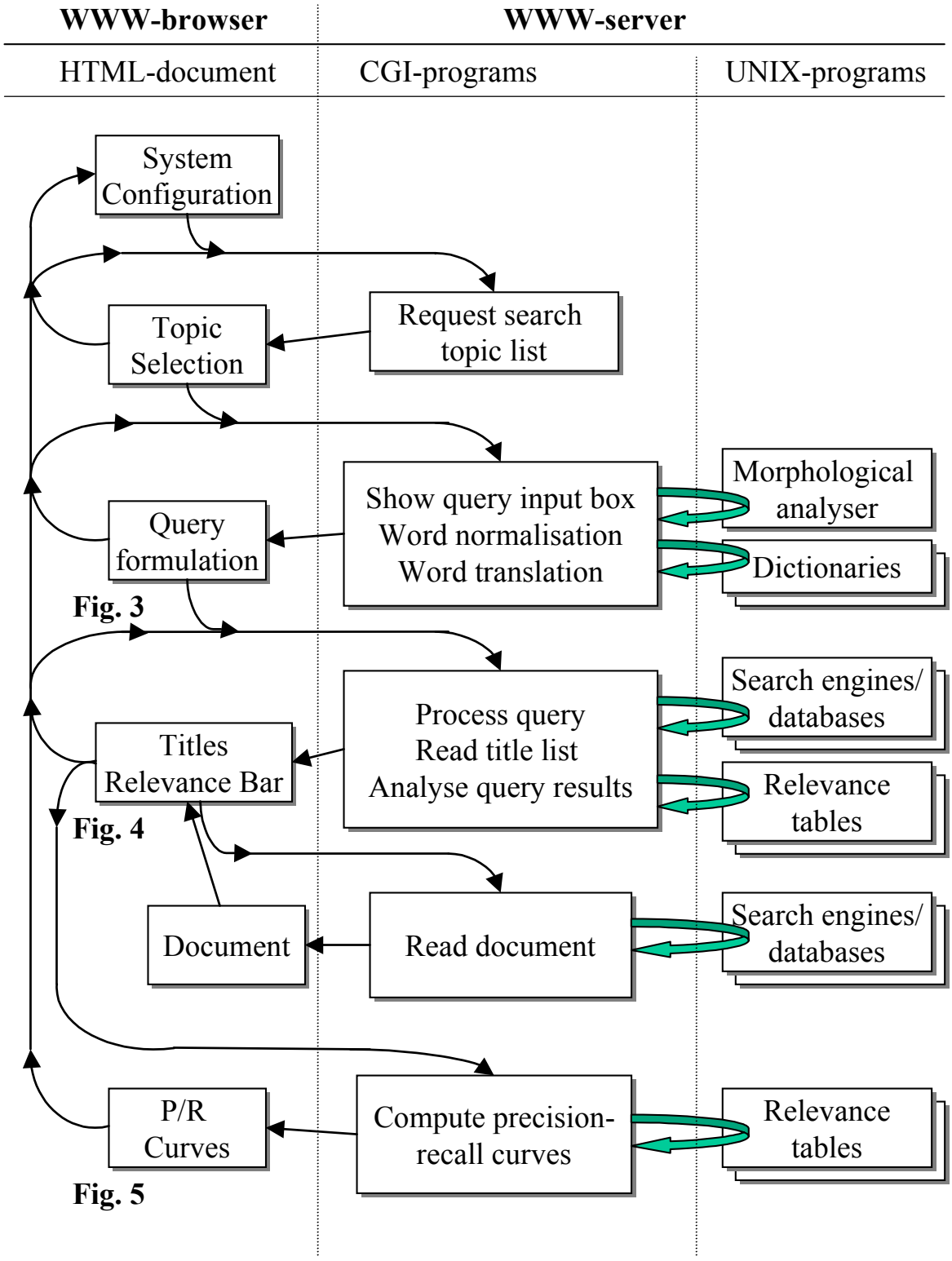

Figure 2. General view of the flow of operation in the Query Performance Analyser, version 3.5.

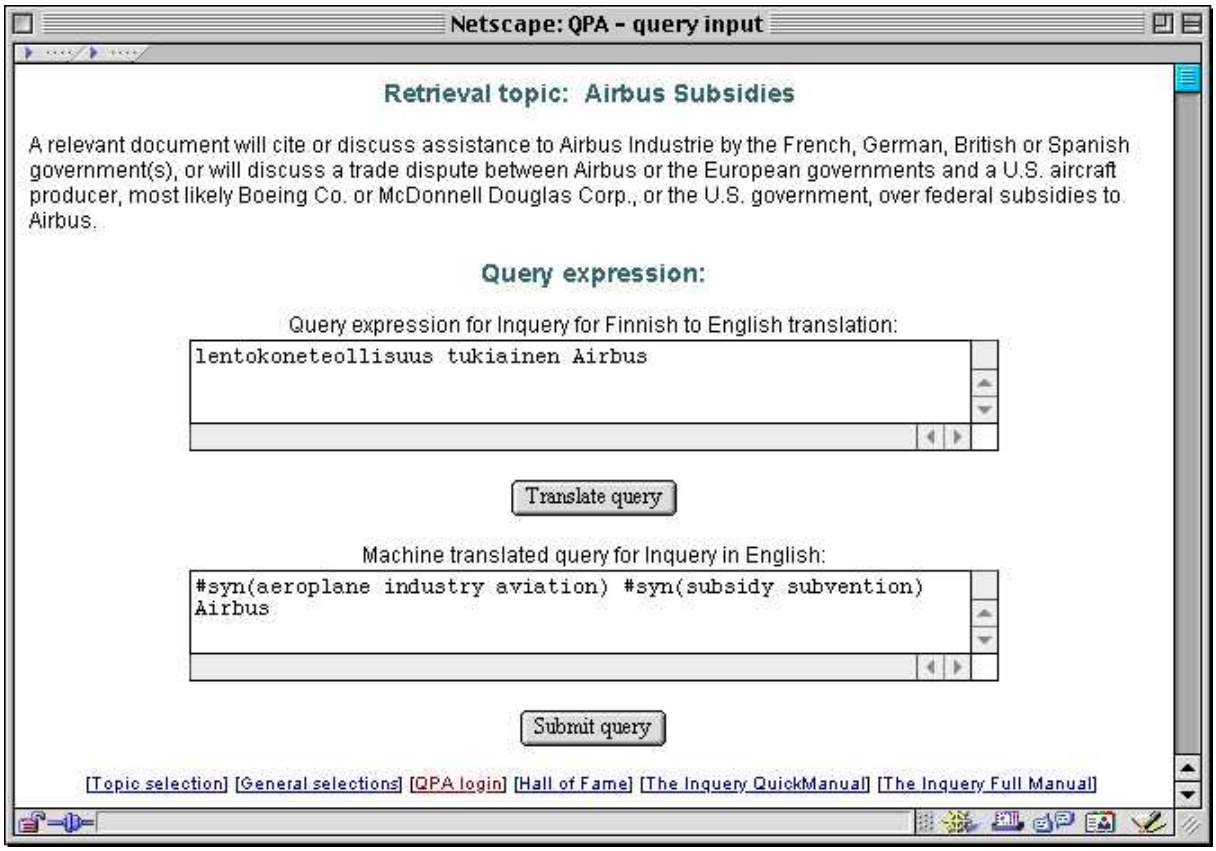

Figure 3. Query input.

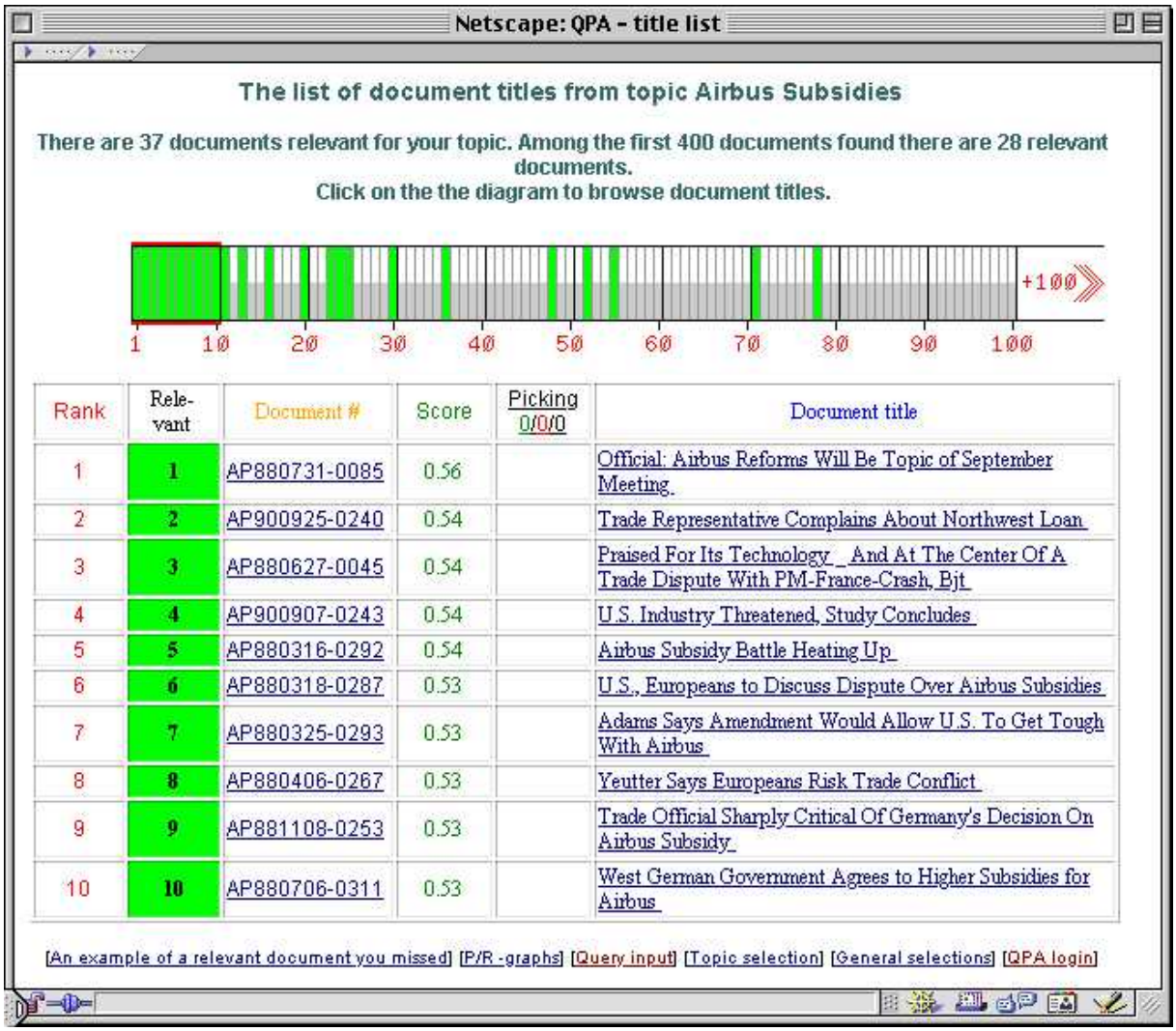

Figure 4. Display of retrieved document titles with the relevance bar.

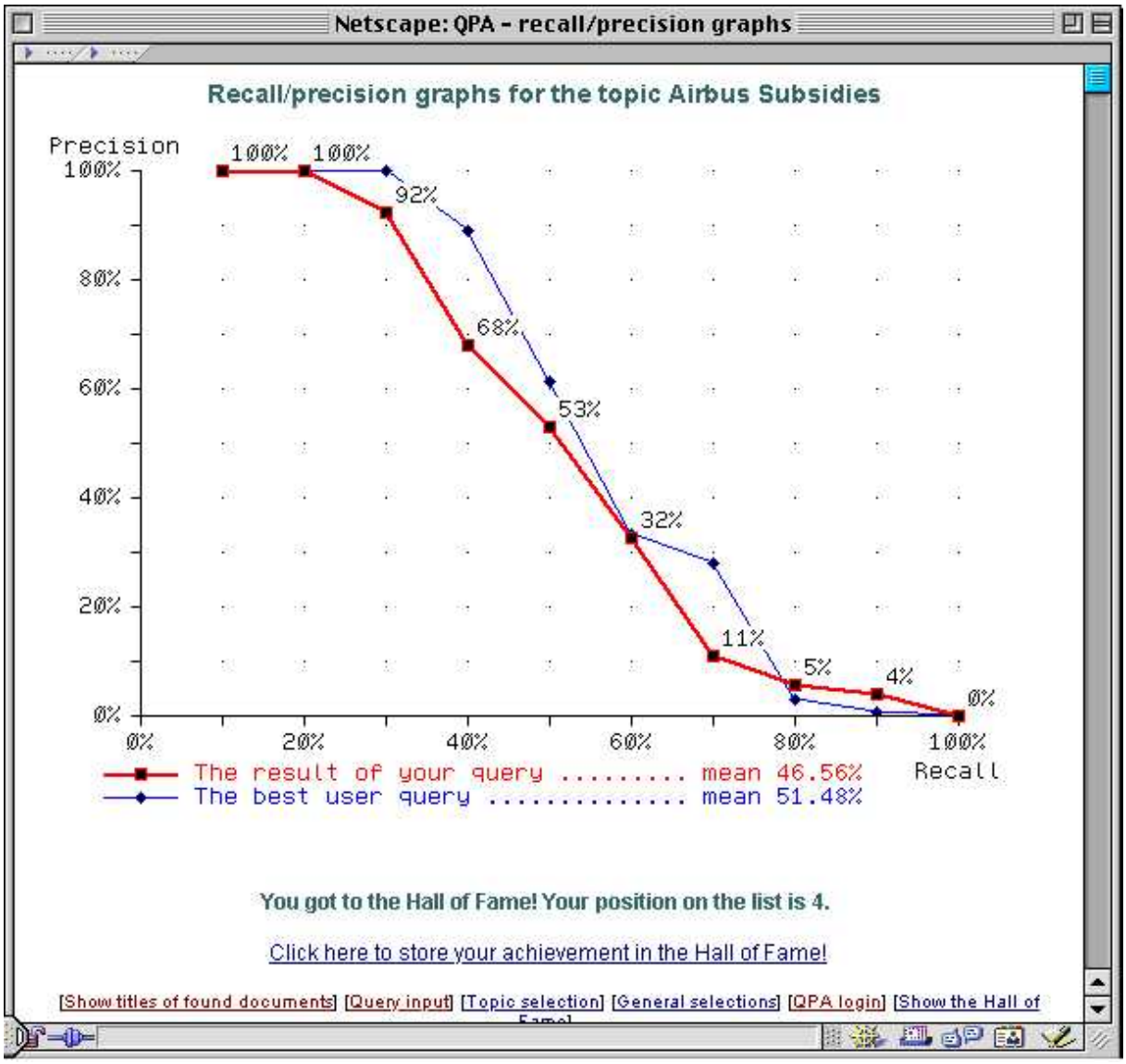

Figure 5. Performance visualisation as a P/R graph.