Blog Snippets: A Comments-Biased Approach

Javier Parapar
Information Retrieval Lab
Dept. of Computer Science
University of A Coruña
javierparapar@udc.es

Jorge López-Castro
Information Retrieval Lab
Dept. of Computer Science
University of A Coruña
irlab@udc.es

Álvaro Barreiro
Information Retrieval Lab
Dept. of Computer Science
University of A Coruña
barreiro@udc.es

ABSTRACT

In the last years Blog Search has been a new exciting task in Information Retrieval. The presence of user generated information with valuable opinions makes this field of huge interest. In this poster we use part of this information, the readers' comments, to improve the quality of post snippets with the objective of enhancing the user access to the relevant posts in a result list. We propose a simple method for snippet generation based on sentence selection, using the comments to guide the selection process. We evaluated our approach with standard TREC methodology in the Blogs06 collection showing significant improvements up to 32% in terms of MAP over the baseline.

Categories and Subject Descriptors: H.3.3 [Information Search and Retrieval]: search process
General Terms: Experimentation
Keywords: Blogs, Comments, Snippets.

1. INTRODUCTION AND BACKGROUND

In the last years the rise of the blogs has induced a new branch within the field of Information Retrieval: blog searching. Indeed, a new TREC Track has been created in order to advance in this area, the Blog Track starting in 2006 [7] with a new standard data-set, the Blogs06 collection [5].

Basically in blog search we can distinguish the post-page, the individual web-page that contains a post; the post by itself, the text entry that has been written by the blog's author; and the comments, opinions of the readers about the post. For instance, in the opinion retrieval task in the TREC Blog Track the retrieval unit is the post-page. Although most of the information should be present in the post, several works have already presented results demonstrating the utility of the readers' comments in the post-pages.

In [6] the authors demonstrated that by indexing the comments and the post, the recall of blog search is increased. It has to be remarked that recall is more important in blog than in web search because searchers are usually interested in all the recent posts about a specific topic. Although the precision values are basically maintained, highly discussed relevant post-pages seem to appear before in the rank. These results were obtained in a small corpus of 225MB where the 15% of the posts are commented. Another way of exploiting the comments information is in opinion finding. In [6] the comments are also used as indicators of discussion, estimating the presence of comments and using this information to formulate query independent document priors.

Post comments have also been used in order to improve blog's summaries as a way of capturing readers feedback to the post entries. In [3] several original approaches are presented exploiting comments' content to summarize a post. In this work the evaluation was done only over 100 posts of two different blogs and using the criteria of four human evaluators under the ROUGE methodology [4]. The conclusion is that the summaries guided by the comments are more accurate than the generic summaries.

The objective of this poster is to demonstrate that the post snippets biased by the readers' comments can be used to improve the accessibility to relevant posts. The quality of the snippets in web search has been already demonstrated as a determinant factor to improve the user access to relevant information [8]. We present a simple sentence extraction technique to generate the snippet for a post, introducing the comments' information to select good sentences from the post text. We propose an indirect evaluation of the snippets quality. We will compare the performance of our approach against generic summaries in terms of retrieval effectiveness. We assume that better snippets can be more effective in a retrieval task. This assumption allows a large scale evaluation that can be performed with TREC methodology, comparing the effectiveness of using or not the comments in terms of topic relevance, as defined in the TREC Blog Track.

In the next section approaches to generate generic and comments-biased snippets are presented. In section 3 the evaluation and results of both methods are reported. Finally conclusions and future work are commented in section 4.

2. COMMENTS-BIASED SNIPPETS

The way in which we decided to exploit the comments is using them to guide the snippet generation in a similar way of how a query is used in a query-biased summary. We designed a simple algorithm based on sentence extraction from the post text to generate Comments-Biased Snippets (CBS) that works as follows:

1. The comments are splitted in sentences and these are sorted according to their relevance to the post text.
2. Given the comments sentences in that relevance order, a novelty detector is applied to avoid redundant sentences, the most novel sentences (S_η) are selected until reach the 30% of the size of the comments size.
3. The post text is splitted in sentences and these are ranked according to their relevance to S_η.

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.
ACM 978-1-60558-896-4/10/07.

4. The most relevant and novel sentences of the post are selected until reaching 30% of the post original size.

In order to score the relevance of a sentence with respect to a piece of text we used the formulation of $R(s|q)$ as defined in [1], acting the piece of text as q. Cosine distance also as in [1] was used to filter redundant sentences. We have to remark that in no case comments' text is added to the summaries, the comments are only used to guide the selection of important post's sentences.

As baseline we created Generic Snippets (GS) without the use of comments information. To produce these snippets we followed exactly the same steps as presented before but the post text itself was used instead of the comments to guide the sentence selection process.

To test our approach, we had to automatically extract from the collection the posts and comments from the post-pages. This is indeed a messy task because this information is not standardly tagged in the collection. We applied a hybrid approach [2] with static templates for common blogging software platforms and a set of heuristics for the non commonly structured blogs.

3. EVALUATION AND RESULTS

3.1 Settings and Methodology

In order to compare both approaches we decided to use the standard TREC methodology allowing in this way a large-scale evaluation without depending on expensive evaluators' work. As we are dealing with blog snippets we decided to use the Blog Track approach evaluating the results under the topic relevance criteria. We used the Blogs06 collection [5] and topics 851-900 and 901-950 with "title only".

We created two different indices: one with the generic snippets and another one with the comments-biased snippets. Higher retrieval effectiveness using the CBS index should indicate that the relevant documents are more accessible using comments-biased snippets. We decided to use a high performance state-of-the-art retrieval model: BM25. The b value was trained for each index in queries $851-900$ optimizing MAP, for the other BM25 parameters recommended values were used. Topics 901-950 were used for test.

With the objective of performing a fair evaluation we decided to evaluate both approaches only over the subset of the collection where both post and comments were detected by our extraction approach (we have to remark than extraction accuracy can be still improved). Therefore from $3,215,171$ permalinks (post-pages) in the Blogs06 collection we selected $1,754,334$ post-pages. Note than doing this, the effectiveness values are lower than in the whole collection because many relevant documents are not considered.

3.2 Results

The evaluation results are summarized in Table 1. Our approach (CBS) achieves significant improvements for every setting and measure over the generic snippets (GS). In the test topics the improvement in terms of MAP, that was the measure optimized in training, achieved the 32%.

We also tested, in an additional experiment, the combination of both the post and comments to guide the sentence extraction process in the snippet generation. The effectiveness of this approach was only slightly better than using only the post (GS). This is explained because most of the sentences selected to guide the snippet construction were obtained from the post text (they were more relevant).

Table 1: Comparison between GS and CBS for training and test topics. Statistical significant improvements according with Wilcoxon Signed Rank test ($p-value < 0.01$) are starred, best values are bolded.

Topics	*Measure*	*GS*	*CBS*
$851-900$ T	MAP	0.0935	**0.1115***(+19.25%)
	$R-prec$	0.1744	**0.1957***(+12.21%)
	$bPref$	0.1813	**0.2150***(+18.59%)
$901-950$ T	MAP	0.0756	**0.1005***(+32.94%)
	$R-prec$	0.1333	**0.1612***(+20.93%)
	$bPref$	0.1360	**0.1703***(+25.22%)

In additional experiments we indexed the posts' text and the posts' text plus the comments, confirming the preliminary results of Mishne and Glance [6]. For instance, the MAP values in the testing topics were 0.1638 when only indexing the posts against 0.1971 when indexing posts and comments, and 0.1737 when indexing the full text of the web-page. Of course these values are higher than the ones achieved using snippets, but we have to remark that our objective was not to replace the posts as retrieval units but showing that better snippets are more relevant and therefore can improve the user access to the relevant posts.

4. CONCLUSIONS AND FUTURE WORK

In this poster we presented the use of readers' comments in blog search to generate better snippets an thus to improve the accessibility to relevant post in blog search results lists. The evaluation confirmed the suitability of the approach significantly improving the baseline method with gains up to 32%. As future work we would like to combine both the comments and the user queries to guide the snippet generation and perform an user involved evaluation with DUC methodology or crowdsourcing evaluation [9].

Acknowledgments: This work was funded by FEDER, *Ministerio de Ciencia e Innovación* and *Xunta de Galicia* under projects TIN2008-06566-C04-04 and 07SIN005206PR.

5. REFERENCES

[1] J. Allan, C. Wade, and A. Bolivar. Retrieval and novelty detection at the sentence level. In *Proc. of ACM SIGIR'03*, pp. 314–321, 2003.

[2] G. Attardi and M. Simi. Blog mining through opinionated words. In *TREC*, Special Publication 500-272. NIST, 2006.

[3] M. Hu, A. Sun, and E. Lim. Comments-oriented document summarization: understanding documents with readers' feedback. In *Proc. of ACM SIGIR'08*, pp. 291–298, 2008.

[4] C.-Y. Lin and E. Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proc. of NAACL'03*, pp. 71–78, 2003.

[5] C. Macdonald and I. Ounis. The TREC Blogs06 collection: Creating and analysing a blog test collection. *DCS TR, University of Glasgow*, 2006.

[6] G. Mishne and N. Glance. Leave a reply: An analysis of weblog comments. In *Third annual workshop on the Weblogging ecosystem*, Edinburgh, Scotland, 2006.

[7] I. Ounis, M. de Rijke, C. Macdonald, G. A. Mishne, and I. Soboroff. Overview of the TREC-2006 blog track. In *TREC*, Special Publication 500-272. NIST, 2006.

[8] R. W. White, J. M. Jose, and I. Ruthven. Using top-ranking sentences to facilitate effective information access. *JASIST*, 56(10): pp. 1113–1125, 2005.

[9] O. Alonso, R. Baeza-Yates, and M. Gertz. Effectiveness of Temporal Snippets. In *Proc. of WWW'09*, pp. 1113–1125, 2005.