Visual Re-Ranking for Multi-Aspect Information Retrieval

Khalil Klouche[1,3], Tuukka Ruotsalo[2], Luana Micallef[2]
Salvatore Andolina[2], Giulio Jacucci[1,2]
[1]Helsinki Institute for Information Technology HIIT, Department of Computer Science,
University of Helsinki, PO Box 68, 00014 University of Helsinki, Finland
[2]Helsinki Institute for Information Technology HIIT, Aalto University,
PO Box 15600, 00076 Aalto, Finland
[3]Aalto University, School of Arts, Design and Architecture, Media Lab Helsinki,
Hämeentie 135 c, 00560 Helsinki, Finland
[1]first.last@helsinki.fi, [2]first.last@aalto.fi

ABSTRACT

We present visual re-ranking, an interactive visualization technique for multi-aspect information retrieval. In multi-aspect search, the information need of the user consists of more than one aspect or query simultaneously. While visualization and interactive search user interface techniques for improving user interpretation of search results have been proposed, the current research lacks understanding on how useful these are for the user: whether they lead to quantifiable benefits in perceiving the result space and allow faster, and more precise retrieval. Our technique visualizes relevance and document density on a two-dimensional map with respect to the query phrases. Pointing to a location on the map specifies a weight distribution of the relevance to each of the query phrases, according to which search results are re-ranked. User experiments compared our technique to a uni-dimensional search interface with typed query and ranked result list, in perception and retrieval tasks. Visual re-ranking yielded improved accuracy in perception, higher precision in retrieval and overall faster task execution. Our findings demonstrate the utility of visual re-ranking, and can help designing search user interfaces that support multi-aspect search.

Keywords

Information visualization; information retrieval; multi-aspect search; multi-dimensional ranking

1. INTRODUCTION

Multi-aspect search refers to activities in which the information need of the user consists of more than one aspect or query simultaneously. Such situation arises in contexts such as exploratory search, item selection and multi-criteria decision making. In exploratory search activities, the user's goal is not clearly defined, and the information space is usually unfamiliar to the user. In such scenarios, the user might start from a small set of notions, with the intent of learning and making sense of the related document space. In this case, conventional result lists offer little insight of the data and nothing indicates how the given results relate to the multiple aspects of the query. For example, a user looking for recent literature on physiological measurements might want to search for aspects such as 'Electroencephalography', 'Electrodermal Activity', 'Electromyography' and quickly be able to assess how the result space is distributed and how the retrieved documents relate to each aspect.

Item or product selection is currently widely supported by faceted search and search result clustering. Such systems are widespread in e-commerce and library catalogs. These techniques allow the user to investigate the results through the use of multiple filters, but they offer limited support for perceiving the result space and weighting the aspects accordingly. Conventional query-based search tools usually visualize results as a one-dimensional ranked list, and offer limited support for multi-aspect retrieval. Another example is multi-criteria decision making, a well researched process that often requires multi-aspect search [8]. Take the example of a user looking online for a new car. Usual faceted tools allow her to select filters to narrow down the offering: e.g., a manufacturer, a price range, a fuel type. Such criteria require the user to have

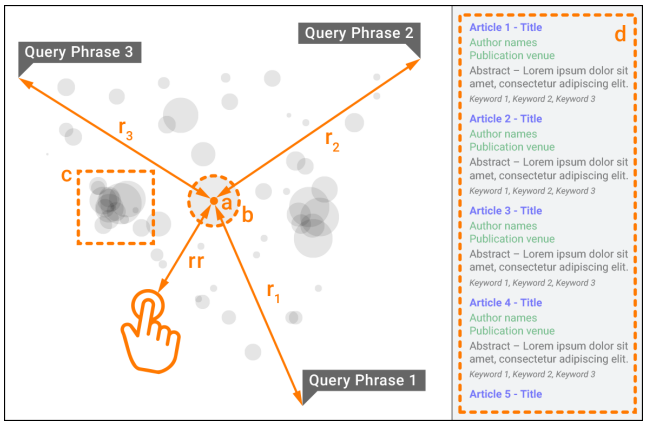

Figure 1: Interactive relevance map visualization. (a) Position of a document marker is computed as a weighted linear combination of relevance to individual query phrases $r1, r2, r3$. (b) Radius of a document marker encodes the overall relevance of the corresponding document to all query phrases. (c) Opacity encodes the density of document mass in a certain position of the 2D plane. (d) The result list can be re-ranked by relevance and the distance to the selected position rr.

CHIIR '17, March 07-11, 2017, Oslo, Norway

DOI: http://dx.doi.org/10.1145/3020165.3020174

a specific goal in mind, whereas a typical user would be inclined to come up with more vague criteria such as: *good gas-mileage* (no specific threshold in mind), *family-friendly*, and/or *fun to drive*. Such criteria are not binary, and the user can expect to find on the market several satisfying solution with different tradeoffs, instead of one ideal car. On the other hand, looking for such criteria using one single unified query on a conventional search engine returns a list of results that does not reflect the user's preferences and does not allow for conscious tradeoffs.

In all these cases, the user should be able to quickly assess distribution of the results with respect to how they relate to each researched aspect, and then be able to rapidly inspect them, which is possible if the user 1) perceives the distribution to understand which parts of the result space contain interesting information (i.e. what are the tradeoffs between the query phrases) and 2) is able to determine the tradeoff rapidly using the visualization.

We present a visual re-ranking technique that uses multi-dimensional ranking and two-dimensional interactive visualization. Inspired by earlier work on visual information retrieval and seeking [33, 2], the technique allows the user to perceive the relevance distribution with respect to multiple query phrases by using a *relevance map* visualization. A novel feature of this technique is that it allows the user to investigate specific areas on the map by re-ranking the results through pointing at the map. The method estimates document relevance with respect to user-specified query phrases in a multi-dimensional space in which the query phrases define the dimensionality. The method then computes a layout for the documents on a two-dimensional plane where relative distances of document markers to each query phrases are defined by their respective relevance, overall relevance of each document is visualized as the radius, and higher document density translates in darker areas. (see Figure 1). The visualization allows the user to perceive how the result space is populated with respect to both density and relevance to some query phrases.

Rather than relying only on a one-dimensional ranking algorithm to select the documents most relevant to a query, the role of the system is to organize and present information about many documents and multi-dimensional query phrases in a way that makes comparison possible. Re-ranking by pointing allows users to rank documents with respect to relative relevance weights to the query phrases. For example, expressing that a user wants the ranking to be based a little on both query phrases *interaction* and *interfaces*, but mainly on the phrase *design* can be done simply by pointing to an area on the map that is inside a triangle of the query phrases but closer to the concept *design*.

The approach was evaluated in a controlled laboratory study with 20 participants performing two tasks: perception and retrieval. In the perception task, participants were asked to find out how a document space was populated and organized with respect to specific topics, such as whether there was more research about *interaction* or *design*. In the retrieval task, the participants were asked to find documents with varying relevance to several topics, such as a document that was mainly related to design, but slightly related to interaction and interfaces.

Our results show significant improvement in task completion time as well as improved accuracy in perception, and improvement in task completion time in retrieval, without compromising effectiveness measured as the quality of the task outcome. These results suggest that relevance mapping and re-ranking is effective in cases when the initial one-dimensional result list is not enough for the user to analyze the information.

The contributions of this paper are: (1) We present a visual re-ranking approach to multi-aspect information retrieval in which users can perceive the result space and rapidly re-rank the result list by pointing to the visualization. (2) We demonstrate that users can complete perception and re-ranking tasks significantly faster without compromising the effectiveness. (3) While different approaches for search result visualization have been proposed in the past, up to our knowledge, this is the first study that empirically verifies the benefits of interactive visualization for multi-aspect information retrieval.

2. RELATED WORK

2.1 Visual Information Retrieval and Seeking

Information spaces can be huge and thus hard to comprehend. However, visualizing the space and allowing the user to directly interact with and manipulate objects in the space facilitates comprehension. For instance, when the results of actions are shown immediately and when typing is replaced with pointing or selecting, exploration and retention increase while errors decrease [46]. For information seeking, the following visualization and interaction features are of particular importance [43]: (a) dynamic querying for rapid browsing and filtering to view how results change; (b) a starfield display for the immediate, continuous, scalable display of result sets as different queries are processed; (c) tight coupling of queries to easily use the output of one query as input to another [1]. For instance, a user study indicates that dynamic querying significantly improves user response time and enthusiasm. Using such techniques, systems like FilmFinder [1] support querying over multiple varying attributes such as time, while showing the changing query results in the context of the overall data. User studies also indicate that user interfaces that show the result list together with an overview of the result categories encourage a deeper and more extensive exploration of the information space [25], especially when the system allows relevance feedback to be given on such categories to direct the exploration [40, 39].

2.2 Document Collection Visualization

Various visualizations have been proposed for large document collections [24]. Most of these techniques adopt the visual information seeking mantra [44] to provide an overview at first and details only on demand. The documents are often visualized on a 2D plane, in the form of a map based on a similarity metric. Higher-level entities, such as topics, are also displayed on the map for immediate and better understanding of the document space organization.

Document Atlas [12] uses Latent Semantic Indexing and multi-dimensional scaling (MDS) to extract semantic concepts from the text and position the documents with respect to the concepts. Document densities around concepts are visualized as a heat map. On mouse hover, common keywords in the area are listed, and on zoom in, more details are shown.

Self-Organizing Maps have also been used by systems like WEBSOM [20] and Lin's maps [26] to position the documents on the 2D plane. WEBSOM also suggests areas in the map that could be relevant to the user's search query. Lin's maps are further split up into regions whose area indicates the number of documents with specific related terms.

Other techniques visualize the documents as glyphs to indicate additional inter-document relationships and metadata on the map (e.g., [38, 29]). Various metaphors have also been adopted; examples include the terrain metaphor, in which dense regions in the map are seen as mountains with valleys in between [9, 49]; the galaxy metaphor, in which documents are seen as stars in different constellations (document clusters) [16]; and the physical metaphor,

in which documents are considered to be moving particles and the inter-particle forces move similar documents closer to each other and dissimilar documents apart [10]. Visualizations with two dimensions and meaningful axes (e.g., categories vs. hierarchies [45], query results vs. query index [6, 23], production vs. popularity [1]) have also been proposed.

These visualizations provide an overview of the entire document collection, but they do not allow the user to direct and focus the exploration as required. A user-driven rather than a data-driven technique could be more helpful when searching for documents relevant to multiple keywords. To that end, such a technique should visualize the ranking of documents with respect to multiple keywords so the user can easily judge the relevance of documents to each of the keywords of interest [33]. However, most of the current techniques only visualize whether a document is relevant or not to a keyword using set visualizations [4], without showing the document's degree of relevance to each keyword.

2.3 Multi-Aspect Search

In multi-aspect search the information need of the user consists of more than one aspect or query simultaneously. As a consequence, an item in a collection needs to be ranked differently based on its multiple attributes. The Graphics, Ranking, and Interaction for Discovery (GRID) principles and the corresponding rank-by-feature framework state that interactive exploration of multi-dimensional data can be facilitated by first analyzing one- and two-dimensional distributions and then by exploring relationships between the dimensions, using multi-dimensional rankings to set hypotheses and statistics to confirm them [41]. However, comparing, analyzing and relating different ranks is difficult and requires an interactive visualization that supports the various requirements identified by Gratz et al. [13].

Multi aspect search support is provided in Song et al. [47], with the proposal of a strategy for multi-aspect oriented query summarization task. The approach is based on a composite query strategy, where a set of component queries are used as data sources for the original query. Similarly Kang et al. [19] propose a multi-aspect relevance formulation, but in the context of vertical search.

LineUp [13] is an interactive visualization that uses bar charts to support the ranking of objects with respect to multiple heterogeneous attributes. Stepping Stones [11] visualizes search results for a pair of queries, using a graph to show relationships between the two sets of results. Sparkler [14] allows to visually compare results sets for different queries on the same topic.Tilebars [15] visualizes the frequency of different words in various sections of documents as a heat map and ranks the documents accordingly. Similarly, HotMap uses a two-dimensional grid layout to augment a conventional list of search results with colors indicating how hot (relevant) specific search terms are with respect to the document [18]. Ranking cube [50] is a novel rank-aware cube structure that is capable of simultaneously handling ranked queries and multi-dimensional selections. RankExplorer [42] uses stack graphs for time-series data. Techniques for incomplete and partial data have also been proposed [22]. TreeJuxtaposer [31] was primarily devised to compare rankings.

For document collections, the vector space model could be used, such that each document and search query is a vector in a multi-dimensional space, each axis is a term, and the document position is determined by the frequencies of each term in that document (e.g., [36]). Visualizations of such a model could aid understanding of the document space, but more research is required, particularly for user-driven approaches that allow the user to specify the dimensions of interest [33].

2.4 User-driven Visualization

VIBE [33] is one of the most well-known user-driven multi-dimensional ranking visualization for large document collections. To indicate the subspace of interest, the user first enters two or more query terms, known as "points of interest" (POIs). POIs are then shown (as circles) on a 2D plane, together with documents (as rectangles) related to at least one POI, forming a map. The position of each rectangle indicates the relevance of the corresponding document to each of the POIs. The size of a rectangle indicates the relevance of that document to the search query. Citation details of documents selected from the map are listed; clicking on an item in the list opens the full document. Any time a POI is added, removed or moved, the map is updated accordingly. However, regions of the map with numerous close-by documents are not easily detectable because the rectangles are not color filled; using semi-transparent color filled shapes reduces overplotting [28] and facilitates perceptual ordering of different regions in the map by their density [27]. Also, documents are not re-ranked as the user navigates over the map.

Variants of VIBE include: WebVIBE [30], in which POIs act like magnets that attract documents containing related terms; VR-VIBE [7], which visualizes the space in 3D (for more space to view documents between POIs) and depicts relevance by color; and Adaptive VIBE [3], in which POIs are query terms (as in VIBE) but also user profile terms that are automatically extracted from user notes.

Similar to VIBE, GUIDO [32], DARE [53] and TOFIR [52] also allow users to specify POIs and display documents based on their relevance to the POIs. However, in GUIDO each POI is an axis (not an icon on a 2D plane) and documents are positioned based on their absolute rather than relative distances from the POIs. In DARE and TOFIR, relevance to POIs is indicated by both distance and angle.

Other user-driven systems, like combinFormation [21], TopicShop [5] and InfoCrystal [48], retrieve and display search results related to user-defined keywords but do not visualize the results' multi-dimensional ranks. Similarly, HotMap [18] supports a weighted re-ranking of the search results, but without leveraging a graphical interactive approach for specifying the weights. WordBars [17] also supports re-ranking of the search results, but uses additional terms extracted from the search results rather than relying on the query terms.

While similar techniques of mapping data to 2D visualization for better user interpretation have been proposed, the current research lacks understanding on (1) how useful these are for the user and (2) whether they lead to quantifiable benefits in specific tasks related to search activity. This work is the first to demonstrate a technique where the visualization can be effectively used for re-ranking search results. It is also the first that empirically verifies that users perceive the document space faster and are able to execute retrieval faster without compromising the quality of retrieved information.

3. RELEVANCE MAPPING

The method for relevance mapping is first illustrated with an overview from the user perspective. Then the computation of the layout and document visualization is explained.

3.1 Overview

Figure 2 shows an example of the relevance map visualization. Here, a user investigates a document space delimited by three query phrases with corresponding markers on the map: *design*, *interaction* and *interface*. A fourth query marker, *exploration*, is greyed out because it has been disabled to permit a temporary focus on the three remaining query markers. The user has positioned the

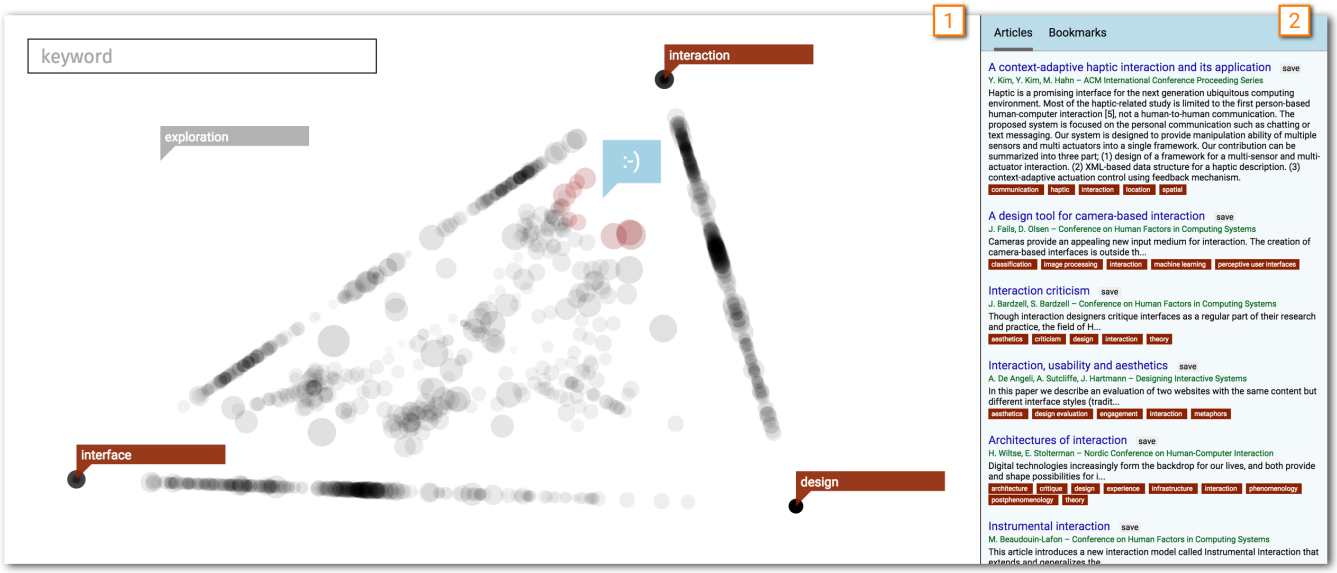

Figure 2: The relevance map (1) displays documents in relation to multiple query phrases, displayed as red text labels. Here, a fourth query phrase is greyed out (disabled). The exploration cursor (in blue) is located at the user-specified position to be used for the re-ranking. Red document markers indicate the position of articles currently on display in the result list (2).

pointer (blue flag with a smiley face) close to *interaction* to investigate a collection of documents highly related to *interaction* and more loosely related to *interface* and *design*. As a result, the list shows articles ranked with a specific focus on the selected area.

Query markers are created by inputting keywords in the query box in the top left. Each query marker can be activated or disabled by clicking it. Documents returned by the system are visualized on the map as semi-opaque dots scattered between the query markers with respect to their individual relevance. The overall relevance of a document is indicated by the radius of the dot. The partial opacity translates overlapping into a darkened tint that cues the user on the number of document markers in any given area. Query markers can be moved/dragged around on the map, which updates the position of the document markers. The position of the pointer can be positioned by dragging or tapping on the map. Any change in the pointer position or query marker organization triggers a re-ranking of documents based on their overall relevance and proximity to the pointer.

The ranked articles appear in a conventional one-dimensional list layout in the result list (2). Documents being displayed in the result list are shown as red dots on the map. The result list is scrollable. Each document is displayed with its title, authors, publication venue, abstract and keywords. Abstracts are first shown partially but can be displayed in full at a click or a tap. Keywords are interactive, as they can be added to the map as new query markers on a tap.

3.2 Layout

The data used to compute the relevance map layout consists of a set of m query phrases $q_{1...m} \in Q$, a set of k documents $d_{1...k} \in D$ and relevance estimates $r_{1...k} \in R$ for each of the k documents according to each of the m query phrases.

Each query marker and each document marker has a position on the plane, pos_{q_x}, pos_{q_y} and pos_{d_x}, pos_{d_y} respectively. The position of each query phrase marker is defined by the user by moving it to the desired position on the plane. The position of each of the document markers is computed as a weighted linear combination of the relevance scores to each query phrase and the relative position of the query marker. Intuitively, document markers are positioned proportional to their relevance to each of the query phrases. Formally, the position of an jth document marker on dimension dim is:

$$pos_{d_{j_{dim}}} = \frac{\sum_{i}^{|Q|} r_{q_i d_j} \cdot pos_{q_{dim}}}{|Q|} \quad (1)$$

so that $pos_{d_{i_{dim}}}$ is the coordinate of document d_i with respect to dimension dim. On a two-dimensional plane dim can be x or y. The relevance estimation $r_{q_i d_j}$ of a document to a query phrase is explained in the next section.

3.3 Document Marker Visualization

The radius of the document marker is directly the relevance $r_{q_i d_j}$. That is, the size of the dot is defined by the relevance.

The opacity of overlapping document markers is used to visualize the density of the document mass in a particular position on the plane. We use a standard computation of opacity [35] in which opacity of o of a pixel on the plane is computed as:

$$o = 1 - (1 - f)^n \quad (2)$$

where n is the number of overlapping layers and f is a constant setting of an opacity effect of an individual layer and was set to $f = 0.95$.

4. RELEVANCE ESTIMATION

The relevance estimation used in ranking and computing the document marker layout and size are explained in this section.

4.1 Relevance Estimation

Given the document collection and a set of query phrases that specify the multiple dimensions to be used in ranking and visualization, the relevance estimation method results in a set of probabilities $r_{1...k} \in R$ for each document d of k documents in the collection according to each query phrase $q_{1...m} \in Q$.

To estimate the probabilities from the query phrases Q and documents D, we utilize the language modeling approach of information retrieval [34]. We use a multinomial unigram language model. The vector Q of query phrases is treated as a sample of a desired document, and document d_j is ranked according to a query phrase q_i by the probability that q_i would be generated by the respective language model M_{d_j} for the document; with the maximum likelihood estimation we get

$$P(q|M_{d_j}) = \prod_{i=1}^{m} \hat{P}_{mle}(q_i|M_{d_j})^{w_i}, \tag{3}$$

where w_i is the weight of each of the query phrases and is set as $w_i = \frac{1}{|Q|}$ as default. In case of interactive re-ranking w_i is weighted based on user interactions as explained in the next section.

To estimate the relevance $r_{q_{i d_j}}$ of an individual document d_j with respect to an individual dimension defined by each query phrase q_i and avoid zero probabilities, we then compute a smoothed relevance estimate by using Bayesian Dirichlet smoothing for the language model so that

$$r_{q_{i d_j}} = P_{mle}(q_i|M_{d_j}) = \frac{c(q_i|d_j) + \mu p(q_i|C)}{\sum_k c(q|d_j) + \mu}, \tag{4}$$

where $c(d_i|d_j)$ is the count of a query phrase q_i in document d_j, $p(q_i|C)$ is the occurrence probability (proportion) of a query phrase q_i in the whole document collection, and the parameter μ is set to 2000 as suggested in the literature [51].

4.2 Ranking

Given the probability estimates for each of the documents, we apply a probability ranking principle [37] to rank the documents in descending order of their probabilities for the query phrases. These are then used to compute the total ordering of the document list. The top-k ranking computation remains efficient by making use of priority queue with complexity log(k) of k search results with presorted inverted index.

The user can interactively re-rank the result list by selecting a point on the relevance map. The point for the desired re-ranking is defined by its two-dimensional coordinates rr_x and rr_y with respect to the two-dimensional coordinates of the query markers $pos_{q_{i_x}}$ and $pos_{q_{i_y}}$ for the $i = 1 \ldots |Q|$ query phrases.

The re-rank weighting for an ith query marker is computed as the Euclidean distance between the $pos_{q_{i_x}}$ and $pos_{q_{i_y}}$ and the rr_x and rr_y. Formally,

$$w_i = \frac{\sqrt{(pos_{q_{i_x}} - rr_x)^2 + (pos_{q_{i_x}} - rr_y)^2}}{\sum_i^{|Q|} q_i}, \tag{5}$$

The re-ranking of the documents is then computed using these distances by Formula 3 by setting the weight w_i accordingly. Intuitively, the distance from the query marker is used as the importance of the query phrase in the ranking of the documents.

5. EXPERIMENTS

The current research lacks understanding on the end-user benefits of interactive visualization in multi-aspect search scenarios. The perceived simplicity and overall familiarity of well-studied conventional search system interfaces – like the current de-facto search interface with typed query and a ranked result list – have not been challenged in experiments that measure the quantifiable benefits of task completion time and effectiveness.

We conducted a controlled laboratory experiment in which the relevance mapping and re-ranking were compared to a conventional ranked list visualization in two basic tasks that searchers have to perform when using an information retrieval system: perception and retrieval.

The perception task sought understanding on the benefits of the visualization in perceiving the distribution and density of resulting documents with respect to the multi-aspect query phrases. The retrieval task sought understanding on the benefits of the visualization in re-ranking the results according to a user specified distribution over the importance of the different query phrases (see Figures 3b,3c_1, and 3c_2). The benefits were measured with respect to task completion time and effectiveness (quality of the perception or retrieval). The following subsections explain the details of the experiments.

5.1 Hypotheses

The study tested the following four hypotheses:

- $H1$: Efficient perception hypothesis: The relevance map allows faster perception of the result set.
- $H2$: Efficient retrieval hypothesis: The relevance map allows faster retrieval of relevant information.
- $H3$: Effective perception hypothesis: The relevance map allows more accurate perception of the result set.
- $H4$: Effective retrieval hypothesis: The relevance map allows retrieval of more highly relevant information.

5.2 Experimental Design

The experiment used a 2×2 within-subjects design with two search tasks and two systems. The conditions were counterbalanced by varying the order of the systems and tasks.

5.3 Baseline

A baseline system, shown in Figure 3a, was implemented to enable comparability and as to ensure that the evaluation revealed the effects solely on the features enabling relevance mapping and re-ranking. The baseline used the same data collection as well as the same document ranking model. All retrieved information in the baseline system was displayed with a ranked list layout. The baseline did not feature a relevance map, and the ranking was based on a single query at a time. The baseline was using the same hardware, i.e. a multi-touch-enabled desktop computer with a physical keyboard.

5.4 Tasks

The experiment consisted of two tasks, perception and retrieval, which are explained below and exemplified in Figures 3b,3c_1, and 3c_2. Both tasks used a common set of four topics, either (1) interaction, tabletop, tangible, and prototyping, or (2) surfaces, exploration, visualization, and sound. The two set of topics were formed by two researchers who were experts on human-computer interaction. The same researchers were then asked to assess the task outcomes of the participants.

5.4.1 Perception Task

The perception task aimed to measure task completion time and effectiveness, to help understand how a document space is populated and organized with respect to specific query topics. Participants were asked the two following questions: (1) "Out of the 4 topics provided, which 2 topics are related to the highest amount of

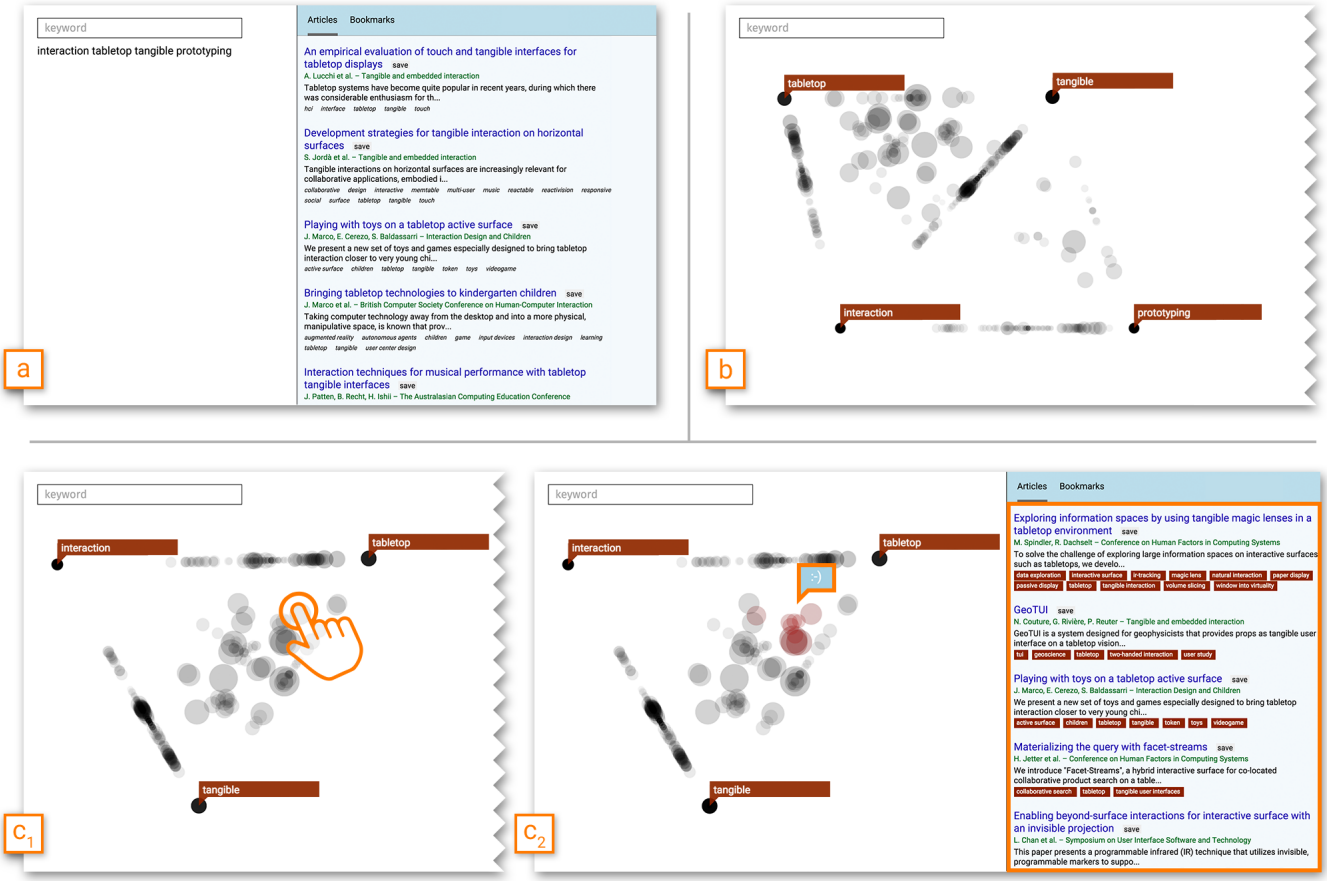

Figure 3: In the perception task, participants must identify the two and three keywords out of four that are the most related to relevant information. In the baseline (a), they must skim through the ranked list of results to infer the most prevalent keywords from the top articles. Using the relevance map, they must interpret the distribution of document markers. In the retrieval task, participants must find an article that shows a high relevance to one keyword (say, *tabletop*), and a lesser relevance with two other keywords (say, *tangible* and *interaction*). Using the baseline (a), they must query the three keywords, then find a fitting article in the result list. Using the relevance map, they point (by tapping on the touch-enabled monitor) at an area between the three keywords (c1), somewhere closer to *tabletop* than *tangible* or *interaction*, which triggers a re-ranking of retrieved articles based on the selected position (c2). The participant should be able to select one of the top articles as a fitting task outcome.

relevant documents?", and (2) "Out of the 4 topics provided, which 3 topics are related to the highest number of relevant documents?".

An example visualization from which the user had to select the topics is shown in Figure 3*b*. In that case, we can see that the space delimited by *tabletop*, *tangible*, and *interaction* is the most densely populated through sheer amount of document markers, making them part of the answer. To find the two keywords, they must then compare pair-wise document density by focusing on the edges between query phrase markers, with a slight but noticeable lead in density (encoded as darkness) between *tangible* and *interaction*.

5.4.2 *Retrieval Task*

The retrieval task aimed to measure task completion time and effectiveness in finding documents with varying multi-dimensional relevance toward several topics. Participants were given the following instruction: "Find one article that is highly relevant to 'Topic A' and slightly related to 'Topic B' and 'Topic C'.". The task was then repeated one more time with a different topic priority: "Find one paper that is highly relevant to 'Topic B' and slightly related to 'Topic A' and 'Topic C'."

An example sequence of a visualization, user pointing to the visualization to re-rank the document list from which the user had to select the documents is shown in Figures $3c_1$ and $3c_2$.

5.5 Measures

We used two performance measures: task completion time and effectiveness. *Task Completion Time* measured the time required to complete the task. *Effectiveness* measured the quality of the task outcome.

5.5.1 *Task Completion Time*

Task completion time was computed directly as the duration in seconds from the beginning of the task to the completion of the task.

5.5.2 *Effectiveness*

Effectiveness was computed differently for the two tasks and the corresponding ground truths for the task outcomes were defined differently.

In the perception task, effectiveness was measured as the accuracy of the participants answer. The ground truth was available

from the relevance estimation and was computed as a sum of the relevance scores associated to each query phrase representing the topic. The topics were then ordered based on the sum of relevance scores and the top 2 and top 3 topics corresponding to the task description were selected as the ground truth to which each answer was then compared. Accuracy was computed for each answer, resulting in a grade of 1 for a match, 0 for a mismatch, and – in the case two topics selected out of four – 0.5 for a partial match. Each participant having returned two answers, effectiveness was then measured as the mean of both grades.

In the retrieval task, effectiveness was measured as precision on the documents selected by the participants. All documents chosen by any of the participants in any of the two system conditions were pooled. Two experts then assessed the actual relevance of each document to each topic. The experts being authors of the experiment design and having themselves devised the topics, potential bias in the assessment was addressed by following a strict double-blind procedure (i.e. experts had no knowledge of the participant, the system or concurrent assessment) and balancing the use of each set of topics across both conditions. The experts assigned for each document a grade between 0 (non-relevant) and 5 (highly relevant) to each of the topics, which were then averaged (mean) into a final grade. The topic defined as highly relevant was given a double coefficient so that the final grade reflected the weighted aspect of the task. The final grade indicated the expert opinion on how relevant the document was for the task. The inter-annotator agreement between the experts was measured by using Cohen's Kappa for two raters who provided three relevance assessments per document. Agreement was found to be substantial (Kappa = 0.684, $Z = 7.04$, $p < 0.001$), indicating that the expert assessments were consistent.

Additionally, we collected the position in the result list of each document returned by each participant, to better understand the re-ranking/scrolling tradeoff.

5.6 Data logging and data collection

For the purpose of the task completion time measurement, we recorded (1) the task duration from the start button press to the end button press. For the purpose of the effectiveness measurement, we recorded (2) bookmarked documents. After completion of both tasks in both conditions, participants were given a questionnaire to collect data on their age, gender, academic background and research experience.

We used a document set including all articles available at the Digital Library of the Association of Computing Machinery (ACM) as of the end of 2011. The information about each document consists of its title, abstract, author names, publication year, and publication venue. Articles with missing information in the metadata were excluded during the indexing phase, resulting in a database with over 320,000 documents. Both the baseline and the proposed system used the same document set and the users were presented with the top 2000 documents.

5.7 Participants

Twenty researchers in computer science (40% females) from two universities, ranging in age from 21 to 36 years old and from 1 to 8 years in research experience, volunteered to participate in the experiment. The participants were all compensated with a movie voucher that they received at the end of the experiment. All participants were assigned the same experimental tasks on both systems with systematic varying order between the systems. In this experiment, informed consent was obtained from all participants.

5.8 Apparatus

Participants performed the experiment on a desktop computer with a 27" multi-touch-enabled capacitive monitor (Dell XPS27). The computer was running Microsoft Windows 8 and both systems – being Web based – were used on a Chrome Web browser version 45.0.2454.85 m. A physical keyboard was provided for text input, whereas pointing, dragging and scrolling were performed through touch interaction. The search engine implementing the relevance estimation method was running on a virtual server and the document index was implemented as an in-memory inverted index allowing very fast response times with an average latency of less than one second.

5.9 Procedure

The tasks were described on individual instruction sheets that incorporated one of the two sets of keywords, to which we will refer as the task versions. The duration of the tasks was not constrained. To avoid introduction of confounding variables, we counterbalanced the tasks by systematically changing the order of the systems, the order of the task versions, and which task version was allocated to each system.

Considering the novelty aspect of the visualization, a training version of the tasks was devised, allowing participants to use both system with comparable proficiency. Training tasks had to be done using each system, right before the main task, using a separate set of four keywords: *creativity*, *collaboration*, *children* and *robotics*. The training started with the participant receiving a tutorial on how to use the system, then, while performing the training task, she could ask questions about either the task or the system. As soon as the training task was completed and the participant had no more questions, the participants started the actual experiment.

Participants were asked to underline the chosen answers on the instruction sheet. In the retrieval task, we asked the participants to bookmark the chosen articles. A Start/Submit button was added to both systems in the upper right corner. To be able to use each system, participants had to tap Start when ready to perform each task and Submit when they had completed it.

6. RESULTS

The results of the experiment regarding performance are shown in Table 1 and illustrated in Figure 4 with respect to the selected measures: task completion time and effectiveness, and reported according to both tasks, perception and retrieval. The mean position of selected articles in the result list is also illustrated in Figure 4. The results are discussed in detail in the following sections.

6.1 Task Completion Time

Significant differences were found between the systems in both tasks, which are discussed as follows.

6.1.1 Perception Task

The results of the perception task show that participants spent substantially less time completing the perception task when using the relevance map than when using the baseline system. The mean task duration for the relevance map was 84.23 seconds, while the mean task duration for the baseline system was 177.72 seconds. The differences between the systems were found statistically significant (Wilcoxon pair-matching ranked-sign test: $Z = 3.27$; $p < 0.001$). In conclusion, the relevance map shows 111% improvement, and was therefore more efficient for the perception task, confirming $H1$.

	Task Completion Time					Effectiveness				
	Baseline (B)		Map (M)		B vs. M	Baseline (B)		Map (M)		B vs. M
	M	SD	M	SD	Wilcoxon Test	M	SD	M	SD	Wilcoxon Test
Perception	177.72	116.20	**84.23**	39.38	**p < 0.001**	0.75	0.23	**0.89**	0.17	**p = 0.013**
Retrieval	137.53	101.66	**80.93**	70.65	**p < 0.001**	0.70	0.13	0.71	0.12	p = 0.95

Table 1: Task completion time and effectiveness results for both tasks. Task completion time is reported as a duration of the task averaged over participants. Effectiveness in the perception task is reported by mean quality of topics averaged over participants, and effectiveness in the retrieval task by mean quality of documents averaged over participants. Results showing significant improvement over the baseline are shown in bold.

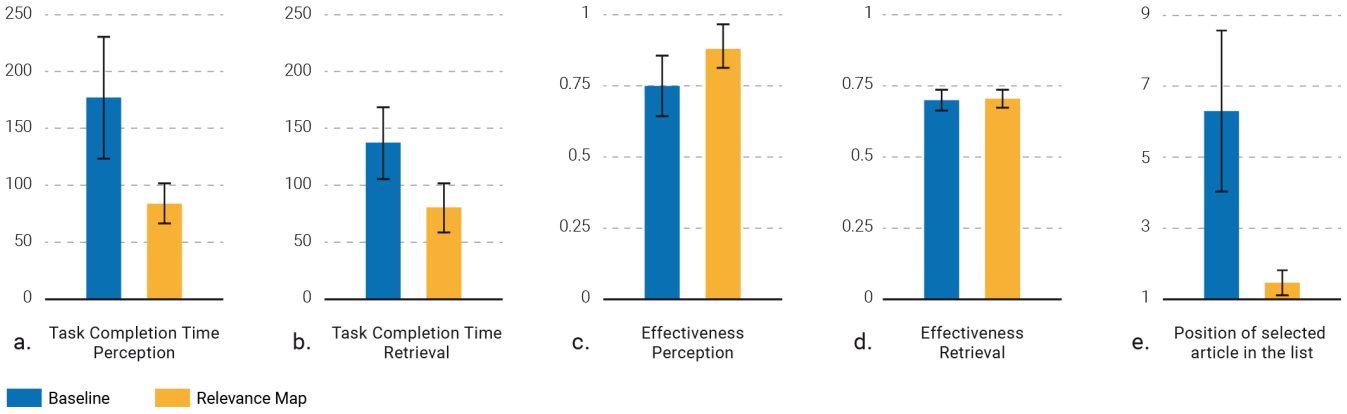

Figure 4: Results from the performance measures displayed for both systems with confidence intervals for: (a) task completion time in the perception task and (b) task completion time in the retrieval task with the mean duration (lower is better), (c) effectiveness in the perception task with the mean topic quality, and (d) effectiveness in the retrieval task with the mean document quality (higher is better). (e) Mean position in the result list of selected articles in the retrieval task.

6.1.2 *Retrieval Task*

In the retrieval task, participants spent substantially less time completing the task when using the relevance map than when using the baseline system. The mean task duration for the relevance map was 80.93 seconds, while the mean task duration for the baseline system was 137.53 seconds. The differences between the systems were found to be statistically significant (Wilcoxon pair-matching ranked-sign test: $Z = 3.87$; $p < 0.001$). In conclusion, relevance map shows 70% improvement and was therefore more efficient for the retrieval task, confirming $H2$.

Using the relevance map, participants selected articles close to the top in the result list, with a mean position of 1.48 (SD = 1.20), while the mean position of the selected article for the baseline system was 6.33 (SD = 7.20). The differences between the systems were found statistically significant (Wilcoxon pair-matching ranked-sign test: $Z = 4.77$; $p < 0.001$).

6.2 Effectiveness

6.2.1 *Perception Task*

In the perception task, the effectiveness as measured by the accuracy of the topics selected by the participants on the relevance map is 0.89, while accuracy on the baseline system is 0.75. The differences between the systems were found to be statistically significant (Wilcoxon pair-matching ranked-sign test: $Z = -2.46$; $p = 0.013$). In conclusion, relevance map was more effective for the perception task, confirming $H3$.

6.2.2 *Retrieval Task*

No statistically significant difference in the relevance of retrieved documents was found in the retrieval task (Wilcoxon pair-matching ranked-sign test: $Z = -0.07$ and $p = 0.95$). The fourth chart in figure 4 shows very similar results for both systems. This result fails to confirm $H4$, but it shows that the improvement in task completion time observed in the retrieval task did not impair the quality of the retrieved documents.

7. DISCUSSION

The results of the experiments show significant improvements in task completion time in both perception and retrieval, without compromising effectiveness. These results confirm hypotheses $H1$, $H2$ and $H3$.

In the perception task, participants were able to use the relevance map visualization to make decisions with greater accuracy, 111% faster. The visualization allowed the participants to understand more accurately the distribution of information with respects to the multiple aspects of the query.

In the retrieval task, documents fitting complex criteria were retrieved 70% faster using re-ranking through interaction with the relevance map. While finding documents with different relevance to several topics requires users to go through long lists of results and assess the relevance of individual documents, our proposed method for re-ranking through pointing at the map successfully narrows down the top results to documents that fit the criteria.

The quality of the task outcome was the same in both conditions in the retrieval task, which failed to confirm hypothesis $H4$. A possible reason for equal performance is the absence of strict time

constraints for participants to complete the tasks. It is possible that a constrained time to complete the task would have negatively impacted the quality of the task outcome for the baseline, as the participants would not have been able to carefully examine the list to find a fitting article, but would have been forced to skim, resulting in possibly lower quality of selected topics and articles.

While our results show substantial improvements over the baseline, there is a tradeoff between the perceived simplicity of a result list and the added visual complexity of a relevance map. Interaction-wise, a result list is explored by scrolling, while a relevance map requires more complex behavior, justifying the use of a training session and tutorial. In the context of the present experiment, the necessity for a tutorial introduces a risk of influencing participants towards optimal behaviors that may outperform self-devised strategies. While our results suggest that the design of such visual interfaces can make both retrieval and perception faster, simpler interfaces may be more effective when the cost of interactions is higher, e.g. smaller devices and mobile scenarios.

We see further research directions to be addressed. First, different task complexity could be investigated and open-ended tasks explored, in which users would have more control over the search process. Second, more realistic search situations outside of our present laboratory experiment could be exploited to investigate interaction with relevance mapping and re-ranking functionality in situations in which users would have the possibility to try their own areas of interest and determine whether the suggestion effectively met their preferences and expectations.

8. CONCLUSION

Conventional systems for information retrieval are not designed to provide important insights of the data, such as relevance distribution of the results with respect to the user's query phrases. In this paper, we introduced visual re-ranking, an interactive visualization technique for multi-aspect information retrieval that helps overcome such limitations. The method proved successful in substantially improving performance over complex analytical tasks. Evaluation showed that users are able to make sense of the relevance map and take advantage of the re-ranking interaction to lower the time required to make analytic decisions or retrieve documents based on complex criteria. These results suggest that the conventional one-dimensional ranked list of results may not be enough for complex search-related tasks that go beyond simple fact finding.

9. ACKNOWLEDGMENTS

This research was partially funded by the European Commission through the FP7 Project MindSee 611570 and the Academy of Finland (Multivire, 255725, 278090 and 305739). The data used in the experiments is derived from the ACM Digital Library.

10. REFERENCES

[1] C. Ahlberg and B. Shneiderman. Visual information seeking: Tight coupling of dynamic query filters with starfield displays. In *Proc. CHI'94*, pages 313–317. ACM, 1994.

[2] C. Ahlberg, C. Williamson, and B. Shneiderman. Dynamic queries for information exploration: An implementation and evaluation. In *Proc. CHI'92*, pages 619–626. ACM, 1992.

[3] J.-W. Ahn and P. Brusilovsky. Adaptive visualization of search results: Bringing user models to visual analytics. *Information Visualization*, 8(3):167–179, 2009.

[4] B. Alsallakh, L. Micallef, W. Aigner, H. Hauser, S. Miksch, and P. Rodgers. Visualizing sets and set-typed data: State-of-the-art and future challenges. In *EuroVis– State of The Art Reports*, pages 1–21. Eurographics, 2014.

[5] B. Amento, W. Hill, L. Terveen, D. Hix, and P. Ju. An empirical evaluation of user interfaces for topic management of web sites. In *Proc. CHI'99*, pages 552–559. ACM, 1999.

[6] S. Andolina, K. Klouche, J. Peltonen, M. Hoque, T. Ruotsalo, D. Cabral, A. Klami, D. Głowacka, P. Floréen, and G. Jacucci. Intentstreams: smart parallel search streams for branching exploratory search. In *Proc. IUI'15*, pages 300–305. ACM, 2015.

[7] S. Benford, D. Snowdon, C. Greenhalgh, R. Ingram, I. Knox, and C. Brown. Vr-vibe: A virtual environment for co-operative information retrieval. In *Computer Graphics Forum*, volume 14, pages 349–360. Wiley, 1995.

[8] P. P. Bonissone, R. Subbu, and J. Lizzi. Multicriteria decision making (MCDM): a framework for research and applications. *Computational Intelligence Magazine*, 4(3):48–61, 2009.

[9] K. W. Boyack, B. N. Wylie, and G. S. Davidson. Domain visualization using VxInsight® for science and technology management. *JASIST*, 53(9):764–774, 2002.

[10] M. Chalmers and P. Chitson. Bead: Explorations in information visualization. In *Proc. SIGIR'92*, pages 330–337. ACM, 1992.

[11] F. Das-Neves, E. A. Fox, and X. Yu. Connecting topics in document collections with stepping stones and pathways. In *Proc. CIKM'05*, pages 91–98. ACM, 2005.

[12] B. Fortuna, M. Grobelnik, and D. Mladenic. Visualization of text document corpus. *Informatica*, 29(4):497–502, 2005.

[13] S. Gratzl, A. Lex, N. Gehlenborg, H. Pfister, and M. Streit. Lineup: Visual analysis of multi-attribute rankings. *TVCG*, 19(12):2277–2286, Dec 2013.

[14] S. Havre, E. Hetzler, K. Perrine, E. Jurrus, and N. Miller. Interactive visualization of multiple query results. In *Information Visualization*, page 105. IEEE, 2001.

[15] M. A. Hearst. Tilebars: Visualization of term distribution information in full text information access. In *Proc. CHI'95*, pages 59–66. ACM, 1995.

[16] E. Hetzler and A. Turner. Analysis experiences using information visualization. *IEEE CG&A*, 24(5):22–26, 2004.

[17] O. Hoeber and X. D. Yang. Interactive web information retrieval using wordbars. In *International Conference on Web Intelligence*, pages 875–882. IEEE, 2006.

[18] O. Hoeber and X. D. Yang. The visual exploration of web search results using hotmap. In *Proc. IV'06*, pages 157–165. IEEE Computer Society, 2006.

[19] C. Kang, X. Wang, Y. Chang, and B. Tseng. Learning to rank with multi-aspect relevance for vertical search. In *Proc. WSDM'12*, pages 453–462. ACM, 2012.

[20] S. Kaski, T. Honkela, K. Lagus, and T. Kohonen. WEBSOM–self-organizing maps of document collections. *Neurocomputing*, 21(1):101–117, 1998.

[21] A. Kerne, E. Koh, B. Dworaczyk, J. M. Mistrot, H. Choi, S. M. Smith, R. Graeber, D. Caruso, A. Webb, R. Hill, and J. Albea. combinformation: A mixed-initiative system for representing collections as compositions of image and text surrogates. In *Proc. JCDL'06*, pages 11–20. ACM, 2006.

[22] P. Kidwell, G. Lebanon, and W. S. Cleveland. Visualizing incomplete and partially ranked data. volume 14, pages 1356–1363. IEEE, 2008.

[23] K. Klouche, T. Ruotsalo, D. Cabral, S. Andolina, A. Bellucci, and G. Jacucci. Designing for exploratory search on touch devices. In *Proc. CHI'15*, pages 4189–4198. ACM, 2015.

[24] K. Kucher and A. Kerren. Text visualization techniques: Taxonomy, visual survey, and community insights. In *PacificVis'15*, pages 117–121. IEEE Computer Society, 2015.

[25] B. Kules and B. Shneiderman. Users can change their web search tactics: Design guidelines for categorized overviews. *IPM*, 44(2):463–484, 2008.

[26] X. Lin. Map displays for information retrieval. *JASIS*, 48(1):40–54, 1997.

[27] J. Mackinlay. Automating the design of graphical presentations of relational information. *TOG*, 5(2):110–141, 1986.

[28] J. Matejka, F. Anderson, and G. Fitzmaurice. Dynamic opacity optimization for scatter plots. In *Proc. CHI'15*, pages 2707–2710. ACM, 2015.

[29] N. E. Miller, P. C. Wong, M. Brewster, and H. Foote. Topic islands tm-a wavelet-based text visualization system. In *Proc. VIS*, pages 189–196. IEEE, 1998.

[30] E. Morse and M. Lewis. Why information retrieval visualizations sometimes fail. In *Proc. IEEE SMC'97*, volume 2, pages 1680–1685, Oct 1997.

[31] T. Munzner, F. Guimbretière, S. Tasiran, L. Zhang, and Y. Zhou. TreeJuxtaposer: scalable tree comparison using focus + context with guaranteed visibility. *TOG*, 22(3):453–462, 2003.

[32] A. Nuchprayoon and R. R. Korfhage. Guido, a visual tool for retrieving documents. In *Proc. VL/HCC'94*, pages 64–71. IEEE, 1994.

[33] K. A. Olsen, R. R. Korfhage, K. M. Sochats, M. B. Spring, and J. G. Williams. Visualization of a document collection: The VIBE system. *IPM*, 29(1):69–81, 1993.

[34] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Proc. SIGIR'98*, pages 275–281. ACM, 1998.

[35] T. Porter and T. Duff. Compositing digital images. In *Proc. SIGGRAPH'84*, pages 253–259. ACM, 1984.

[36] V. V. Raghavan and S. M. Wong. A critical analysis of vector space model for information retrieval. *JASIS*, 37(5):279–287, 1986.

[37] S. E. Robertson. Readings in information retrieval. chapter The Probability Ranking Principle in IR, pages 281–286. Morgan Kaufmann Publishers Inc., 1997.

[38] R. M. Rohrer, D. S. Ebert, and J. L. Sibert. The shape of Shakespeare: visualizing text using implicit surfaces. In *Proc. InfoVis*, pages 121–129. IEEE, 1998.

[39] T. Ruotsalo, G. Jacucci, P. Myllymäki, and S. Kaski. Interactive intent modeling: Information discovery beyond search. *Communications of the ACM*, 58(1):86–92, 2015.

[40] T. Ruotsalo, J. Peltonen, M. Eugster, D. Głowacka, K. Konyushkova, K. Athukorala, I. Kosunen, A. Reijonen, P. Myllymäki, G. Jacucci, et al. Directing exploratory search with interactive intent modeling. In *Proc. CIKM'13*, pages 1759–1764. ACM, 2013.

[41] J. Seo and B. Shneiderman. A rank-by-feature framework for interactive exploration of multidimensional data. *Information Visualization*, 4(2):96–113, 2005.

[42] C. Shi, W. Cui, S. Liu, P. Xu, W. Chen, and H. Qu. RankExplorer: Visualization of ranking changes in large time series data. *TVCG*, 18(12):2669–2678, 2012.

[43] B. Shneiderman. Dynamic queries for visual information seeking. *Software*, 11(6):70–77, Nov 1994.

[44] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Proc. VL/HCC'96*, pages 336–343. IEEE, 1996.

[45] B. Shneiderman, D. Feldman, A. Rose, and X. F. Grau. Visualizing digital library search results with categorical and hierarchical axes. In *Proc. DL'00*, pages 57–66. ACM, 2000.

[46] B. Shneiderman, C. Plaisant, M. Cohen, and S. Jacobs. *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. Addison-Wesley Publishing Company, 5th edition, 2009.

[47] W. Song, Q. Yu, Z. Xu, T. Liu, S. Li, and J.-R. Wen. Multi-aspect query summarization by composite query. In *Proc. SIGIR'12*, pages 325–334. ACM, 2012.

[48] A. Spoerri. InfoCrystal: A visual tool for information retrieval & management. In *Proc. CIKM'93*, pages 11–20. ACM, 1993.

[49] J. Wise, J. J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, V. Crow, et al. Visualizing the non-visual: spatial analysis and interaction with information from text documents. In *Proc. Information Visualization*, pages 51–58. IEEE, 1995.

[50] D. Xin, J. Han, H. Cheng, and X. Li. Answering top-k queries with multi-dimensional selections: The ranking cube approach. In *Proc. VLDB'06*, pages 463–474. VLDB, 2006.

[51] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *TOIS*, 22(2):179–214, 2004.

[52] J. Zhang. Tofir: A tool of facilitating information retrieval–introduce a visual retrieval model. *IPM*, 37(4):639–657, 2001.

[53] J. Zhang and R. R. Korfhage. Dare: Distance and angle retrieval environment: A tale of the two measures. *JASIS*, 50(9):779–787, 1999.