Comparative Study of Search Engine Result Visualisation: Ranked Lists Versus Graphs

Casper Petersen
Dept. of Computer Science
University of Copenhagen
cazz@diku.dk

Christina Lioma
Dept. of Computer Science
University of Copenhagen
c.lioma@diku.dk

Jakob Grue Simonsen
Dept. of Computer Science
University of Copenhagen
simonsen@diku.dk

ABSTRACT

Typically search engine results (SERs) are presented in a *ranked list* of decreasing estimated relevance to the user query. While familiar to users, ranked lists do not show any inherent connections between SERs, e.g. whether SERs are hyperlinked or authored by the same source. Such potentially useful connections between SERs can be displayed as *graphs*. We present a preliminary comparative study of ranked lists versus graph visualisations of SERs. Experiments with TREC data from the domain of web search and a small user study of 10 participants show that ranked lists result in more precise and also faster search sessions than graph visualisations.

Categories and Subject Descriptors

H.3.3 [**INFORMATION STORAGE AND RETRIEVAL**]: Information Search and Retrieval

Keywords

Search Engine Result Visualization, Ranked List, Graph

1 Introduction and Motivation

Typically search engine results (SERs) are presented in a ranked list of decreasing estimated relevance to the user query. Drawbacks of ranked lists include showing only a limited view of the information space, not showing how similar the retrieved documents are and/or how the retrieved documents relate to each other [4, 6]. Such potentially useful information could be displayed to users in the form of *SER graphs*; these could present at a glance an overview of any clusters or isolated documents among the SERs, features not typically integrated into ranked lists. For instance, directed/undirected and weighted/unweighted graphs could be used to display the direction, causality and strength of various relations among SERs. In addition, various graph properties (see [7]), such as the average path length, clustering coefficient or degree, could be also displayed to the user, reflecting potentially useful or interesting features about how the retrieved data is connected. Our motivation comes from the observation, that while traditional IR systems successfully support known-item search [5], what should users do if they want to locate something from a domain where they have a general interest but no specific knowledge? [8]. Such exploratory searching is not supported by contemporary IR systems [5], prompting users to "develop coping strategies which involves [...] the interactive exploration of the retrieved document space, selectively following links and passively obtaining cues about where their next steps lie." [9]. A step towards exploratory search is thus to make explicit the hyper-linked structure of the ordered list used by e.g. Google and Yahoo. Investigation of such a representation does not exist according to our knowledge, but is comparable to Google's Knowledge Graph whose aim is to guide users to other relevant information from an initial selection.

In this paper, we present a user study comparing ranked list versus graph-based SER visualisation interfaces. We use a web crawl of approximately 50 million documents in English with associated hyperlink information and 10 participants. We find that ranked lists result in overall more accurate and faster searches than graph displays, but that the latter result in slightly higher recall. We also find overall higher inter-rater agreement about SER relevance when using ranked lists instead of graphs.

2 Previous Work

Earlier work on graph-based SER displays includes Beale et al.'s (1997) visualisation of sequences of queries and their respective SERs, as well as the work of Shneiderman & Aris (2006) on modelling semantic search aspects as networks (both overviewed in [10]). Treharne et al. (2009) present a critique of ranked list displays side by side a range of other types of visualisation, including not only graphs, but also cartesian, categorical, spring and set-based displays [6]. This comparison is analytical rather than empirical. Closest to ours is the work of Donaldson et al. (2008), who experimentally compare ranked lists to graph-based displays [2]. In their work, graphs model social web information, such as user tags and ratings, in order to facilitate contextualising social media for exploratory web search. They find that users seem to prefer a hybrid interface that combines ranked lists with graph displays.

3 Interface Design

This section presents the two different SER visualisations used in our study. Our goal is to study the effect of display-

SIGIR 2013 Dublin, Ireland

ing exactly the same information to the user in two different ways, using *ranked list* and *graph* visualisations, respectively.

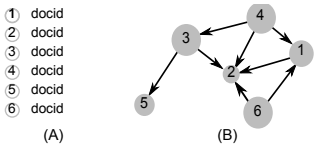

Figure 1: Ranked list (A) and graph (B) representation of the top-k documents from a query.

3.1 Ranked List (RL) Display

We use a standard ranked list SER display, where documents are presented in decreasing order of their estimated relevance to the user query. The list initially displays only the top-k retrieved document ids (docids) with their associated rank (see Figure 1 (A)). When clicked upon, each document expands to two mini windows, overlaid to the left and right of the list:

- The left window shows a document snippet containing the query terms. The snippet provides a brief summary of the document contents that relate to the query in order to aid the user to assess document relevance prior to viewing the whole document [4]. We describe exactly what the snippet shows and how it is extracted in Section 3.3.
- The right window shows a graph of the top-k ranked SERs (see Section 3.2). The position of the clicked document in the graph is clearly indicated, so users can quickly overview its connections, if any, to other top-k retrieved documents.

Previously visited documents in the list are colour-marked.

3.2 Graph (GR) Display

We display a SER graph $G = (V, E)$ as a directed graph whose vertices $v \in V$ correspond to the top-k retrieved documents, and edges $e \in E$ correspond to links (hyperlinks in our case of web documents) between two vertices. Each vertex is shown as a shaded circle that displays the rank of its associated document in the middle, see Figure 1 (B). The size of each vertex is scaled according to its out-degree, so that larger vertex size indicates more outlinks to the other top-k documents. Edge direction points towards the outlinked document. Previously visited documents are colour-marked.

When clicked upon, each vertex expands to two mini windows, overlaid to the left and right of the graph:

- The left window shows the same document snippet as in the RL display.
- The right window shows the ranked list of the top-k SERs. The position of the clicked document in the list is clearly marked.

We display the SER graph in a standard force-directed layout [1]. Our graph layout does not allow for other types of interaction with the graph apart from clicking on it. We reason that for the simple web search tasks we consider, layouts allowing further interaction may be confusing or time-consuming, and that they may be more suited to other search tasks, involving for instance decision making, navigation and exploration of large information spaces.

3.3 Document Snippets

Both the RL and GR interfaces include short query-based summaries of the top-k SERs (*snippets*). We construct them as follows: We extract from each document a window of $\pm$ 25 terms surrounding the query terms on either side. Let a query consist of 3 terms q_1, q_2, q_3. We extract snippets for all ordered but not necessarily contiguous sequences of query terms: $(q1, q2, q3)$, (q_1, q_2), (q_1, q_3), (q_2, q_3), (q_1), (q_2), (q_3). This way, we match all snippets containing query terms in the order they appear in the query (not as a bag of words), but we also allow other terms to occur in between query terms, for instance common modifiers.

Several snippets can be extracted per document, but only the snippet with the highest TF-IDF score is displayed to the user. The TF-IDF of each window is calculated as a normalised sum of the TF-IDF weights for each term:

$$S_{s(D)} = \frac{1}{|w|} \sum_{t=0}^{|w|} tf(t, D) \times \log \left(\frac{|C|}{|D \in C \ : \ t \in D|} \right)$$

where $|w|$ is the number of terms in the window extracted, $t \in w$ is a term in the window, tf is the term frequency of t in document D from which the snippet is extracted, C is the collection of documents, and $S_{s(D)}$ is the snippet score for document D. Finally, as research has shown that query term highlighting can be a useful feature for search interfaces [4], we highlight all occurrences of query terms in the snippet.

4 Evaluation

We recruited 2 participants for a pilot study to calibrate the user interfaces; the results from the pilot study were not subsequently used. For the main study, we recruited 10 new participants (9 males, 1 female; average age: 33.05, all with a background in Computer Science) using convenience sampling. Each participant was given a short introduction and demonstration of the two interfaces. Their task was to find and mark as many relevant documents as possible per query using either interface. For each new query, the SERs could be shown in either interface. Each experiment lasted 30 minutes.

Participants did not submit their own queries. The queries were taken from the TREC Web tracks of 2009-2012 (200 queries in total). This choice allowed us to provide very fast response times to participants ($<$ two seconds, depending on disk speed), because search results and their associated graphs were pre-computed and cached. Alternatively, running new queries and plotting their SER graphs on the fly would result in notably slower response times that would risk dissatisfying participants. However, a drawback in using TREC queries is that participants did not necessarily have enough context to fully understand the underlying information needs and correctly assess document relevance. To counter this, we allowed participants to skip queries they were not comfortable with. To avoid bias, skipping a query was allowed after query terms were displayed, but before the SERs were displayed.

We retrieved documents from the TREC ClueWeb09 cat. B dataset (ca. 50 million documents crawled from the web in 2009), using Indri, version 5.2. The experiments were carried out on a 14 inch monitor with a resolution of 1400 x 1050 pixels. We logged which SERs participants marked relevant, as well as the participants' click order and time spent per SER.

	Ranked List	Graph
MAP@20	0.4195	0.3211
MRR	0.4698	0.3948
RECALL@20	0.0067	0.0069

Table 1: Retrieval performance per visualisation: Mean Average Precision (MAP), Mean Reciprocal Rank (MRR) & Recall of the top 20 results.

4.1 Findings

In total the 10 participants processed 162 queries (89 queries with the RL interface and 73 with the GR interface) with mean $\mu = 16.2$, and standard deviation $\sigma = 7.8$. Four queries (two from each interface) were bypassed (2.5% of all processed queries).

Table 1 shows retrieval effectiveness per interface, aggregated over all queries for the top $k = 20$ SERs. The ranked list is associated with higher, hence better scores than the graph display for MAP and MRR. MAP is +30.6% better with ranked lists that with graph displays, meaning that overall a higher amount of relevant SERs is found by the participants at higher ranks in the ranked list as opposed to the graph display. This finding is in agreement with the MRR scores, which indicate that the first SER to be assessed relevant is likely to occur around rank position 2.13 ($1/2.13 = 0.469 \approx 0.4698$) with ranked lists, but around rank position 2.55 ($1/2.55 = 0.392 \approx 0.3948$) with graph displays. Conversely, recall is slightly higher with graph displays. In general, higher recall in this case would indicate that participants are more likely to find a slightly larger amount of relevant documents when seeing them as a graph of their hyperlinks. However, the difference in recall between ranked lists and graphs is very small and can hardly be seen as a reliable indication.

4.1.1 Click-order

On average participants clicked on 9.46 entries per query in the ranked list (842 clicks for 89 queries) but only on 6.7 entries per query in the graph display (490 clicks for 73 queries). The lower number of clicks in the latter case could be due to the extra time it might have taken participants to understand or navigate the graph. This lower number of clicks also agrees with the lower MAP scores presented above (if fewer entries were clicked, fewer SERs were assessed, hence fewer relevant documents were found in the top ranks).

Figures 2a and 2b plot the order of clicks for the ranked list and graph interfaces respectively on the x-axis, against the frequency of clicks on the y-axis. We see that in the ranked list, the first click of the participant is more often on a relevant document, but in the graph display, the first click is more often on a non-relevant document (as already indicated by the MRR scores shown above). We also see that for the graph display, the majority of participant clicks before the 5th click correspond to non-relevant documents. Even though the MRR scores of the graph display indicate that the first relevant document occurs around rank position 2.5, we see that participants on average click four other documents before clicking the relevant document at rank position 2.5. This indicates that in the graph display, participants click documents not necessarily according to their rank position (indicated in the centre of each vertex), but rather according to their graph layout or connectivity.

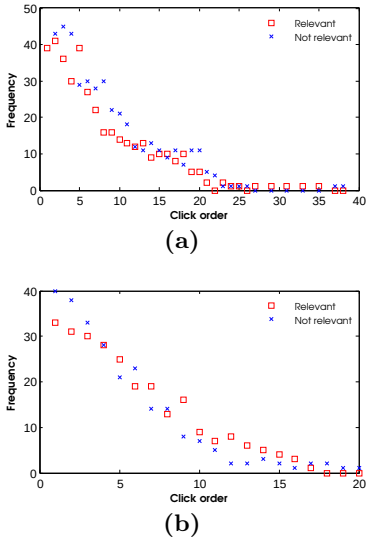

Figure 2: Click-order and participant relevance assessments for the (a) ranked list interface and (b) graph interface

Interface	Min	Max	μ	σ
Ranked List	1.391	25.476	8.228	4.371
Graph	3.322	20.963	9.705	3.699

Table 2: Descriptive statistics from the time spent on each interface in total. All times are in seconds.

4.1.2 Time spent

Table 2 displays statistics about the time participants spent on each interface. We see that overall participants spent less time on the ranked list than on the graph display. This observation, combined with the retrieval effectiveness measures shown in Table 1, indicates that participants conducted overall slightly more precise and faster searches using the ranked lists than using graph displays. The time use also suggests that participants are used to the standard interface of ranked lists; this type of conditioning is not easy to control experimentally.

4.1.3 Inter-participant agreement

To investigate how consistent participants were in their assessments, we report the inter-rater agreement using Krippendorff's α [3]. Table 3 reports the agreement between the participants, and Table 4 reports the agreements between participants and the TREC preannotated relevance assessments per interface. In both cases, only queries annotated more than once by different participants are included (19 queries for the ranked list and 11 for the graph SER).

The average inter-rater agreements between participants vary considerably. For the graph interface, $\alpha = 0.04471$, which suggests lack of agreement between raters. On a query basis, some queries (query 169 and 44) suggest a comparatively much higher agreement whereas others (e.g. query 104 and 184) show a comparatively higher level of disagreement. For the ranked list, inter-rater agreement is higher ($\alpha = 0.19813$). On a per query basis, quite remarkably, query 92 had a perfect agreement between raters, while queries 175 and 129 also exhibited a moderate to high level of agreement. However, most queries show only a low to moderate level of agreement or disagreement.

Overall, the lack of agreement could indicate confusion on behalf of the participants in assessing the relevance of SERs to pre-typed queries. This may be aggreavated by problems in rendering the HTML text of certain snippets into text. Certain HTML documents were ill-formed, meaning that their extracted snippets sometimes included HTML tags or other not always coherent text.

Inter-rater agreements between our participants and the TREC preannotated relevance assessments show an almost complete lack of agreement on average. For both interfaces there is only a weak level of disagreement on average ($\alpha = -0.0750$ and $\alpha = -0.0721$ for the graph and ranked list respectively). On a per query basis there are only two queries (query 169 and 110 for the SER graph and ranked list) exhibiting a moderate level of agreement. For most remaining queries our participants' assessments disagree with the TREC assessments.

Graph			Ranked list		
Query	Raters	α	Query	Raters	α
101	4	0.28696	110	3	0.41000
104	2	-0.21875	119	2	0.00000
132	2	-0.16071	120	2	0.49351
169	2	0.48000	129	2	0.86022
180	2	-0.10031	132	3	-0.08949
184	2	-0.25806	133	2	0.30108
3	2	0.00000	155	2	-0.02632
38	2	-0.07519	175	2	0.49351
44	2	0.49351	180	2	-0.37879
58	2	0.00000	51	2	0.00000
–	–	–	53	2	0.02151
–	–	–	74	2	-0.14706
–	–	–	80	2	0.14420
–	–	–	81	3	-0.12919
–	–	–	92	2	1.00000
–	–	–	95	2	0.15584
–	–	–	96	2	0.15584
–	–	–	97	2	0.30179
Average α: 0.04471			Average α: 0.19813		

Table 3: Krippendorff's α of inter-rater agreement for queries assessed by more than one participant. *Query* refers to the TREC id of each query.

Graph			Ranked List		
Query	Raters	α	Query	Raters	α
101	4	0.09559	110	3	0.38654
104	2	-0.17861	119	2	-0.22370
132	2	0.06561	120	2	0.03146
169	2	0.33625	129	2	0.05600
180	2	-0.08949	132	3	0.01689
184	2	-0.08949	133	2	0.04398
3	2	-0.37209	155	2	-0.21067
38	2	-0.05006	165	2	-0.25532
44	2	-0.05861	175	2	-0.07886
54	2	-0.25532	180	2	-0.17861
58	2	-0.22917	51	2	-0.05006
–	–	–	53	2	-0.24694
–	–	–	74	2	-0.06033
–	–	–	80	2	-0.24694
–	–	–	81	3	-0.13634
–	–	–	92	2	-0.21181
–	–	–	95	2	0.04582
–	–	–	96	2	-0.12919
–	–	–	97	2	0.07813
Average α: -0.0750			Average α: -0.0721		

Table 4: Krippendorff's α of inter-rater agreement between participants and TREC assessments for queries assessed by more than one participant. *Query* refers to the TREC id of each query.

5 Conclusions

In a small user study (10 participants), we compared ranked list versus graph-based search engine result (SER) visualisation. Our motivation was to conduct a preliminary experimental comparison of the two for the domain of web search, where document hyperlinks were used to display them as graphs. We found that overall more accurate and faster searches were done using ranked lists and that inter-user agreement was overall higher with ranked lists than with graph displays. Limitations of this study include: (1) using fixed TREC queries, instead of allowing users to submit their own queries on the fly; (2) having technical HTML to text rendering problems, resulting in sometimes incoherent document snippets; (3) using only 10 users exclusively from Computer Science, which makes for an overall small and rather biased user sample; (4) not using the wider context of the search session in the analysis (e.g. user task, behaviour, satisfaction). Future work includes addressing the above limitations and also testing whether and to what extent these results apply when scaling up to wall-sized displays with significantly larger screen real estate.

6 References

[1] G. D. Battista, P. Eades, R. Tamassia, and I. G. Tollis. *Graph drawing: algorithms for the visualization of graphs*. Prentice Hall PTR, 1998.

[2] J. J. Donaldson, M. Conover, B. Markines, H. Roinestad, and F. Menczer. Visualizing social links in exploratory search. In *Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, HT '08, pages 213–218, New York, NY, USA, 2008. ACM.

[3] A. F. Hayes and K. Krippendorff. Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1(1):77–89, 2007.

[4] M. Hearst. *Search user interfaces*. Cambridge University Press, 2009.

[5] G. Marchionini. Exploratory search: from finding to understanding. *Communications of the ACM*, 49(4):41–46, 2006.

[6] K. Treharne and D. M. W. Powers. Search engine result visualisation: Challenges and opportunities. In *Information Visualisation, 2009 13th International Conference*, pages 633–638, 2009.

[7] S. Wasserman and K. Faust. *Social network analysis: methods and applications*. Structural analysis in the social sciences. Cambridge University Press, 1994.

[8] R. W. White, B. Kules, S. M. Drucker, et al. Supporting exploratory search, introduction, special issue, communications of the acm. *Communications of the ACM*, 49(4):36–39, 2006.

[9] R. W. White, G. Muresan, and G. Marchionini. Report on acm sigir 2006 workshop on evaluating exploratory search systems. *SIGIR Forum*, 40(2):52–60, 2006.

[10] M. L. Wilson, B. Kules, B. Shneiderman, et al. From keyword search to exploration: Designing future search interfaces for the web. *Foundations and Trends in Web Science*, 2(1):1–97, 2010.