Topic-Relevance Map: Visualization for Improving Search Result Comprehension

Jaakko Peltonen[1,2] **Kseniia Belorustceva**[2] **Tuukka Ruotsalo**[2,3]
[1]Faculty of Natural Sciences, University of Tampere, Tampere, Finland
[2]Department of Computer Science, Aalto University, Espoo, Finland
[3]Department of Computer Science, University of Helsinki, Finland
{jaakko.peltonen, tuukka.ruotsalo}@aalto.fi

ABSTRACT
We introduce topic-relevance map, an interactive search result visualization that assists rapid information comprehension across a large ranked set of results. The topic-relevance map visualizes a topical overview of the search result space as keywords with respect to two essential information retrieval measures: relevance and topical similarity. Non-linear dimensionality reduction is used to embed high-dimensional keyword representations of search result data into angles on a radial layout. Relevance of keywords is estimated by a ranking method and visualized as radiuses on the radial layout. As a result, similar keywords are modeled by nearby points, dissimilar keywords are modeled by distant points, more relevant keywords are closer to the center of the radial display, and less relevant keywords are distant from the center of the radial display. We evaluated the effect of the topic-relevance map in a search result comprehension task where 24 participants were summarizing search results and produced a conceptualization of the result space. The results show that topic-relevance map significantly improves participants' comprehension capability compared to a conventional ranked list presentation.

ACM Classification Keywords
H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous; H.3.3. Information Storage and Retrieval: Information Search and Retrieval

Author Keywords
Sense-making; Dimensionality Reduction; Visualization; Exploratory Search

INTRODUCTION
Search systems often present the retrieved information for the users as a ranked list of documents in a descended relevance order [34]. Presenting results as a ranked list forces the user to scan through the ranked list and pick sufficiently relevant documents from the top of the ranking. Often, however, users are not looking for a single highest-ranked document, but the initial query is issued to broadly comprehend the information space [39]. This requires proficiency in information literacy based on the result list and confronts users with problems detecting topical subspaces within the initial ranked list [7]. A promising solution to assist the user is to visualize the search result space for the user [9]. In contrast to a ranked list, which uses one dimension (relevance ranking), the advantage of various visualization techniques is that they allow additional dimensions to represent the search results. The methods proposed for search result visualization can be divided into two main types: *visualizing topics or terms within retrieval results* (i.e. indicating the content of a result document with the terms attached to that document) and *visualizing an overview of results* (i.e. using a separate visualization component that provides an overview of the whole result space) [20].

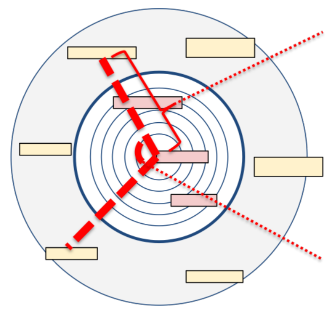

Figure 1. An illustration of the Topic-Relevance Map. The radius, i.e. the distance from the center of the circle represents relevance estimated for a keyword. The angle between keywords represents their projected topical similarity.

We present *topic-relevance map*, a novel overview visualization approach considering two dimensions essential for search results: topical similarity and relevance. Topic-relevance map organizes keywords representing the search results onto a radial layout. Non-linear dimensionality reduction is used to embed high-dimensional keyword representations into angles on the radial layout. Relevance of keywords is estimated by a ranking method and visualized as radiuses on the radial layout. As a result, similar keywords are modeled by nearby points, dissimilar keywords are modeled by distant points, more relevant keywords are closer to the center of the radial display, and less relevant keywords are distant from the center of the

IUI 2017 March 13-16, 2017, Limassol, Cyprus

ACM ISBN 978-1-4503-4348-0/17/03.
DOI: http://dx.doi.org/10.1145/3025171.3025223

radial display. Figure 1 shows an illustration of the underlying principle of the topic-relevance map.

We study the effectiveness and user behavior of topic-relevance map in a search result comprehension task; skimming and understanding the structure and conceptualization of search results by creating a conceptual structure organized as main topics and their subtopics. We perform user experiments in which 24 users comprehend search results in eight tasks. The results, evaluated by several measures, suggest that the topic-relevance map yields a more comprehensive understanding of search results than a comparison system that visualized topics within the search result list. Our findings can help to design more effective sense-making tools that assist users in rapid comprehension and understanding of large search result spaces.

BACKGROUND

Our research is related directly to search result visualization and topic visualization that we review below. We also give a short introduction to result comprehension and sensemaking and position our methodological and empirical contributions.

Search Result Comprehension and Sensemaking

Comprehending search results is a task where users skim and try to understand the structure and conceptualization of the search results. Comprehension might mean the same as the venerable notion of sensemaking; the human ability to understand individual stimuli, especially words, sentences, or chunks [24]. Comprehension is important in broad information search scenarios where a user tries to gain understanding of a topic in order to retrieve more specific or related information in subsequent search iterations [39, 29]. A simple ranked list can support simple look-up search scenarios where users focus on finding one or a few highly relevant documents. However, linear scanning of search result list can be insufficient for broader information needs. Comprehension of search results is needed especially when the aim of the information seeking is not to look up an individual relevant document but to gain an overall understanding of varied information across multiple relevant documents. Only part of the information in each document may be relevant, and information content over multiple documents may be interrelated; the user must then comprehend not only individual documents but a wider body of knowledge spread across documents. Successfully comprehending information content across search results would let the user to relate the result documents to each other, the query, and the underlying information need, and to exploit the information content appropriately in further processing of the found information [24].

Search Result Visualization

Instead of resorting only to traditional ranked document lists, researchers have proposed a variety of techniques to present and visualize search results and allowing more efficient comprehension than exhaustively going over each result document. The main presentation approaches are visualizing topics or terms within retrieval results or visualizing an overview of results either in text or on a separate visualization element [20], and interactive support to direct search and information exploration [35, 6].

Visualization of topics or terms within retrieval results was first proposed in the TileBars system [18]. TileBars visualizes explicit term distribution information in a full text information access system. The representation indicates relative document length, query term frequency, and query term distribution and can be used to order the results.

TextTiling is a technique for visualizing topical structure of a document within the text. It subdivides texts into multi-paragraph units that represent passages, or subtopics [19]. SenseMaker is another good example of early work on supporting information exploration process by visualizing the information space. The interface of SenseMaker approximates the current information context and provides a set of user-centered actions for examining the current context [6]. Another approach has been search result clustering in which the initial result set is clustered and an overview of each cluster is presented to the user [21]. Empirical results on using search result clustering show that relevant documents tend to be more similar to each other than to non-relevant documents and that users can utilize the grouping of information by the topical similarity in selecting relevant results. A similar approach for two-dimensional representation has been proposed for image search allowing browsing on a 2D canvas [23].

A line of recent work has focused on interactive support for making sense of search results and several interactive interfaces have been recently proposed. ExplorationWall [25] is an interface that allows incremental exploration and sensemaking of large information spaces by visualizing documents and related entities as search streams. PivotPaths [13] is another recently proposed interface for exploring faceted information resources by visualizing facets as paths and supporting pivot operations as lightweight interaction techniques that trigger gradual transitions between the facets.

Interactive interfaces that enable transparent control on user models have recently become popular [35, 36, 2, 4, 42, 32]. The idea behind these approaches is that, as opposite to visualizing results, the user model is visualized and the user can interactively provide feedback on the search intentions using the visualization. Similar visual controls for user modeling have also been proposed for recommender systems [5, 11, 41].

Topic Visualization

Topic model visualization has recently been extensively studied. The key idea is close to our approach in that a lower-dimensional representation of the original document space is first computed and then it is visualized in order to assist making sense of the complete data collection [17].

Topic visualizations leverage word lists or word clouds to visualize topic models. For example, Chaney and Blei [8] employed word lists to illustrate the hidden structure discovered by a topic model. The HierarchicalTopics system [15] hierarchically organizes the learned topics and thus can represent a large number of topics without being cluttered. TopicPanorama [28] is a topic visualization that visualizes a picture of relevant topics discussed in multiple sources. It combines a radially stacked tree visualization with a density-based graph visualization to facilitate the examination of the matched topic graph from multiple perspectives. Another topic visualization system is Serendip [3] which focuses on supporting multi-level discovery in text corpora, including the corpus, passage, and word levels. Topic visualization has also been studied in cases of evolving text corpora [10] and real-time adaptive context [12].

Despite extensive research in visualization and interactive support for search result ordering, *there is no conclusive evidence that these visualization approaches would lead to improved search result comprehension in the hands of users* [37, 20]. Previous studies mainly focus on reporting either behavioral findings on the exploration behavior [14, 26], or finding more diverse information or increased coverage of variables in the exploration process [43]. Moreover, existing approaches focus on visualizing different data dimensions as separate widgets [14, 1, 43], but not topical similarity and relevance simultaneously.

While the previous work highlights the utility and applicability of visualization techniques for interactively browsing information collections in general, we lack understanding of the benefits of the visualizations; do the visualizations improve users effectiveness or efficiency in understanding or comprehending the information collections, or are they rather complex proxies for simpler interactions that the users perform as a part of their information exploration processes?

Contributions

The contributions of this article are both methodological and empirical. First, we present topic-relevance map, a novel visualization intended for comprehension of search results by simultaneously visualizing relevance and topical similarity. The visualization is optimized by data-driven machine learning and presents the results via keywords. Second, we show empirical evidence on improved search result comprehension in controlled user experiments comparing our approach to a baseline which visualizes the same topical information within search results. In the experiments, the users were given a short time to analyze a set of search results and their task was to produce a conceptualization that represents the search result space. In this task users who used the topic-relevance map had improved comprehension output and were able to produce better and more broader conceptualization of the search result space. The differences can be attributed to the topic-relevance map visualization.

TOPIC-RELEVANCE MAP

The design principles behind the topic-relevance map are:

1. The topic-relevance map should illustrate, in a compact manner, the topical information across all search results.

2. The map should efficiently show the estimated relevance of each topical content.

3. The map should compactly show relationships among the topical content.

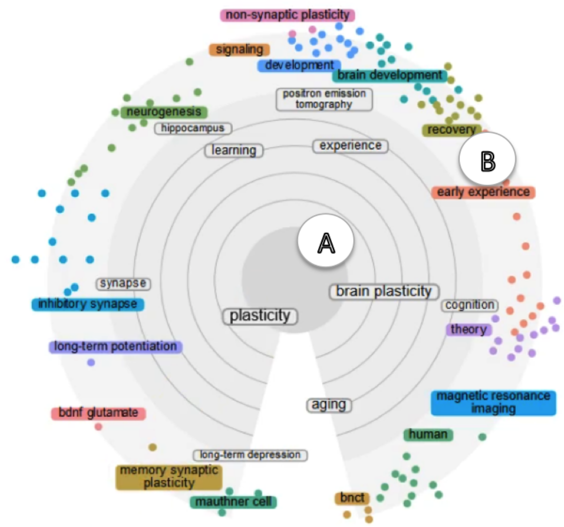

Figure 2. An example of the topic-relevance map visualization in response to a search query "Brain Plasticity". The *inner area* (marked with A) show the top ten keywords related to the search query. The *outer ring* (marked with B) shows a set of other keywords clustered and the most important in each cluster highlighted.

An illustration of the topic-relevance map is shown in Figure 2. The map represents topical content of search results based on keywords contained in the result documents, and organizes the keywords as directions. The map uses a polar coordinate system and radial layout, thus focusing more on relations between keywords than their exact relevance weights, yielding a good tradeoff between the amount of shown information and comprehensibility: a simple list of keywords only uses one degree of freedom and does not show keyword relationships, whereas higher than two-dimensional visualizations could make interaction with the visualization more difficult [16].

The map consists of two main areas: an *inner area* showing the top ten keywords representing the most relevant topics, and an *outer ring* showing the complete set of other keywords representing other topics occurring in the search results and clustered as directions on the radial layout.

In more detail, the *inner area* of the visualization represents the search: the closer a keyword is to the center the more relevant it is. The *outer ring* shows other keywords arranged so that topically keywords are shown with similar angles, thus keywords form angular clusters that represent directions in the information space. The interface colors the keywords according to such clustering, and label each cluster by highlighting the most relevant keyword. The rest of the keywords are shown as dots to preserve clarity. The *inner area* is also arranged along angular directions, and shown with radial positions representing their relevances; the top keywords act as rough signposts along their angular directions, directing the user towards the larger clusters of keywords found along those directions. A user can inspect the clusters with a *fish-eye lens*: the lens zooms into keywords in a small circular area and shows their full text.

Our interface directly supports several key tasks in Shneiderman's [38] taxonomy of tasks for information visualizations, including the visual information seeking mantra *overview first, zoom and filter, details on demand, and relate*:
Overview: the map shows an at-a-glance overview of keyword content in the search results: top keywords, keyword relevances, groups of keywords as directions, and an example keyword from each direction.
Zoom and filter: the fisheye lens functionality in the interface allows the user to zoom into a small part of the visualization.
Details on demand: hovering over a top of a dot representing a keyword highlights the keyword both on the radar and in all documents where the keyword appears.
Relate: the map shows data-driven relationships of keywords by their co-location along similar directions in the visualization, as well as their direct co-occurrences in specific documents.

Preliminaries

Search results contain a set of documents, which in turn contain a set S_{all} of keywords. In our experiments, author-provided keywords are available; we use author keywords and augment keywords if they are missing from the keyword set, but occur in the article abstract. Alternatively, keywords can be extracted by a variety of automated techniques from the text of the documents. Suppose for each keyword k_i we have available an estimated *relevance value* $r_i \in [0,1]$ to the user's information need. From the relevances we identify subset S_{top} of most relevant keywords $k_i \in S_{top}$ by highest relevance; in experiments we took the 10 top keywords. For the rest of the keywords, called the non-top keywords $k_i \in S_{all} \setminus S_{top}$, suppose we have available *topical features* $\mathbf{x}_i$ describing the relationship of this keyword to the potential information needs covered by the current search results.

The keyword relevances and the topical features are all easily estimated, as shown in the "Estimation of relevance and associations" subsection later in this document. In particular, we will use *association strengths* of keyword k_i to each of the top keywords to represent topical content of k_i as described later. The most powerful solution is to directly infer these associations by interaction (lookups) with a search system; however, if such interaction is unavailable we support simpler solutions described in the "Estimation of relevance and associations" subsection.

Layout Computation

We optimize a data-driven layout for keywords by probabilistic modeling-based nonlinear dimensionality reduction, in two stages: first, the outer ring which contains the majority of keywords is laid out, and then the keywords of the inner area are placed following the layout of the outer ring. Our low-dimensional layout summarizes the essential parts of topical variation in the keyword set; it can be seen as a nonlinear and nonparametric reduced-dimensional model of topical content in the results.

Neighborhood preservation: An intuitive goal for dimensionality reduction of keywords. Neighborhood preservation is a powerful concept recently adopted in several dimensionality reduction methods. The idea is that a dimensionality reduction method must place data items from a high-dimensional space into a low-dimensional scatter plot, and preserve neighborhood relationships of the data items. This is a useful goal because 1) it is more flexible than trying to preserve the detailed distances among items, and 2) analyzing neighborhood relationships (finding similar data items) is a natural way to *relate items*, one of the main tasks for visualization in Shneiderman's [38] taxonomy; since users need to perform this task to explore and comprehend search results, optimizing dimensionality reduction for this subtask makes the visualization directly useful for a task of the user.

We quantify the goodness of a visualization of keywords by analyzing what happens to neighborhood relationships between the high-dimensional topical-features space and the low-dimensional angles (directions) on the topic-relevance map. Perfect low-dimensional representation of all high-dimensional relationships is usually not possible; it is then crucial to quantify the errors and optimize the display to minimize them. We first define an input neighborhood and an output neighborhood, and then quantify their differences.

Defining neighborhoods of keywords. Let $\{\mathbf{x}_i\}_{i=1}^N$ be a set of keywords i having a vector of high-dimensional real-valued topical features $\mathbf{x}_i$. We define the features later in this paper. Two keywords can be called *neighbors* if their topical features are similar. Let each keyword i have a neighborhood p_i, which represents which other keywords would be considered similar by a user who inspected their high-dimensional features. We represent p_i as a probability distribution $p_i = \{p(j|i)\}$, telling for each other keyword j the probability $p(j|i)$ that if the user had to pick an example neighbor she would pick j. The probabilities should be high for neighbors with features similar to i and small for neigbors with dissimilar features. We take a simple definition,

$$p(j|i) = \frac{\exp(-||\mathbf{x}_i - \mathbf{x}_j||^2/\sigma_i^2)}{\sum_{j'} \exp(-||\mathbf{x}_i - \mathbf{x}_{j'}||^2/\sigma_i^2)} \tag{1}$$

which is a Gaussian falloff with respect to distance between features of keywords i and j, normalized so that the probabilities sum to one. The constant σ_i lets us set the falloff rate; we set it as in [40]. This is a simple definition of a probabilistic neighborhood, which already yields good results.

On the outer ring of the topic-relevance map, keywords have angles $\{a_i\}_{i=1}^N$, and two keywords appear similar if they have close-by angles. We define for each keyword i an *output neighborhood* q_i telling how similar other keywords appear to be based on their angles. The q_i will again be a probability distribution $q_i = \{q(j|i)\}$ where we define the probabilities analogously to the input feature space, as

$$q(j|i) = \frac{\exp(-|a_i - a_j|^2/\sigma_i^2)}{\sum_{j'} \exp(-|a_i - a_{j'}|^2/\sigma_i^2)} . \tag{2}$$

Errors in neighborhoods. When high-dimensional keywords are placed onto a low-dimensional display two kinds of errors will happen (Figure 3): *misses* are keywords j that were neighbors of i in the input space but not on the display, having

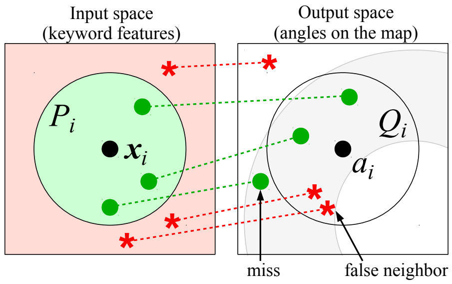

Figure 3. Top: errors in visualization of neighborhood relationships of a keyword i. The input space denotes high-dimensional descriptions $\mathbf{x}_i$ of keywords. The set P_i denotes other keywords whose neighborhood probability with keyword i is high. The output space denotes low-dimensional visual locations of keywords, here angles a_i on a radial display. The set Q_i denotes other keywords whose neighborhood probability with keyword i is high on the display. Two kinds of errors are illustrated: *misses* are keywords that are neighbors in the input space but not on the display, and *false neighbors* are keywords that are neighbors on the display but are not neighbors in the input space.

high $p(j|i)$ but low $q(j|i)$, whereas *false neighbors* are keywords that look like neighbors of i on the display but are not neighbors in the input space, having high $q(j|i)$ but low $p(j|i)$. The costs of misses and false neighbors can be quantified by generalizations of the information retrieval measures *precision* (penalizes false neighbors) and *recall* (penalizes misses). It can be shown that the measure

$$D_{KL}(p_i, q_i) = \sum_{j \neq i} p(j|i) \log \frac{p(j|i)}{q(j|i)} \tag{3}$$

which is a Kullback-Leibler divergence between the distributions p_i and q_i generalizes recall (the divergence is high when there are many misses; see [40] for a proof of the connection) and the divergence $D_{KL}(q_i, p_i)$ similarly generalizes precision (it is high when there are many false neighbors). The sum of these two divergences over all neighborhoods p_i and q_i is then a suitable measure for quality of a visualization.

We next describe the construction of the topic-relevance map using the above dimensionality reduction definitions. Our principle is to place each keyword k_i on a radius proportional to its estimated relevance r_i (the larger, the closer to the center). We first estimate the topical features (Stage 1), then we create the outer ring (stage 2) which has less radial space but more angular space to show directions, and then (stage 3) we create the inner area.

Stage 1: Estimation of relevance and associations

The topic-relevance map is applicable to several search systems. We characterize keywords k_i by relevances and topical features. If a system can directly provide relevance scores r_i for keywords, we use those, and in experiments we did that. For topical features, we do not resort to some generic feature set, as generic features could turn out irrelevant or distracting for information needs of a particular search. Instead, for each keyword k_i we create topical features by *associations of keywords to top keywords*: we use features $\mathbf{x}_i = \{x_{il}\}$ where x_{il} is the association strength of keyword k_i to an information need characterized by top keyword k_l.

Topical features as associations from lookup searches. For each top keyword k_l we estimate associations x_{il} of all keywords by a lookup. In principle, we could simply append k_l to the current query as a query suggestion, and re-estimate keyword relevances using the search system; for each keyword k_i the re-estimated keyword relevance would be used as the x_{il}. However, this would require carrying out new searches, and we instead perform a lookup within the current result set.

In detail, we use Bayesian linear regression to estimate relevances for each keyword in a lookup, based on their document occurrences. Using the term frequency-inverse document frequency (TF-IDF) matrix of the current result set, for each keyword we use its TF-IDF values across the documents as an input feature vector for the regression. We set the lth top keyword k_l as a known target, assigning it the highest relevance $r_l = 1$; this represents a lookup where the user explores intent related to k_l. Bayesian linear regression trained with this target yields estimated relevances $\hat{r}_i$ for each keyword; we set $x_{il} = \hat{r}_i$ as the association strength, and repeat for lookups with other top keywords k_l.

The above Bayesian linear regression has two advantages: 1) if the search results arise as as part of an ongoing search session, any available previous keyword relevance feedback from users can be used as additional targets in addition to each lookup target; 2) the Bayesian relevance estimate can be made to balance exploitation (of relevances estimated from sparse data) and exploration (of uncertain keywords that have potential to be relevant) by taking upper confidence bounds as relevance estimates instead of using mean estimates.

Normalization. After the lookups, we use the norm of the association vector $\mathbf{x}_i$ as the radius r_i for non-top keywords on the topic-relevance map; it represents relevance across all the lookups instead of just the current relevance. We then normalize the $\mathbf{x}_i$ to have norm one; the normalized association tells which of the top keywords are most associated with each non-top keyword k_i. The $\mathbf{x}_i$ are a high-dimensional topical description of the keyword; we ue dimensionality reduction to reduce them to angles on the topic-relevance map.

Simplified relevances. If relevance scores for keywords are not available from the system, a simple alternative is to estimate r_i, e.g., proportional to average TF-IDF value across result documents, or average weighted by document ranks. Thus the topic-relevance map can be constructed whenever search results and their document-keyword TF-IDF matrix are available. Our system in experiments directly provided keyword relevances and we used those.

Stage 2: Layout of Outer Keywords as Directions in the Information Space

To create the outer ring, we take the non-top keywords k_i and use their relevances as radiuses, compressed to the width of the outer ring. We use dimensionality reduction to create angles: the task of the layout algorithm is to place keywords so that keywords with neighboring angles have neighboring topical features. We quantify the goodness of the result as the sum $(\sum_s D_{KL}(p_i, q_i) + \sum_s D_{KL}(q_i, p_i))/2$ of Kullback-Leiblers divergences between angular neighborhoods q_i and topical neighborhoods p_i, which quantifies with equal interest both misses and false neighbors. This total divergence is a function of the angles a_i of keywords in the outer ring: we optimize the a_i by gradient descent to minimize the total divergence. This method is based on [40] (see also [22, 31] for discussion and variants) but in a new context; the difference is that the work in [40] is noninteractive visualization of fixed data sets with no integration of information retrieval, whereas in our work data arise from momentary search result sets; topical input features are extracted from the search system instead of using fixed features, by feedback based estimation of relevance and association features; and a radial output is used that couples dimensionality reduction output with estimated relevances. This layout approach can be shown [40] to correspond to *optimizing information retrieval of neighboring keywords from the display layout* (minimizing misses and false positives of such retrieval).

Highlighting of keywords in the outer circle. To highlight the structure in the outer circle layout, we apply a simple agglomerative clustering to angles a_i of keywords in the outer circle. In detail, start a cluster from the keyword with the smallest angle, and iteratively add the keyword with the next largest angle into the cluster as long as the angle difference is below a treshold and the size of the cluster is smaller than a specified percentage of all keywords in the outer circle; when either condition fails start the next cluster. We show clusters with different colors, and show for each cluster the label of the predicted most relevant keyword (having largest r_i).

Stage 3: Layout of Inner Keywords

The top keywords in the inner area represent the current search intent; for each such top keyword k_l, its radius naturally represents its current estimated relevance $r_l \in [0,1]$. The angles a_l of keywords in the inner circle must be placed consistently with the layout of the outer ring. Since we represent topical features in the outer ring by estimated associations with the top keywords k_l, that is, keywords k_i in the outer ring have features $\mathbf{x}_i = \{x_{il}\}$ where x_{il} is an association strength to the lth top keyword, then the angle a_l should act as a signpost to represent which outer-ring keywords are most associated with k_l. For the lth top keyword we thus set a_l to the highest weighted mode of angles a_i of outer-ring keywords k_i, where the angle of each outer-ring keyword is weighted by the association strength x_{il}. The resulting angle a_l of each keyword k_l in the inner area indicates which outer-ring keywords should be explored in relation to the lth top keyword.

Scalability. Results were retrieved by scalable Lucene search. The interface and visualization are computed for top-300 articles and their keywords; top-k pruning is simple as results were already ranked. Visualization optimization is fast and can be made $O(N * log(N))$ with respect to number of keywords N by recent strategies ([44]; see also [45, 46]).

EXPERIMENTS

The purpose of the experiment was to measure the effect of the visualization component to the user's information comprehension performance. The experiment followed a within-subject design. The independent variable of the experiment was the system configuration: a system with the visualization component and a system without the visualization component. The order of presentation of the systems and comprehension tasks was randomized and counterbalanced.

System Configurations

Two systems were configured: a system with the visualization component and a system without the visualization component. The system with the visualization component is illustrated in Figure 4. The system without the visualization component is exactly the same, but pertains only the conventional search result listing and the visualization component is removed (left side of Figure 5). Both systems had the keywords visualized within search results (i.e. a list of keywords describing the document was placed under each result). The systems were augmented with a workspace that was used by the participants to collect the information. This allowed a simple interaction to select information by dragging from the actual interface without switching to another application. The workspace also enabled accurate logging and data collection.

Task and Topics

The participants were situated in a simulated work task in which they had to comprehend and summarize the search results. The participants were asked to use two-level hierarchical conceptualization:
1. Find as many *main topic keywords*, but at least two, that you find important to cover the overall topic.
2. Find as many *subtopic keywords* under each main keyword that you find important to cover the main keyword.

The work task scenario was: "You are searching information about a pre-defined topic using an information retrieval system. Your task is to comprehend the topic by describing, at least two, main concepts related to the overall topic and describe as many as possible subconcepts related to each main concept." Eight topics were used: Human Memory, Web Design, Cognition, Distributed Systems, Language Processing, Kernel Function, Wearable Sensors, and Compiler Design.

Participants

We recruited 24 participants from two universities. Six were females. Participants reported their age within these age intervals: Seven participants were 17-24 years old, eleven were 25-32 years old, and three were 33-40 years old. Three participants had a PhD degree, eight had a MSc degree and the rest had an undergraduate degree. As the text in the user interface was in English, only participants with a self-reported good knowledge of English were eligible to take part. Participants were told they could ask the experimenter for clarification at any time during the experiment. All participants had a limited experience with interactive search engines. Users were recruited by word of mouth and received no compensation for participation.

Procedure

Prior to the experiment, the participants were asked to read written instructions. The instructions explained the purpose of the experiment and the the task that the participants were expected to perform. The participants were then informed that

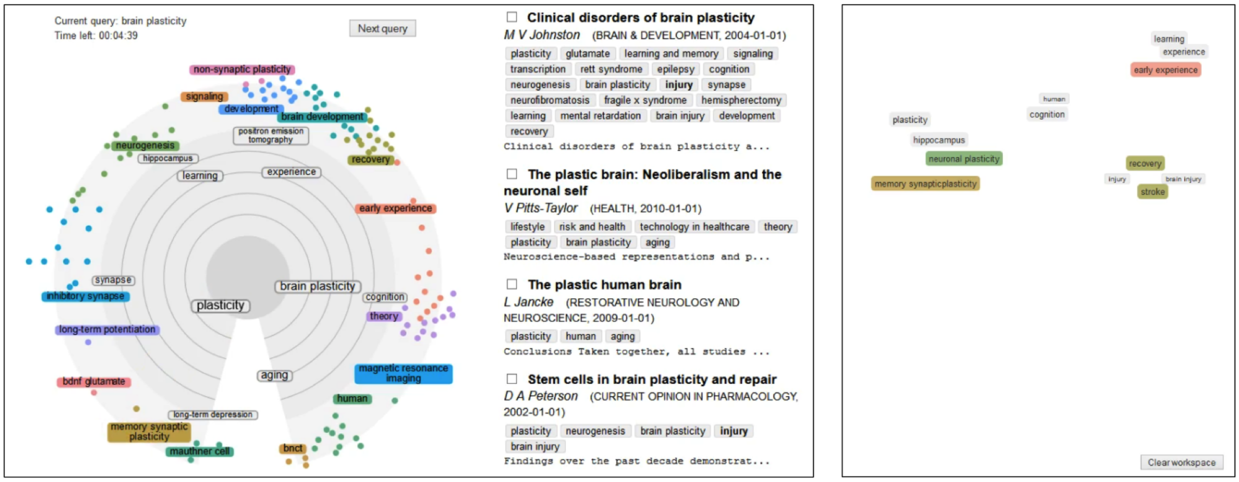

Figure 4. A screenshot of the user interface that was used by the participants. The visualization component and the ranked list of search results (left) and the workspace (right). In the experiment, the workspace was placed under the visualization component as a floating element to ensure equal screen estate with the baseline system.

the system would automatically launch queries and return and present search results, and the participant was only expected to gather information by examining the presented search results using the given system. The participants were informed that they will use two different *systems* to gather information and store their conceptualization in a *workspace* component, which will be the same for both systems. Then the participants watched a 2-minute video illustrating an exemplar task with the system variants and they had a 3-minute trial using the systems via a pre-defined query that was different form the ones used in the actual experiment.

After this phase, the actual experiment started. Participants performed eight tasks corresponding to the eight topics (see the previous subsection for the list of topics). Each task had two phases: comprehension phase and composition phase.

The rationale of the comprehension phase was to study the quality of the concepts that the participants were able to produce given a system variant. The rationale in separating the composition phase was to let the users to concentrate in producing the conceptualization as fast as possible and to allow them to organize the selected concepts in a separate composition phase. Previous research has shown that humans often spend significant amount of time in composing their answers instead of looking for information to support their answers and that these two tasks are interleaved and cause task-switching costs [30]. The experimental design where these tasks were separated ensured that the participants focused on collecting the best possible concepts comprehensing the result space in the given time without having to interleave this activity with composition of their answer.

In the comprehension phase, the system automatically launched queries corresponding to the tasks by executing a client side Java Script code. Each query run either on timer (5 minutes), or when the participant clicked the "next" button. The query that was automatically issued by the system was exactly the name of the topic. For example, for the topic *Web Design*, the system automatically issued a query "Web design".

This allowed to remove possible variance originating from participants' subjective interpretations of the topics. The time limit was used to make sure that there was no variance in the time that the participants used in the comprehension phase. The timer was visible for the participants so that they were aware of how much time they had left to complete the task. The time was limited to two minutes for each task. The participants had two minutes to read and collect the information on the screen and after two minutes all the information on the screen disappeared, except the information on the workspace which was used to store the results to be used in the composition phase.

In the composition phase, the participants could still use the information that they had collected to the *workspace* to compose a written answer that comprehended the search result space. The workspace was visible for additional 3 minutes.

Such experimental procedure ensured that the participants were working under strict time limits in order to complete the task as fast as possible and use the preferred interface element when they knew that their time is limited.

After the experiment, the participant filled in post-task questionnaires selected from the ResQue questionnaires [33].

Apparatus

The experiment was run on a desktop PC connected to a vertically mounted 24-inch wide-screen monitor. The vertical position of the monitor was chosen because the workspace was placed under the result list and visualization component and the additional screen estate allowed fair comparison to the baseline system. The system was implemented as a Web application accessed using the Google Chrome browser. During the experiment, participants could use a mouse and a keyboard to operate the interface. The search engine automatically logged the timestamp, the action performed by the user: selected keyword to represent the main or subtopic, the position of the documents in the ranked list that contained the particular keyword, and the state of the workspace. Participants' subjec-

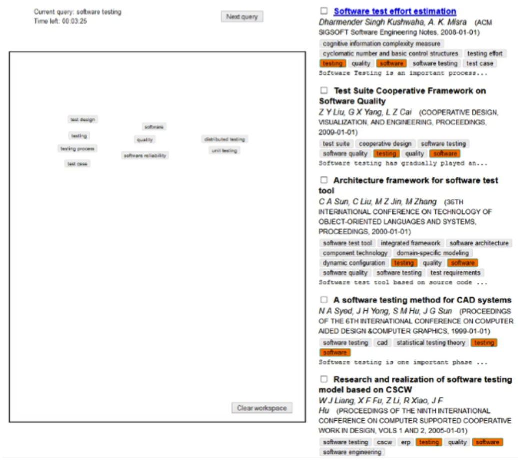

Figure 5. A screenshot of the baseline system without the visualization component. The ranked list of search results (right) and the workspace (left).

tive evaluations of the systems were recorded by means of a post-use questionnaire presented on the PC.

Search Engine Index and Data

The search engine was built using the Apache Lucene (version 4.1)[1] and indexed over 50 million scientific documents from these sources: The Web of Science prepared by THOMSON REUTERS, Inc., the digital library of the Association of Computing Machinery (ACM), the Digital Library of the Institute of Electrical and Electronics Engineers (IEEE), and the digital library of Springer. The dataset contains the following information about each document: title, abstract, keywords, author names, publication year and publication forum. Both system variants used the same document set. Both variants ranked the documents using a unigram language model with Dirichlet smoothing [27, 47]. The number of retrieved documents at each iteration by the language model ranking was set to 100 and we used $\mu = 2000$ for the Dirichlet smoothing.

Relevance Assessments and Ground Truth

After the experiment all responses from all participants and systems were pooled so that each *main topic keywords* and each *subtopic keyword* associated with the main topic keyword were listed in a matrix. Two assessors assessed the relevance of the main topic keywords and the subtopic keywords using a graded relevance on a 5-point Likert scale. Scores were created by one expert and checked by another, resolving disagreements by consensus.

Research Questions

The focus of the study was twofold. First, to study if the visualization would assist users in the comprehension process. Second, to study if the visualization would improve the output of the comprehension process. We defined the following research questions:
Comprehension process: Did the participants inspect the search result space using the topic-relevance map more often than using the result list? Did the participants select keywords from the topic-relevance map visualization more often than from the result list?
Comprehension outcome: Did the topic-relevance map result in improved comprehension outcome?

[1] https://lucene.apache.org/core/4_1_0/

Measures

We used the following measures for the *Comprehension process*. *Fluency of keyword selection:* how many keywords are selected from the visualization compared to the result list? *Share of information inspection source:* how much time is spent browsing the visualization compared to the result list? *Share of information selection source:* how many keywords are selected from the visualization compared to the result list?

We used the following measures for the *Comprehension outcome*. *Cumulative gain of selected keywords:* how good quality are the selected keywords. *Distribution of the selected keywords in the result list:* how comprehensively the result list is covered. *Subjective preference:* how useful and easy to use were the compared systems.

RESULTS

Fluency of keyword selection

Users of the baseline system dragged in total 1286 keywords from under articles in the result list to the workspace (on average 6.7 per user and task). Users of the system with the topic-relevance map dragged a comparable total number, 1068 keywords to the workspace (on average 5.6 per user and task). Figure 7 shows the result graphically.

Share of information inspection source

Recorded mouse movements over time were available from 18 of the users (mouse-movements of 6 users were not available due to a technical problem). Figure 6 shows the locations as a scatter plot. Based on the mouse cursor locations, users on the system with the topic-relevance map spent 30.5% of time browsing the map, and 55.4% of time browsing the search result list. In comparison, users on the baseline system spent 83.4% of time browsing the search result list. This further illustrates the fluency of the topic-relevance map: users spent a reasonable portion of time browsing the map, and while less time was spend over the map than over the familiar result list, nevertheless the majority of keywords chosen to be dragged were ultimately dragged from the map as discussed below.

Share of information selection source

In the baseline system keywords can only be dragged to the workspace from the list of document results. Users of the system with the topic-relevance map strongly preferred to use the map: 914 (85.6%) of the dragged keywords were dragged from the map to the workspace, and 154 (14.4%) were dragged from under articles in the result list to the workspace. Figure 7 shows the result graphically.

Distribution of selected keywords

On average over the eight tasks, in the baseline system the 24 users overall dragged 55.5 unique keywords per task from the search result list onto the workspace. In the system with the topic-relevance map, the 24 users dragged 45.8 unique keywords onto the workspace, from the map onto the workspace

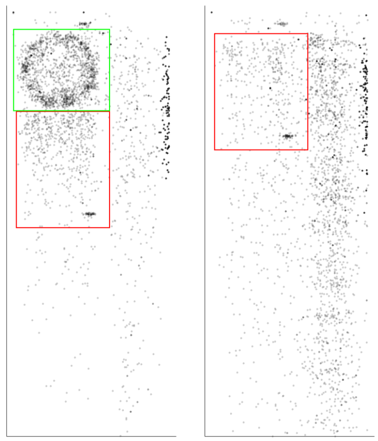

Figure 6. Mouse position scatter plots from 18 users over the two systems: the one with the topic-relevance map (left) and the baseline (right). Dots are mouse positions recorded at 3-second intervals. The areas of the workspace and topic-relevance map are outlined in red and green respectively. Dots in the bottom-half of the figures are from situations where the user has scrolled the screen to see more results.

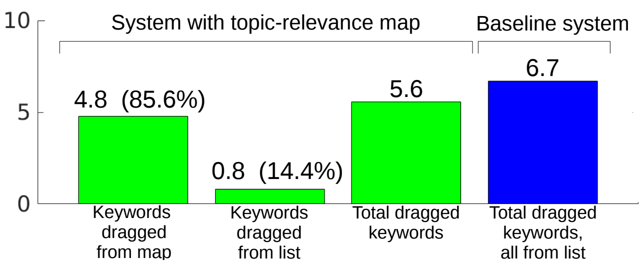

Figure 7. Sources of information selected to be dragged to the workspace. Numbers are average amounts of keywords dragged to the workspace, for both systems, on average over all tasks and users. For the system with the topic-relevance map, we also report separately the amounts of keywords dragged from the map and from the document list, and give their relative percentages. On the system with the map, users drag from the map most of the time.

and 120 unique keywords from the list; the difference is statistically significant over the tasks. Thus, even though users overall dragged slightly less keywords onto the workspace in the proposed system than in the baseline, the topic-relevance map allowed them to reach a more varied and comprehensive selection of the keywords than the baseline system. Since users of the topic-relevance map system gained a better expert score than users of the baseline system, this comprehensiveness was beneficial and not simply random variation.

On the system with the topic-relevance map, 89.8% of the unique keywords dragged to the workspace by the 24 users were dragged from the map, thus not only did the users prefer to use the map, it accounted for a majority of the comprehensive keyword selection they discovered.

Cumulative gain of selected keywords

Users had to provide answers organized as main-topic keywords and sub-topic keywords under each main topic, and both were evaluated separately by cumulative gain of expert-given scores for the keywords. Average within-task standard deviation over users was 3.9 for main-topic scores and 17.2

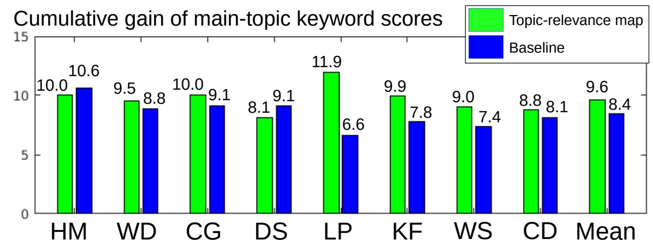

Figure 8. Cumulative gain of main-topic keyword scores from experts. Numbers are cumulative-gain scores averaged over the users, for each task and each system. The tasks are: Human memory (HM), Web design (WD), Cognition (CG), Distributed systems (DS), Language processing (LP), Kernel functions (KF), Wearable sensors (WS), Compiler design (CD). The rightmost bars are the mean over all tasks per system. The system with the topic-relevance map is statistically significantly better than the baseline system, by right-tailed two-sample t-test at the $p = 0.05$ threshold ($p = 0.0192$).

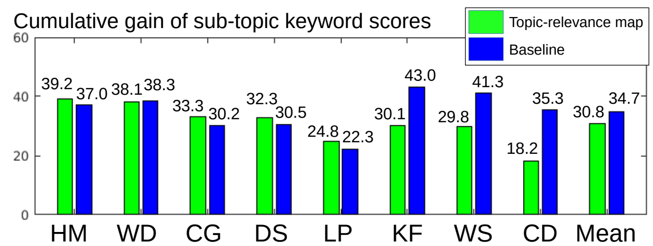

Figure 9. Cumulative gain of sub-topic keyword scores from experts, with respect to their corresponding main-keywords. Numbers are cumulative-gain scores averaged over the users, for each task and each system. The tasks are the same as in Figure 8, and the rightmost bars are the mean over all tasks per system. The overall difference between systems is not statistically significant.

for sub-topics; we focus on between-system difference. Figure 8 shows the topic-relevance map yielded a statistically significantly improved score for main-topic keywords; since the main topics represent the breadth of information content discovered from the results, the topic-relevance map helped users reach a more comprehensive understanding of the results. Figure 9 shows the results for sub-topic keywords, representing depth of understanding for each main topic; the difference between systems was not statistically significant, hence the topic-relevance map increased overall comprehension without sacrificing depth of comprehension.

Subjective preference

In the post-task questionnaires (Table 1), users indicated through ratings of several questions that the simpler and more familiar interface was found easier to use, learn, and they felt more confident using the conventional system with only search result listing. This is natural when comparing a traditional interface to a new one with only a small amount of training time. However, users clearly felt the topic-relevance map influenced their selection of topics and subtopics. While users' overall satisfaction score for systems was similar, in a separate question about system preference, a two thirds majority of the 24 users preferred to use the system with the topic-relevance map. In summary, the subjective user experience with the topic-relevance map was found more cumbersome than with the conventional baseline. However, it influenced participants'

Question	TM	BL	P-value
I found the system unnecessarily complex	2.5	**1.7**	0.004
I thought the system was easy to use	3.8	**4.4**	0.002
I think that I would need the support of a technical person to be able to use this system	1.9	**1.5**	0.03
I found the various functions in this system were well integrated	3.1	**3.6**	0.03
I thought there was too much inconsistency in this system	2.8	**1.9**	0.0002
I would imagine that most people would learn to use this system very quickly	3.3	**4.2**	0.003
I found the system very cumbersome to use	2.9	**2.1**	0.04
I felt very confident using the system	3.0	**3.8**	0.01
I needed to learn a lot of things before I could get going with the system	2.3	**1.7**	0.03
The system can be trusted	3.0	**3.8**	0.003
I became familiar with the system very quickly	3.7	**4.4**	0.008
The labels/keywords/information provided by the system are clear	3.0	**3.7**	0.03
The system influenced my selection of topics and subtopics	**4.1**	3.0	$2 \cdot 10^{-4}$
Which system do you prefer?	**16**	8	

Table 1. Post-task questionnaires, selected from the ResQue questionnaires, in which significant differences were found. Numbers are 1-5 Likert scale agreement scores (higher=stronger agreement) with the statements in the question column, averaged over the 24 users for each system: Topic-relevance map (TM), Baseline (BL), and the t-test p-value of the difference. The better score for each question is in bold. The last line is the question of system preference, where we directly list how many users preferred each system; 67% preferred the topic-relevance map.

selections of topics and subtopics. Together with the improved comprehension outcome, this result suggests that users experience additional cognitive load using the topic-relevance map, but it yields improved task outcome.

DISCUSSION AND CONCLUSIONS

Information spaces grow in size and richness and information searches are increasingly conducted to explore and learn as opposite to only looking up information. As a consequence, conventional search interfaces fall short in supporting comprehension of large search result spaces.

We introduced topic-relevance map, a visualization that assists users in comprehension of search results. In contrast to visualization approaches that focus on grouping search results or visualizing topical relevance within search result listing, out approach is based on simultaneously visualizing topical relevance and topical similarity. We evaluated the effect of the topic-relevance map in a search result comprehension task where participants were summarizing search results and produced a conceptualization of the result space.

Answers to the Research Questions

Here we reflect the research questions that we defined earlier:

Did the participants inspect the search result space using the topic-relevance map more often than using the result list? Yes, the participants inspected the search results using the topic-relevance map. The recordings of the mouse movements showed that over one third of the time the users inspected the search results using the map.

Did the participants select keywords from the topic-relevance map visualization more often than from the result list? Yes, participants selected significantly more and more comprehensive set of keywords from the map than from the result list.

Did the topic-relevance map result in improved comprehension outcome? Yes, the topic-relevance map improved the comprehension outcome for the main topic keywords, but no overall difference was found in case of the subtopic keywords. This suggests that the visualization enabled participants to obtain a broader view on the search results, but did not help gather better information under an individual subtopic.

Limitations and Future Work

While our work shows novel visualization that was found to improve search result comprehension over a baseline that used within search result visualization, it could be extended in several ways. We only studied the search result comprehension task; a larger variety of tasks, such as diversity or novelty of information, or topic spotting, could allow measuring a wider set of objectives that could be important for the users. We compared our method to a within search results visualization that had exactly the same keywords visualized under each search result; while this visualization has become the de-facto method in many search engines, comparison to larger variety of baselines and other visualizations could reveal additional benefits or weaknesses of our approach. We restricted the time for the tasks to few minutes; while this lets us mimic a time constrained result comprehension task, the time constraint was arbitrarily chosen; different constraint types (time or other task constraints), or comprehension across a series of searches, could yield valuable information about possible tradeoffs the visual complexity may impose.

Nevertheless, our results show that topic-relevance map significantly improves participants' comprehension capability compared to a conventional ranked list presentation. In general, our results echo the recent developments in visual search engines that promote human control over search results, personal data processing, and more generally putting users at the center and in control of their information search processes.

ACKNOWLEDGMENTS

Authors belong to HIIT and JP and TR to COIN. The work was funded in part by TEKES (Re:Know2 project) and Academy of Finland decisions 252845, 256233, 295694, 305739, and 278090.

REFERENCES

1. Christopher Ahlberg and Ben Shneiderman. 1994. Visual Information Seeking: Tight Coupling of Dynamic Query

Filters with Starfield Displays. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '94)*. ACM, New York, NY, USA, 313–317.

2. Jae-wook Ahn, Peter Brusilovsky, and Shuguang Han. 2015. Personalized Search: Reconsidering the Value of Open User Models. In *Proceedings of the 20th International Conference on Intelligent User Interfaces (IUI '15)*. ACM, New York, NY, USA, 202–212.
3. Eric C. Alexander, Joe Kohlmann, Robin Valenza, Michael Witmore, and Michael Gleicher. 2014. Serendip: Topic model-driven visual exploration of text corpora.. In *IEEE VAST*, Min Chen, David S. Ebert, and Chris North (Eds.). IEEE Computer Society, 173–182.
4. S. Andolina, K. Klouche, J. Peltonen, M. Hoque, T. Ruotsalo, D. Cabral, A. Klami, D. Glowacka, P. Floreen, and G. Jacucci. 2015. IntentStreams: Smart Parallel Search Streams for Branching Exploratory Search. In *Proc. IUI'15*. 300–305.
5. Fedor Bakalov, Marie-Jean Meurs, Birgitta König-Ries, Bahar Sateli, René Witte, Greg Butler, and Adrian Tsang. 2013. An Approach to Controlling User Models and Personalization Effects in Recommender Systems. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces (IUI '13)*. ACM, New York, NY, USA, 49–56.
6. Michelle Q. Wang Baldonado and Terry Winograd. 1997. SenseMaker: An Information-exploration Interface Supporting the Contextual Evolution of a User's Interests. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI '97)*. ACM, New York, NY, USA, 11–18.
7. Christine Susan Bruce. 1999. Workplace experiences of information literacy. *International Journal of Information Management* 19, 1 (1999), 33 – 47.
8. A. J.-B. Chaney and D. M. Blei. 2012. Visualizing topic models. In *Proceedings of ICWSM*. 419–422.
9. CHAOMEI CHEN and YUE YU. 2000. Empirical Studies of Information Visualization. *Int. J. Hum.-Comput. Stud.* 53, 5 (Nov. 2000), 851–866.
10. Weiwei Cui, Shixia Liu, Li Tan, Conglei Shi, Yangqiu Song, Zekai Gao, Huamin Qu, and Xin Tong. 2011. TextFlow: Towards Better Understanding of Evolving Topics in Text. *IEEE Transactions on Visualization and Computer Graphics* 17, 12 (Dec. 2011), 2412–2421.
11. Cecilia di Sciascio, Vedran Sabol, and Eduardo E. Veas. 2016. Rank As You Go: User-Driven Exploration of Search Results. In *Proceedings of the 21st International Conference on Intelligent User Interfaces (IUI '16)*. ACM, New York, NY, USA, 118–129.
12. Karthik Dinakar, Jackie Chen, Henry Lieberman, Rosalind Picard, and Robert Filbin. 2015. Mixed-Initiative Real-Time Topic Modeling & Visualization for Crisis Counseling. In *Proceedings of the 20th International Conference on Intelligent User Interfaces (IUI '15)*. ACM, New York, NY, USA, 417–426.
13. M. Dork, N.H. Riche, G. Ramos, and S. Dumais. 2012. PivotPaths: Strolling through Faceted Information Spaces. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (Dec. 2012), 2709–2718.
14. Marian Dörk, Carey Williamson, and Sheelagh Carpendale. 2012. Navigating Tomorrow's Web: From Searching and Browsing to Visual Exploration. *ACM Trans. Web* 6, 3, Article 13 (Oct. 2012), 28 pages.
15. W. Dou, L. Yu, X. Wang, Z. Ma, and W. Ribarsky. 2013. HierarchicalTopics: Visually Exploring Large Text Collections Using Topic Hierarchies. *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (Dec 2013), 2002–2011.
16. Geoffrey M. Draper, Yarden Livnat, and Richard F. Riesenfeld. 2009. A Survey of Radial Methods for Information Visualization. *IEEE Transactions on Visualization and Computer Graphics* 15, 5 (2009), 759–776.
17. Matthew J Gardner, Joshua Lutes, Jeff Lund, Josh Hansen, Dan Walker, Eric Ringger, and Kevin Seppi. 2010. The Topic Browser an Interactive Tool for Browsing Topic Models.
18. Marti A. Hearst. 1995. TileBars: Visualization of Term Distribution Information in Full Text Information Access. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '95)*. ACM Press/Addison-Wesley Publishing Co., New York, NY, USA, 59–66.
19. Marti A. Hearst. 1997. TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages. *Comput. Linguist.* 23, 1 (March 1997), 33–64.
20. Marti A. Hearst. 2009. *Search User Interfaces* (1st ed.). Cambridge University Press, New York, NY, USA.
21. Marti A. Hearst and Jan O. Pedersen. 1996. Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '96)*. ACM, New York, NY, USA, 76–84.
22. S. Kaski and J. Peltonen. 2011. Dimensionality Reduction for Data Visualization. *IEEE Signal Processing Magazine* 28, 2 (2011), 100–104.
23. Yanir Kleiman, Joel Lanir, Dov Danon, Yasmin Felberbaum, and Daniel Cohen-Or. 2015. DynamicMaps: Similarity-based Browsing Through a Massive Set of Images. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 995–1004.
24. Gary Klein, Brian Moon, and Robert R. Hoffman. 2006. Making Sense of Sensemaking 1: Alternative Perspectives. *IEEE Intelligent Systems* 21, 4 (July 2006), 70–73.

25. Khalil Klouche, Tuukka Ruotsalo, Diogo Cabral, Salvatore Andolina, Andrea Bellucci, and Giulio Jacucci. 2015. Designing for Exploratory Search on Touch Devices. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 4189–4198.

26. Bill Kules, Robert Capra, Matthew Banta, and Tito Sierra. 2009. What Do Exploratory Searchers Look at in a Faceted Search Interface?. In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '09)*. ACM, New York, NY, USA, 313–322.

27. John Lafferty and Chengxiang Zhai. 2001. Document Language Models, Query Models, and Risk Minimization for Information Retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '01)*. ACM, New York, NY, USA, 111–119.

28. S. Liu, X. Wang, J. Chen, J. Zhu, and B. Guo. 2014. TopicPanorama: A full picture of relevant topics. In *Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on*. 183–192.

29. Gary Marchionini. 2006. Exploratory Search: From Finding to Understanding. *Commun. ACM* 49, 4 (April 2006), 41–46.

30. Stephen Monsell. 2003. Task switching. *Trends in Cognitive Sciences* 7, 3 (2003), 134–140.

31. J. Peltonen and S. Kaski. 2011. Generative Modeling for Maximizing Precision and Recall in Information Visualization. In *Proc. AISTATS'11*. 579–587.

32. J. Peltonen, J. Strahl, and P. Floreen. 2017. Negative Relevance Feedback for Exploratory Search with Visual Interactive Intent Modeling. In *Proc. IUI'17*.

33. Pearl Pu, Li Chen, and Rong Hu. 2011. A User-Centric Evaluation Framework for Recommender Systems. In *Proceedings of the fifth ACM Conference on Recommender Systems*. ACM, New York, NY, USA, 157–164.

34. S. E. Robertson. 1997. Readings in Information Retrieval. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, Chapter The Probability Ranking Principle in IR, 281–286.

35. Tuukka Ruotsalo, Giulio Jacucci, Petri Myllymäki, and Samuel Kaski. 2014. Interactive Intent Modeling: Information Discovery Beyond Search. *Commun. ACM* 58, 1 (Dec. 2014), 86–92.

36. Tuukka Ruotsalo, Jaakko Peltonen, Manuel Eugster, Dorota Glowacka, Ksenia Konyushkova, Kumaripaba Athukorala, Ilkka Kosunen, Aki Reijonen, Petri Myllymäki, Giulio Jacucci, and Samuel Kaski. 2013. Directing Exploratory Search with Interactive Intent Modeling. In *Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management (CIKM '13)*. ACM, New York, NY, USA, 1759–1764.

37. Marc M. Sebrechts, John V. Cugini, Sharon J. Laskowski, Joanna Vasilakis, and Michael S. Miller. 1999. Visualization of Search Results: A Comparative Evaluation of Text, 2D, and 3D Interfaces. In *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99)*. ACM, New York, NY, USA, 3–10.

38. Ben Shneiderman. 1996. The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the IEEE Symposium on Visual Languages*. IEEE.

39. Jaime Teevan, Christine Alvarado, Mark S. Ackerman, and David R. Karger. 2004. The Perfect Search Engine is Not Enough: A Study of Orienteering Behavior in Directed Search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '04)*. ACM, New York, NY, USA, 415–422.

40. Jarkko Venna, Jaakko Peltonen, Kristian Nybo, Helena Aidos, and Samuel Kaski. 2010. Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *Journal of Machine Learning Research* 11 (2010), 451–490.

41. Katrien Verbert, Denis Parra, Peter Brusilovsky, and Erik Duval. 2013. Visualizing Recommendations to Support Exploration, Transparency and Controllability. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces (IUI '13)*. ACM, New York, NY, USA, 351–362.

42. C. Wongchokprasitti, J. Peltonen, T. Ruotsalo, P. Bandyopadhyay, G. Jacucci, and P. Brusilovsky. 2015. User Model In a Box: Cross-System User Model Transfer for Resolving Cold Start Problems. In *Proc. UMAP'15*. 289–301.

43. Kanit Wongsuphasawat, Dominik Moritz, Anushka Anand, Jock Mackinlay, Bill Howe, and Jeffrey Heer. 2016. Voyager: Exploratory Analysis via Faceted Browsing of Visualization Recommendations. *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)* (2016).

44. Z. Yang, J. Peltonen, and S. Kaski. 2013. Scalable Optimization of Neighbor Embedding for Visualization. In *Proc. ICML'13*. 127–135.

45. Z. Yang, J. Peltonen, and S. Kaski. 2014. Optimization Equivalence of Divergences Improves Neighbor Embedding. In *Proc. ICML'14*. 460–468.

46. Z. Yang, J. Peltonen, and S. Kaski. 2015. Majorization-Minimization for Manifold Embedding. In *Proc. AISTATS'15*. 1088–1097.

47. Chengxiang Zhai and John Lafferty. 2004. A Study of Smoothing Methods for Language Models Applied to Information Retrieval. *ACM Trans. Inf. Syst.* 22, 2 (April 2004), 179–214.