Using Word Clouds to Summarize Multiple Search Results

Rianne Kaptein[1,2] Jaap Kamps[1]

[1] University of Amsterdam, The Netherlands
[2] Oxyme, Amsterdam, The Netherlands

ABSTRACT

Search engine result pages (SERPs) are known as the most expensive real estate on the planet. Most queries yield millions of organic search results, yet searchers seldom look beyond the first handful of results. To make things worse, different searchers with different query intents may issue the exact same query. An alternative to showing individual web pages summarized by snippets is to represent a whole group of results. In this paper we investigate if we can use word clouds to summarize groups of documents, e.g. to give a preview of the next result page, or of clusters of topically related documents. We experiment with three word cloud generation methods (full-text, query biased and anchor text based clouds) and evaluate them in a user study. Our findings are: First, biasing the cloud towards the query does not lead to test persons better distinguishing relevance and topic of the search results, but test persons prefer them because differences between the clouds are emphasized. Second, anchor text clouds are to be preferred over full-text clouds since they contain fewer noisy words. Third, we obtain moderately positive results on the relation between the selected world clouds and the underlying search results: there is exact correspondence in 70% of the subtopic matching judgments and in 60% of the relevance assessment judgments.

1. INTRODUCTION

In this paper we investigate the use of word clouds to summarize multiple search results. We study how well users can identify the relevancy and the topic of search results by looking only at the word clouds. Search results can contain thousands or millions of potentially relevant documents. In the common search paradigm of today, you go through each search result one by one, using a search result snippet to determine if you want to look at a document or not. We want to explore an opportunity to summarize multiple search results which can save the users time by not having to go over every single search result.

Documents are grouped by two dimensions. First of all, we summarize complete SERPs containing documents returned in response to a query. Our goal is to discover whether a summary of a SERP can be used to determine the relevancy of the search results on that page. Secondly, documents are grouped by subtopic of the search request. Search results are usually documents related to the same topic, that is the topic of the search request. Faceted or ambiguous queries are have multiple distinct interpretations, and most likely a user interested in one interpretation would not be interested in the others.

*This is an extended abstract of Kaptein and Kamps [1].

DIR-2012 February 23–24, 2012, Gent, Belgium.

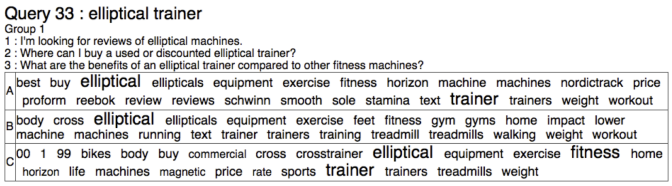

Figure 1: Full-text clouds for the query 'Elliptical Trainer" of the subtopic matching task

The snippets used in modern web search are query biased, and are proven to be better than static document summaries. We want to examine if the same is true for word clouds, hence we want to find out if query biased word clouds are preferred over static word clouds? Besides the text on a web page, web pages can be associated with anchor text, i.e. the text on or around links on web pages linking to a web page. Is anchor text a suitable source of information to generate word clouds?

2. APPROACH

We conduct a user study were we simulated result pages based on known relevance or known subtopics using the TREC 2009 Web Track's data, and try to find out if test persons can recover the relevancy or subtopic. Specifically, we had 21 test persons completing each 10 queries for two tasks based on three types of word clouds.

Subtopic Matching Task When queries are ambiguous or multi faceted, can the word cloud be used to identify the clusters? Test persons have to match the subtopics to the corresponding word clouds. An example topic for this task can be found in Figure 1.

Relevance Assessment Task How well can test persons predict if results are relevant by looking at a word cloud? Test persons have to grade word clouds on a three-point scale (Relevant, Some relevance, Non relevant). An example topic for this task can be found in Figure 2.

Full text Clouds Our standard model uses the full text of documents to generate word clouds using a parsimonious language model that incorporates multiword terms [2], as shown in Figure 1.

Query biased Clouds The surrogate documents used to

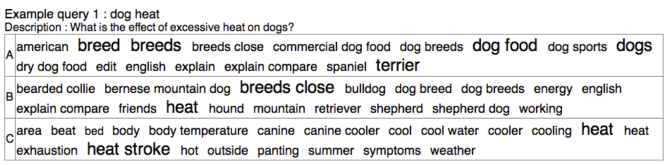

Figure 2: Query biased clouds for the query 'Dog Heat' of the relevance assessment task

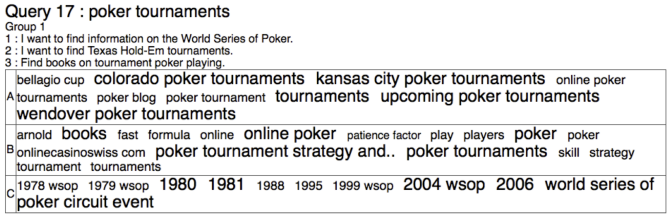

Figure 3: Anchor text clouds for the query 'Poker tournaments' of the subtopic matching task

Table 1: Percentage of correct assignments on the relevance assessments task

Model	Relevant	Half	Non Relevant	All
FT	0.42	0.36	0.44	0.40
QB	0.42^{-}	0.39^{-}	0.50^{-}	0.44^{-}

generate query biased clouds contain only terms that occur around the query words. In our experiments all terms within a proximity of 15 terms to any of the query terms are included, as shown in Figure 2

Anchor Text Clouds For each document, we only keep the most frequently occurring anchor text terms. Maximum likelihood estimation is used to estimate the probability of an anchor text term occurring, dividing the number of occurrences of the anchor text by the total number of anchor text terms in the document set.

3. RESULTS

Query Bias Are query biased word clouds to be preferred over static word clouds?

Our test persons perform the subtopic matching significantly better using the full-text model (significance measured by a 2-tailed sign-test at significance level 0.05). The full-text clouds judgments match the ground truth in 67% of all assignments, the query biased clouds match in 58% of the cases.

The results of this relevance assessment task are in Table 1. On the relevance assessment task the query biased model performs better than the full-text model, but the difference is not significant.

Anchor Text Is anchor text a suitable source of information to generate word clouds?

On the subtopic matching task, the anchor text model performs slightly better than the full-text model on the subtopic task, with an accuracy of 72% versus an accuracy of 68% of the full text model.

Results of the relevance assessment task are in Table 2. The anchor text model performs best, with almost 60% of the assignments correctly made. The clouds with some relevance are the hardest to recognize.

Table 2: Percentage of correct assignments on the relevance assessments task

Model	Relevant	Half	Non Relevant	All
FT	0.61	0.47	0.56	0.54
AN	0.62^{-}	0.50^{-}	0.63^{-}	0.59^{-}

Table 3: Pairwise preferences of test persons over word cloud generation models

		Subtopic		Relevance	
$Model_1$	$Model_2$	$Model_1$	$Model_2$	$Model_1$	$Model_2$
AN	FT	**47**$^{\circ}$	21	**43**$^{\circ}$	23
AN	QB	39	**47**	34	34
FT	QB	29	**41**	23	**43**$^{\circ}$

User Preference For each query, the test persons assess two groups of word clouds without knowing which word cloud generation method was used, and they selected a preference for one of the clouds. The totals of all these pairwise preferences are shown in Table 3. The full-text model performs worst on both tasks. On the subtopic task, the query biased model outperforms the anchor text model, but the difference is not significant.

4. CONCLUSIONS

In this paper we investigated whether word clouds can be used to summarize multiple search results to convey the topic and relevance of these search results, and experimented with using anchor text as an information source and biasing the clouds towards the query. We achieve moderately positive results on the correspondence between the selected word clouds and the underlying pages. Word clouds to assess the relevance of a complete SERP achieve an accuracy of around 60% of the assignments being correct, while subtopics are matched with an accuracy of around 70%. It is clear however that interpreting word clouds is not so easy. This may be due in part to the unfamiliarity of our test persons with this task, but also due to the need to distinguish between small differences in presence of noise and salient words. Especially the word clouds based on varying degrees of relevant information seem remarkably robust. This can also be regarded as a feature: it allows for detecting even a relatively low fraction of relevant results.

Acknowledgment This research was supported by the Netherlands Organization for Scientific Research (NWO projects # 612.066.513, 639.072.601, and 640.005.001) and by the European Community's Seventh Framework Program (FP7 2007/2013, Grant Agreement 270404).

PS In case you are wondering: the correct assignments of the clouds in Figures 1, 2 and 3 respectively are: 1-A, 2-C, 3-B; A-Non Rel., B-Some Rel., C-Rel.; and 1-C, 2-A, 3-B.

REFERENCES

[1] R. Kaptein and J. Kamps. Word clouds of multiple search results. In *IRFC 2011*, volume 6653 of *LNCS*, pages 78–93, 2011. `http://dx.doi.org/10.1007/978-3-642-21353-3`.

[2] R. Kaptein, D. Hiemstra, and J. Kamps. How different are language models and tag clouds? In *ECIR 2010*, volume 5993 of *LNCS*, pages 556–568, 2010.