Comments-Oriented Blog Summarization by Sentence Extraction*

Meishan Hu, Aixin Sun and Ee-Peng Lim
Centre for Advanced Information Systems
School of Computer Engineering
Nanyang Technological University, Singapore
{hu0004an,axsun,aseplim}@ntu.edu.sg

ABSTRACT

Much existing research on blogs focused on posts only, ignoring their comments. Our user study conducted on summarizing blog posts, however, showed that reading comments does change one's understanding about blog posts. In this research, we aim to extract representative sentences from a blog post that best represent the topics discussed among its comments. The proposed solution first derives representative words from comments and then selects sentences containing representative words. The representativeness of words is measured using *ReQuT* (i.e., *Re*ader, *Qu*otation, and *T*opic). Evaluated on human labeled sentences, *ReQuT* together with summation-based sentence selection showed promising results.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information filtering; H.3.1 [**Content Analysis and Indexing**]: Abstracting methods

General Terms

Experimentation

Keywords

Blog, Comments, Sentence Selection, ReQuT

1. INTRODUCTION

Entries of blogs, also known as blog posts, often contain comments from blog readers. A recent study on blog conversation showed that readers treat comments associated with a post as an inherent part of the post [2]. However, existing research largely ignore comments by focusing on blog posts only. To find out whether the reading of comments would change a reader's understanding about the post, we conducted a user study on summarizing blog posts by labeling representative sentences in those posts. Significant differences between the sentences labeled before and after reading comments were observed.

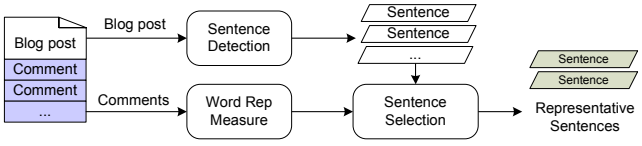

Figure 1: Comments-oriented blog summarization

In this research, we therefore focus on the problem of comments-oriented blog post summarization. The task is to summarize a blog post by extracting representative sentences from the post using information hidden in its comments. The extracted sentences represent the topics presented in the post that are captured by its readers (i.e., commenters). Many applications would benefit from comments-oriented summarization, such as blog search, blog presentation, reader feedback, and others.

Given a blog post and its comments, our solution consists of three modules (see Figure 1): *sentence detection* splits blog post content into sentences; *word representativeness measure* weighs words appearing in comments; and *sentence selection* computes a representativeness score for each sentence based on representativeness of its contained words. In this paper, we mainly focus on the last two modules. For word representativeness measure, we evaluate binary, comment frequency, term frequency, and *ReQuT*, where *ReQuT* measures the representativeness of a word from three aspects including *Re*ader, *Qu*otation, and *T*opic. To select sentences, we propose a summation-based sentence selection method. Together with *ReQuT*, the proposed sentence selection method performed well in our experiments evaluated on manually labeled sentences.

The rest of this paper is organized as follows. Section 2 surveys related work. We formally define the research problem in Section 3. The *ReQuT* model and the sentence selection method are given in Section 4. After presenting our user study and experiments in Section 5, we conclude the paper in Section 6.

2. RELATED WORK

Blogs have received much attention from researchers in recent years. Various studies have been conducted including blog posts tagging, spam blog post detection, and opinion

*This research is partially supported by grant SUG7/06, Nanyang Technological University, Singapore.

CIKM'07, November 6–8, 2007, Lisboa, Portugal.

mining, to name a few. Nevertheless, very few studies on blog comments and blog post summarization have been reported. In a recent study, Mishne and Glance reported that 28% of the collected 36,044 blogs contain comments from readers [6]. Among all blog posts containing comments, an average of 6.3 comments per post was observed. They also reported that comments contributed to the improvement of recall in blog search.

Zhou *et al* viewed a blog post as a summary of online news articles it linked to, with added personal opinions [9]. A summary is generated by deleting sentences from the blog post that are not relevant to its linked news articles. Comments associated with blog posts were however not used. The problem of comments-oriented blog summarization is quite related to the problem of identifying most commented sentences reported in [3]. Comments are represented and clustered using feature vectors, and a human expert is involved to select the clusters of interest. Sentences in blog post are scored and selected using comments in the selected clusters. Our solution, however, differs in two major aspects. First, we do not model comments using feature vectors. Second, our solution is topic neutral and does not involve user judgement.

Sun *et al* used LSA and Luhn's sentence selection methods to generate Web page summaries using clickthrough data [8]. In their work, clickthrough data is believed to provide some human understanding about Web pages. This is similar to our problem setting where comments of a post are utilized in summarizing the blog post.

3. PROBLEM DEFINITION

The problem of comments-oriented blog summarization is formally defined as follows:

DEFINITION 1. *Given a blog post P consisting of a set of sentences $P = \{s_1, s_2, \ldots, s_n\}$ and the set of comments $\mathcal{C} = \{c_1, c_2, \ldots, c_\ell\}$ associated with P, the task of comments-oriented blog summarization is to extract a subset of sentences from P, denoted by S_r ($S_r \subset P$), that best represents the discussion in $\mathcal{C}$.*

Given the problem, one straightforward approach is to compute a representativeness score for each sentence s_i, denoted by $Rep(s_i)$, and select sentences with representativeness scores above a given threshold[1]. As a sentence consists of a set of words, $s_i = \{w_1, w_2, \ldots, w_m\}$, one can derive $Rep(s_i)$ using representativeness scores of all words contained in s_i.

Intuitively, word representativeness can be measured by counting the number of occurrences of a word in comments, such as the following three schemes.

- *Binary.* With binary measure, $Rep(w_k) = 1$ if w_k appears in at least one comment and $Rep(w_k) = 0$ otherwise.
- *Comment Frequency* (CF). Similar to document frequency, $Rep(w_k)$ is defined by the number of comments containing word w_k.
- *Term Frequency* (TF). $Rep(w_k)$ is defined by the number of occurrences of w_k in all comments associated with a blog post.

[1]A threshold could be defined based on the number of sentences to be selected.

All three measures are simple statistics on comment content. Binary captures minimum information; CF and TF capture slightly more. Other information available in comments that could be very useful are ignored, e.g., authors of comments, quotations among comments and so on. Moreover, all three measures suffer from spam comments. For instance, a blog reader (or even the blogger himself) could intentionally write comments containing certain words in order to boost their representativeness, and hence to affect the summary generated. This calls for a measure that could capture more information from comments (besides content) and is less sensitive to spam.

4. REQUT MODEL

A comment, other than its content, is often associated with an author, a time-stamp, and even a permalink. A comment author is also known as a blog reader in this paper. We state three common observations on how comments may link to each other. These observations provide us guidelines on measuring word representativeness.

OBSERVATION 1. *A reader often mentions another reader's name to indicate that the current comment is a reply to previous comment(s) posted by the mentioned reader. A reader may mention multiple readers in one comment.*

OBSERVATION 2. *A comment may contain quoted sentences from one or more comments to reply these comments or continue the discussion.*

OBSERVATION 3. *Discussion in comments often branches into several topics and a set of comments are linked together by sharing the same topic.*

4.1 Reader-, Quotation- and Topic- Measures

Based on the three observations, we believe that a word is representative if it is written by authoritative readers, appears in widely quoted comments, and represents hotly discussed topics.

With Observation 1, given the full set of comments to a blog, we construct a directed *reader graph* $G_R := (V_R, E_R)$. Each node $r_a \in V_R$ is a reader, and an edge $e_R(r_b, r_a) \in E_R$ exists if r_b mentions r_a in one of r_b's comments. The weight on an edge, $W_R(r_b, r_a)$, is the ratio between the number of times r_b mention r_a against all times r_b mention other readers (including r_a). We compute reader authority using a PageRank [1] like algorithm, shown in Equation 1, where $|\mathcal{R}|$ denotes the total number of readers of the blog, and d is the damping factor as in PageRank.

$$A(r_a) = d \cdot \frac{1}{|\mathcal{R}|} + (1-d) \cdot \sum_{r_b} W_R(r_b, r_a) \cdot A(r_b) \quad (1)$$

$$RM(w_k) = \sum_{c_i \leftarrow r_a} tf(w_k, c_i) \cdot A(r_a) \quad (2)$$

The reader measure of a word w_k, denoted by $RM(w_k)$, is given in Equation 2, where $tf(w_k, c_i)$ is the term frequency of word w_k in comment c_i, and $c_i \leftarrow r_a$ means that c_i is authored by reader r_a.

With Observation 2, for the set of comments associated with each blog post, we construct a directed acyclic *quotation graph* $G_Q := (V_Q, E_Q)$. Each node $c_i \in V_Q$ is a comment, and an edge $(c_j, c_i) \in E_Q$ indicates c_j quoted

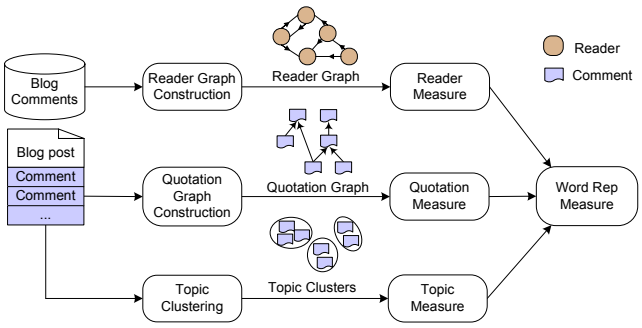

Figure 2: ReQuT Model

sentences from c_i. The weight on an edge, $W_Q(c_j, c_i)$, is 1 over the number of comments that c_j ever quoted.

We derive the quotation degree $D(c_i)$ of a comment c_i using Equation 3. A comment that is not quoted by any other comment receives a quotation degree of $1/|\mathcal{C}|$ where $|\mathcal{C}|$ is the number of comments associated with the given post.

$$D(c_i) = \frac{1}{|\mathcal{C}|} + \sum_{c_j} W_Q(c_j, c_i) \cdot D(c_j) \quad (3)$$

$$QM(w_k) = \sum_{w_k \in c_i} tf(w_k, c_i) \cdot D(c_i) \quad (4)$$

The quotation measure of a word w_k, denoted by $QM(w_k)$, is given in Equation 4 where $w_k \in c_i$ means that word w_k appears in comment c_i.

With Observation 3, given the set of comments associated with each blog post, we group these comments into *topic cluster*s using a Single-Pass Incremental Clustering algorithm presented in [7].

We believe that a hotly discussed topic has a large number of comments all close to the topic cluster centroid. Thus we have Equation 5 to compute the importance of a topic cluster, where $|c_i|$ is the length of comment c_i in number of words, $\mathcal{C}$ is the set of comments, and $sim(c_i, t_u)$ is the cosine similarity between comment c_i and the centroid of topic cluster t_u.

$$T(t_u) = \frac{1}{\sum_{c_j \in \mathcal{C}} |c_j|} \cdot \sum_{c_i \in t_u} |c_i| \cdot sim(c_i, t_u) \quad (5)$$

$$TM(w_k) = \sum_{w_k \in c_i, c_i \in t_u} tf(w_k, c_i) \cdot T(t_u) \quad (6)$$

Equation 6 defines the topic measure of a word w_k, denoted by $TM(w_k)$. In this equation, $c_i \in t_u$ denotes comment c_i is clustered into topic cluster t_u.

4.2 Word Representativeness Score

The representativeness score of a word $Rep(w_k)$ is the combination of reader-, quotation- and topic- measures in $ReQuT$ model, shown in Figure 2. The three measures are first normalized independently based on their corresponding maximum values and then combined linearly to derive $Rep(w_k)$ using Equation 7. In this equation α, β and γ are the coefficients ($0 \leq \alpha, \beta, \gamma \leq 1.0$ and $\alpha + \beta + \gamma = 1.0$).

$$Rep(w_k) = \alpha \cdot RM(w_k) + \beta \cdot QM(w_k) + \gamma \cdot TM(w_k) \quad (7)$$

As both readers and bloggers have no control on authority measure and very minimum control on quotation and topic measure, we argue that $ReQuT$ is less sensitive to spam comments.

4.3 Sentence Selection

Two sentence selection methods are evaluated in our experiments, namely Density-based selection and Summation-based selection.

Density-based selection (DBS) was proposed to rank and select sentences in question answering [5]. Given a set of weighted keywords representing a question, a sentence is scored using Equation 8, where K is the total number of keywords contained in s_i, $Score(w_j)$ is the score of keyword w_j, and $distance(w_j, w_{j+1})$ is the number of non-keywords (including stopwords) between the two adjacent keywords w_j and w_{j+1} in s_i. We adopted DBS in our problem by treating words appearing in comments as keywords and the rest non-keywords.

$$Score(s_i) = \frac{1}{K \cdot (K+1)} \cdot \sum_{j=1}^{K-1} \frac{Score(w_j) \cdot Score(w_{j+1})}{distance(w_j, w_{j+1})^2} \quad (8)$$

Summation-based selection (SBS), proposed in this paper, gives a higher representativeness score to a sentence if it contains more representative words. Nevertheless, SBS does not favor long sentences by considering the number of words in a sentence (see Equation 9). In this equation, $|s_i|$ is the length of sentence s_i in number of words (including stopwords), and $\tau (\tau > 0)$ is a parameter to flexibly control the contribution of a word's representativeness score.

$$Rep(s_i) = \frac{1}{|s_i|} \cdot \left(\sum_{w_k \in s_i} Rep(w_k)^\tau \right)^{\frac{1}{\tau}} \quad (9)$$

5. USER STUDY AND EXPERIMENTS

To the best of our knowledge, no similar user study has been conducted before; hence there is no benchmark dataset. We collected data from two famous blogs, i.e., *Cosmic Variance*[2] and *IEBlog*[3], both having relatively large readership and being widely commented. The former has more loyal but fewer readers with very diverse topics covered in posts; while the latter has less loyal but more readers, with topics mainly in Web development. Table 1 reports statistics of data collected[4].

5.1 User Study

With 10 posts randomly picked up from each of the two blogs and 3 human summarizers recruited from final-year Computer Engineering students, we conducted a user study on the impact of reading comments. Our hypothesis is that one's understanding about a blog post does not change after he or she read the comments associated with the post.

The user study was conducted in two phrases. In the first phrase, we provided 3 summarizers the 20 blog posts without comments and asked them to select approximately 30% of sentences from each post as its summary. The selected sentences served as a labeled dataset known as Reference-Set 1, or *RefSet-1* for short. In the second phrase, 3 human

[2] `http://cosmicvariance.com`
[3] `http://blog.msdn.com/ie`
[4] Note that "Pingback" and "Trackback" comments are excluded in our dataset

Table 1: Statistics of data from two blogs

Parameter	CosmicVariance	IEBlog
Number of blog posts	1114	364
Number of readers	2904	9490
Average post length	508.8	376.4
Average comments per post	22.1	66.8

Table 2: Level of self-agreement

Blog	H_1	H_2	H_3	Average
CosmicVariance	52.4%	40.3%	49.1%	47.3%
IEBlog	29.3%	19.4%	26.0%	24.9%

summarizers were provided the nearly 1000 comments associated with the 20 posts, and were asked to read both the posts and their comments, and again to summarize the posts by labeling approximately 30% of the sentences from each post. We name the second set of selected sentences Reference-Set 2, or *RefSet-2*.

We computed the level of peer-agreement for each pair of human summarizers. The averaged peer-agreement level in *RefSet-1* and *RefSet-2* are 37.8% and 32.6% respectively.

For each human summarizer, we computed the level of self-agreement shown in Table 2. Self-agreement level is defined by the percentage of sentences labeled in both reference sets against sentences in *RefSet-1* by the same summarizer. Recall our hypothesis is that one does not change his/her understanding about a blog post after reading comments, the expected level of self-agreement is 100% for every summarizer. The observed much lower self-agreement level is significant enough to invalidate our hypothesis[5]. That is, reading comments does change one's understanding about blog posts.

5.2 Experimental Results

As the sentences are labeled after reading comments, *RefSet-2* was used to evaluate the two sentence selection methods with four word representativeness measures. We adopted *R-Precision* and *NDCG* (see [4]) as performance metrics. In *NDCG*, the *Relevance Level* of a sentence is defined by the number of human summarizers labeled that sentence in *RefSet-2*. The reported results in Table 3 are averaged over all posts.

In our experiments, the similarity threshold in clustering comments was empirically set to 0.4; and parameter τ in SBS was set to 0.2. The three coefficients α, β, and γ in combining reader-, quotation-, and topic- measures were all 0.33.

As shown in Table 3, with either R-Precision or NDCG, SBS achieved better performance than DBS over all four word representativeness measures. For SBS method, *ReQuT* performed the best among the four word measures. Nevertheless, *ReQuT* together with SBS was not significantly better than other combinations according to our significance test. The possible reasons are: (i) the dataset is small and (ii) there is almost no spam comment in our dataset.

[5]The much lower self-agreement level on IEblog (compared with CosmicVariance) could be due to the fact that IEblog posts contain much more comments than that from CosmicVariance.

Table 3: Results in R-Precision and NDCG

R-Precision	Binary	CF	TF	ReQuT
DBS	0.4040	0.4202	0.4122	**0.4712**
SBS	0.4359	0.4496	0.4462	**0.5013**
NDCG	Binary	CF	TF	ReQuT
DBS	0.6526	0.6608	**0.6621**	0.6527
SBS	0.6614	0.6731	0.6769	**0.6794**

6. CONCLUSION

Based on the findings in our user study that reading comments does affect one's understanding about a blog post (and probably other kind of Web objects), we define the problem of comments-oriented blog post summarization. Our proposed solution measures word representativeness using information hidden in comments, and then selects sentences based on the representativeness of the words contained in sentences. Using human labeled sentences, we evaluated two sentence selection methods with four word representativeness measures. Among the latter, *ReQuT* gives the flexibility to measure word representativeness through three aspects, reader, quotation and topic. To study the impact of the three aspects in *ReQuT* is part of our future work.

7. REFERENCES

[1] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proc. of WWW'98*, pages 107–117, Brisbane, Australia, 1998.

[2] A. de Moor and L. Efimova. An argumentation analysis of weblog conversations. In *Proc. of Int'l Working Conf. on the Language-Action Perspective on Communication Modelling (LAP'04)*, New Brunswick, NJ, 2004.

[3] J.-Y. Delort. Identifying commented passages of documents using implicit hyperlinks. In *Proc. of HYPERTEXT'06*, pages 89–98, Odense, Denmark, 2006.

[4] K. Jrvelin and J. Keklinen. IR evaluation methods for retrieving highly relevant documents. In *Proc. of SIGIR '00*, pages 41–48, Athens, Greece, 2000.

[5] G. G. Lee, J. Seo, S. Lee, H. Jung, B.-H. Cho, C. Lee, B.-K. Kwak, J. Cha, D. Kim, J. An, H. Kim, and K. Kim. Siteq: Engineering high performance qa system using lexico-semantic pattern matching and shallow nlp. In *Proc. of TREC'01*, pages 437–446, 2001.

[6] G. Mishne and N. Glance. Leave a reply: An analysis of weblog comments. In *Proc. of WWW'06 Workshop on the Weblogging Ecosystem*, 2006.

[7] D. Shen, Q. Yang, J.-T. Sun, and Z. Chen. Thread detection in dynamic text message streams. In *Proc. of SIGIR '06*, pages 35–42, Seattle, Washington, 2006.

[8] J.-T. Sun, D. Shen, H.-J. Zeng, Q. Yang, Y. Lu, and Z. Chen. Web-page summarization using clickthrough data. In *Proc. of SIGIR'05*, pages 194–201, Salvador, Brazil, 2005.

[9] L. Zhou and E. Hovy. On the summarization of dynamically introduced information: Online discussions and blogs. In *Proc. of AAAI'06 Spring Symposium on Computational Approaches to Analyzing Weblogs*, March 2006.