

Blogs In Brief - Experiments in Query Result Presentation using MEAD

Alex Breuer
abreuer@cs.indiana.edu

Jacob Ratkiewicz
jpr@cs.indiana.edu

ABSTRACT

Traditional search engines provide links to pages - perhaps in conjunction with a few extracted sentences - in response to a query. We propose a system that presents search results over a blog corpus as a summary of the information in posts. We compute this summary using the MEAD multidocument summarization system [2]. We evaluate our results by comparing them to a summary computed by a simpler method, and to unsummarized results; the method for evaluation is a user study.

1. INTRODUCTION

Information and discussion in blogs can be powerful, as recent events concerning CBS news and Dan Rather have demonstrated. We seek to investigate if auto-summarization methods can assist in the consumption of blog data. In particular, we seek to answer the following question: given a topic, can auto-summarization produce a blog digest which contains more information than a single best-matching post, yet is still coherent and useful? To this end we build an archive from a blog crawl, implement summarizer pre-processing tools, and compare summarized results from queries on our archive against the top query hit.

2. CORPUS

2.1 Acquisition

Our initial corpus consisted of 10,000 blogs which we crawled from `livejournal.com` and `blogspot.com`. Our crawl was seeded by 200 initial hyperlinks - 100 from the Google query `political site:livejournal.com` and 100 from the query `political site:blogspot.com`. From these seed links, our crawler followed any links to `blogspot.com` and `livejournal.com` until it had downloaded 10,000 pages.

2.2 Post-Processing

Having collected these 10,000 pages, we used the Perl module `Lingua::EN::Identify` to remove all non-English blogs; from those remaining, we selected the 3,000 largest blogs based on file size.

As a single blog may contain information about a broad range of topics, we split all blogs into posts, and treated individual posts as documents. Such splitting is not entirely trivial; however, Blogspot blog editors tend to use `<div>` tags to delimit posts, and practically all blog editors yet encountered begin posts with timestamps, so we used a date

matching regular expression to split blogs in cases where `<div>` tags are not present.

We removed all navigation text that is not part of a post body, as well as blog-specific artifacts, such as followups and post author information.

2.3 Queries

Our queries are a number of simple, general queries related to current events; some examples are shown in Figure 1. We

chirac	guantanamo	bloomington
blogging	abu gharib	george bush
fox news	scalia	indonesia

Figure 1: Some queries used to select documents for summarization

generated results for these queries using the Lucene full text search engine [1], working over the extracted posts. The top 20 hits of the search constitute the set of documents used for input for our summarization methods.

3. SUMMARIZATION

We compared the results of several summarization methods using MEAD with another “naïve” summarization method and with a baseline method which simply presented the user with the post which was the top hit of the query.

The naïve summarizer produces a summary by simply combining the first sentences from the search results until the desired summary size is reached. MEAD produces its summaries by first ranking sentences according to any given set of features, then re-ranking sentences to avoid redundant sentences in the summary.

To produce a MEAD summary, we used the default MEAD settings, with a sentence feature ranking for query cosine similarity in addition to the standard centroid score and length rankings. Since MEAD is a multi-document summarization system, it can produce a summary of all 20 top post hits; this is what we did for MEAD without preliminary clustering. To address the anticipated problem of topical variance within the top 20 hits, MEAD with preliminary clustering employed an initial clustering of inputs, and MEAD was run on the clusters which were found. Posts were clustered with k -means run for $k \in [2, 4]$, and the value of k

which minimized accumulated centroid-document distance was chosen.

All summaries were produced with absolute compression rates of 200 words. The baseline approach, which simply returned the full text of the top post in the query results, was not restricted in length.

4. EVALUATION

The evaluation of our project was done through two user studies, as follows. The first study presented the user with a query, and four responses to the query, generated by the following methods:

1. No summarization (top hit)
2. Naïve summarization
3. MEAD summarization without preliminary clustering
4. MEAD summarization with preliminary clustering

The users were asked to assign two ratings - “best” and “most information.” We defined “best” to the user as “That which you would like most to see in response to the query given;” we defined “most information” simply as the query containing the most information (even if some of this information was off-topic or extraneous). The user study consisted of 14 queries and their associated response 4-tuples. Users could rate a single response as both “best” and “most information” if they so chose. Six users participated in this first study, ranking a total of 84 query response tuples. The results from this study are shown in Figure 2.

Following comments from some of these users that the first study was too long, we created a second, shorter user study. This study eliminated MEAD without preliminary clustering from the result presentation methods, and reduced the number of queries to seven. Two users participated in this second study, ranking 14 more queries. However, participation in this study was insufficient to draw conclusions, and its results should not be combined with our first study, as it is of a different design.

Method	voted best	voted most info
Top-hit	34%	35%
Naïve	28%	27%
MEAD n/cluster	7%	7%
MEAD w/cluster	31%	31%

Figure 2: Results from first user study

5. CONCLUSIONS & FUTURE WORK

In these experiments we developed several methods for summarizing the results of queries over blogs, an extremely noisy and heterogeneous domain. We evaluate these methods by means of a user study in which users rank the methods on their ability to present the information clearly and concisely. Although we found that MEAD without preliminary clustering does very poorly, we were unable to come to any statistically significant conclusion regarding the performance of the other two methods with respect to the baseline method. In particular, the naïve summarization method, which simply presented the first sentences from posts, performed at

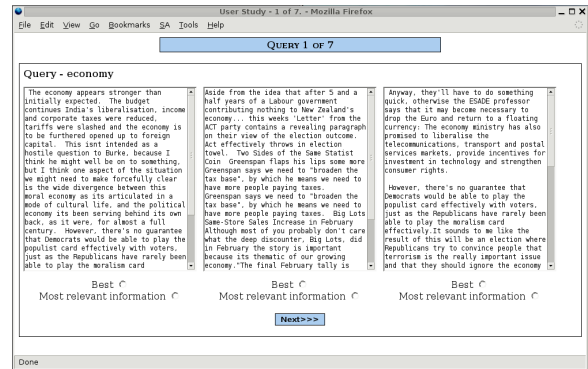


Figure 3: Screenshot of user study

the same level as did the method using MEAD with preliminary clustering, and received some “best” and “most information” votes for most queries. We speculate that this is due to a stylistic similarity between blog posts and news stories; the first few sentences of both tend to contain a large amount of the meaning of the document as a whole.

An issue which may have kept MEAD without preliminary clustering from performing well is that our queries were very general, and admitted posts with a very wide range of topics. If a result set included posts on completely different facets of an issue, the summary generated by MEAD would be less coherent, as the input set centroid would span many topics. MEAD with preliminary clustering performed better; however, we suspect that the preliminary clustering is not consistently effective. Very few documents get reassigned from their initial clusters, and we suspect that this is because the posts selected by the text search engine are already similar in frequency space.

Finally, it should be noted that we used MEAD essentially in its “out of the box” configuration, without changing any but one of its configuration options. A previous and similar application of MEAD, NewsInEssence [3], used a number of optimizations to enhance MEAD’s performance over news data; it could be instructive to examine if these would translate to increasing performance on blog data as well.

6. REFERENCES

- [1] Lucene - a free, open source full-text search engine. <http://lucene.apache.org>.
- [2] RADEV, D., ALLISON, T., BLAIR-GOLDENSOHN, S., BLITZER, J., ÇELEBI, A., DIMITROV, S., DRABEK, E., HAKIM, A., LAM, W., LIU, D., OTTERBACHER, J., QI, H., SAGGION, H., TEUFEL, S., TOPPER, M., WINKEL, A., AND ZHANG, Z. MEAD - a platform for multidocument multilingual text summarization. In *LREC 2004* (Lisbon, Portugal, May 2004).
- [3] RADEV, D. R., OTTERBACHER, J., WINKEL, A., AND BLAIR-GOLDENSOHN, S. NewsInEssence: Summarizing online news topics. *Communications of the ACM*, 2005.