Tag-Oriented Document Summarization*

Junyan Zhu[1], Can Wang[1], Xiaofei He[2], Jiajun Bu[1], Chun Chen[1], Shujie Shang[1],
Mingcheng Qu[1], and Gang Lu[3]
[1,2]College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China
[3]College of information, Zhejiang University of Finance and Economics, Hangzhou 310018, China
[1]{junyan_zhu, wcan, bjj, chenc, henochim, qumingcheng}@zju.edu.cn,
[2]xiaofeihe@cad.zju.edu.cn, [3]hz_lugang@163.com

ABSTRACT

Social annotations on a Web document are highly generalized description of topics contained in that page. Their tagged frequency indicates the user attentions with various degrees. This makes annotations a good resource for summarizing multiple topics in a Web page. In this paper, we present a tag-oriented Web document summarization approach by using both document content and the tags annotated on that document. To improve summarization performance, a new tag ranking algorithm named EigenTag is proposed in this paper to reduce noise in tags. Meanwhile, association mining technique is employed to expand tag set to tackle the sparsity problem. Experimental results show our tag-oriented summarization has a significant improvement over those not using tags.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information filtering

General Terms

Experimentation

Keywords

Document summarization, ranking, tag.

1. INTRODUCTION

The exponential growth of the WWW sees an increasing need of presenting Web pages in a condensed form to facilitate users' assimilation of vast information on the Web. Different from traditional documents, a page on the Web is generally less constrained in its content and organization and consequently may exhibit more diversity in topics contained in it. This poses a new challenge on traditional summarization techniques which focus on local contents of a document and may not be able to capture the true meaning of a Web page.

On the other hand, content users may be concerned with different topics in a Web page and various interaction data including clickthrough data, user comments etc. can be exploited to summarize different topics in a Web page. However, both comments and clickthrough data may contain a lot of noise in that the comments are generally informal and the users may click on irrelevant pages returned by a search engine. Social annotations (also known as tags), most of which being highly generalized descriptions of topics contained in a Web document, provide a good complement for summarizing multiple topics in a Web page.

Like any other forms of interaction data, there exist imprecise tagging and spam in annotations. Identifying quality tags by removing noise is critical in tag-based summarization. We notice that besides tag frequency, the mutually reinforcing property between users and tags can be exploited to evaluate the importance of tags. While a tag closely related to the topics in a Web document is generally regarded as a quality tag, a user producing many quality tags is regarded as a good user. In our tag-oriented summarization approach, we propose a new algorithm using linear transformation similar to that of HITS to estimate the importance of tags. The tags are further expanded to include related words using association mining techniques. The final summary is generated with a sentence evaluation based on expanded tags and TF-IDF of each word in a sentence. Experimental results show the tag-based approaches have achieved significant improvements over those not using tags.

2. SUMMARIZATION WITH TAGS

2.1 Ranking the Tags

A tagging system is comprised of resources, users and tags. A tag frequency ranking, which makes no discrimination among users, does not necessarily represent its usefulness for summarization. This is mainly due to tag spam, i.e. misleading tags which are generated to increase the visibility of some resources or simply to confuse users. Identifying spam users and remove tags annotated by them therefore should be regarded as an important complement to tag frequency ranking. In a word, tags from good users should be given more importance while the effects of the tag spam should be reduced.

Intuitively there exists a mutually reinforcing tagging relationship between users and tags. That is, quality tags are those annotated by many good users and a good user tends to create many quality tags. Based on this observation, we propose a new tag ranking algorithm using linear transformation similar to that used in HITS, namely EigenTag, to calculate the corresponding user scores and tag scores.

*The project is sponsored by S&T Planning Projects of Zhejiang Province, No.2007C13050

WWW 2009, April 20–24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

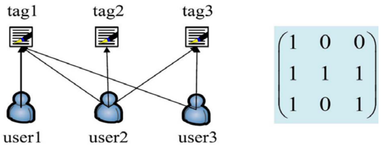

Figure 1: A tagging-graph and its adjacency matrix

Specifically, the relation between users and tags can be further illustrated by the tagging-graph and the corresponding adjacency matrix in Fig. 1. As shown in the Fig. 1, a directed edge from user i to tag j in the tagging-graph suggests user i annotates tag j on some resource. In the corresponding adjacency matrix, the (i, j)-th entry of M equals 1 if there exists an edge from user i to tag j and 0 otherwise. Let U be the vector of user scores $(u_1, u_2, \ldots, u_n)^T$, and T the vector of tag scores $(t_1, t_2, \ldots, t_n)^T$. Then the tag score and user score can be iteratively calculated by the following equations:

$$\mathbf{T} = M^{\mathrm{T}}\mathbf{U} \qquad \mathbf{U} = M\mathbf{T} \tag{1}$$

This iterative process updates tag score and user score repeatedly. In each step, the score of each tag is updated to the sum of the scores of all users annotating it. Then the score of each user is updated to the sum of scores of all tags annotated by him. The algorithm iteratively repeats the above two operations with normalization, until the tag and user scores converge. The final converged tag scores are the eigenvectors of the matrix $M^{\mathrm{T}}M$.

2.2 Tag Expansion And Scoring Associate Tag

Tag scoring for sentences is based on the precise match between tags and words in a sentence. Given that each document in our dataset has about less than 10 quality tags, the odd that an obviously relevant sentence using a related word (but not the tag its self) results in a mismatch is high. Heymann et al. encountered a similar issue in [1] and they used association mining techniques to expand the tag set. Here we use a similar approach, the FP-growth algorithm to expand the seed tag set. We use 101 blog posts from MSDN's IE blog site[1] to mine the related tag words. The seed tag set is expanded with the following associated tag scoring equation:

$$T(Assoc) = \left(\sum_{i=1}^{n} T(Tag_i) \times conf(Assoc|Tag_i) \right) /n \tag{2}$$

In the above equeation, $T(Assoc)$ is the associated tag score, $T(Tag_i)$ is the score for Tag_i in the seed tag set as obtained in the previous ranking step, $conf(Assoc|Tag_i)$ is the $Assoc$'s confidence value for Tag_i, and n is the number of the seed tags.

2.3 Summary Generation

Each word's score $Score(w_i)$ is a linear combination of its tag score and TF-IDF weight:

$$Score(w_i) = \lambda \times \frac{T(w_i)}{max(T(w_i))} + (1 - \lambda) \times \frac{tf \cdot idf(w_i)}{max(tf \cdot idf(w_i))} \tag{3}$$

Table 1: Comparison of three methods to score tags

	Tag TF-IDF	Tag Frequency	EigenTag
recall	0.640243	0.663948	**0.674985**
precision	0.681322	0.706868	**0.717717**
F-measure	0.663948	0.681271	**0.692448**

Table 2: Average summary quality using ROUGE-1

	OTS	Tag	TagEx	TagExDoc
recall	0.624720	0.674985	0.689922	**0.704172**
precision	0.623211	0.717717	0.719978	**0.722553**
F-measure	0.611889	0.692448	0.699850	**0.710302**

The sentence score is then the summed score of all the words in the sentence normalized by the number of the words in the sentence. The final document summary is generated by selecting the top-ranking sentences.

3. EXPERIMENTS AND RESULTS

Without existing benchmark dataset, we download 492 blog posts from MSDN's IE blog site and select 101 posts from them as our dataset. Approximately one-third of sentences from each post will be selected to form the document summary.

Standard Summarizations are provided by 6 evaluators. ROUGE-1 [2] is used to evaluate the effectiveness of the proposed methods.

We compare *EigenTag* with the other two scoring methods: *Tag TF-IDF* and *Tag Frequency*, which use tag's TF-IDF weight and tag frequency respectively to score tags. Table 1 shows the results of summarization using only tags. As can be seen, *EigenTag* outperforms the other two methods.

To compare document-oriented and tag-oriented approaches, the following summarization experiments using *EigenTag* as tag scoring method are performed: (1) *OTS*, Open Text Summarizer[2] that uses only document scoring; (2) *Tag*, in which word score is solely determined by tag score, i.e. $\lambda = 1$; (3) *TagEx*, in which word score is tag score using the expanded tag set, with $\lambda = 1$; (4) *TagExDoc*, in which word score is calculated using both document scoring and expanded tag scoring, i.e. $\lambda \in (0, 1)$; in our experiment we empirically set λ to 0.95. As shown in Table 2, all tag-oriented summarization approaches (i.e., *Tag*, *TagEx*, *TagExDoc*) outperform the document-oriented *OTS*, indicating that tags provide a good complement for summarizing Web documents. Not surprisingly, *TagExDoc*, which considers the information both from tags and web documents, achieves the best performance.

4. CONCLUSIONS

In this paper, we present a new tag-oriented summarization approach that effectively extracts user understanding of a Web document contained in social annotations. Experiments show that social annotations provide a good complement to the document content for summarizing multiple topics in a Web page.

5. REFERENCES

[1] P. Heymann, D. Ramage, and H. Garcia-Molina. Social tag prediction. In *SIGIR '08*.

[2] C. Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In *WAS '04*.

[1] http://blogs.msdn.com/ie/

[2] http://libots.sourceforge.net/