Enhancing Clustering Blog Documents by Utilizing Author/Reader Comments

Beibei Li
Department of Computer Science
University of Kentucky
Lexington, KY 40506-0046
(859)257-9348
beibei@csr.uky.edu

Shuting Xu
Department of Computer Information System
Virginia State University
Petersburg, VA 23906, USA
sxu@vsu.edu

Jun Zhang
Department of Computer Science
University of Kentucky
Lexington, KY 40506-0046, USA
(859)257-3892
jzhang@cs.uky.edu

ABSTRACT

Blogs are a new form of internet phenomenon and a vast ever-increasing information resource. Mining blog files for information is a very new research direction in data mining. We propose to include the title, body, and comments of the blog pages in clustering datasets from blog documents. In particular, we argue that the author/reader comments of the blog pages may have more discriminating effect in clustering blog documents. We constructed a word-page matrix by downloading blog pages from a well-known website and experimented a *k*-means clustering algorithm with different weights assigned to the title, body, and comment parts. Our experimental results show that assigning a larger weight value to the blog comments helps the *k*-means algorithm produce better clustering solutions. The experimental results confirm our hypothesis that the author/reader comments of the blog files are very useful in discriminating blog files.

Categories and Subject Descriptors

H.3.3. [**Information Search and Retrieval**]: Clustering, retrieval models, search process.

General Terms

Performance

Keywords

Blog, blogosphere, data mining, comment, clustering.

1. Introduction

Blogs, or weblogs as they are formally called, are dated, unedited, highly opinionated personal online commentary including hyperlinks to other resources. They are reverse chronological sequences of journal-like entries, maintained and published with blogging software [3]. Blogosphere is the collective term encompassing all blogs on the internet as a community or social network.

The modern blogs evolved from the online diary where people keep a running account of their personal lives and use blogs to publish their thoughts, feeling and viewpoints on whatever topics that may interest them. They are mostly maintained by individuals, though group blogs are increasingly popular as well. Blogs may be dedicated to just any topic, and there are many varieties of blog types and writing styles.

Estimation about the size of the blogosphere varies greatly and is a subject of constant debate. By July 2006, large blog search engines including Technorati were tracking more than 50 million blogs, and about 175,000 blogs were created daily. It is claimed that the size of the blogosphere doubles every six months [14].

Such a large, diversified, and ever-increasing information resource has only recently become the subject of research. A genre analysis of the blogosphere in 2004 [6] showed that more than two-thirds of public blogs are personal journals. Only 12 percent are 'filters', the link-driven blog type that used to be predominant in the early stages of the blogosphere. In the same statistics, knowledge blogs (k-blogs) used for collaborative knowledge management and topic discussion share a mere 3 percent of the public blogosphere. Due to the availability of (usually free) blogging software that makes the starting of blogs extremely easy for even people with limited knowledge of internet, the percentage of individual diary blogs increases steadily. At the time of writing this paper, it is estimated that about 75 percent of the blogs belong to this category. The blogosphere has hyper-accelerated the spread of information [16].

Such an enormous information bank cannot be over-looked by business leaders, government policymakers and researchers. Although blogs do not have a long history yet, research on blogs can first be built on the experience of various adjacent scientific areas. This includes data mining (web mining), social network analysis, economic research, as well as network research [1]. The blogosphere can be mined for the purpose of outreach opinion formation, maintaining online communities, supporting knowledge collaborative environments, monitoring the reaction to public events and is seen as a new alternative to the mass media [12].

Technical Report No. 462-06, Department of Computer Science, University of Kentucky, Lexington, KY, 2006. (December 4, 2006)

Although business blog mining is still in its infancy [7], social blog mining has not yet formally started, to the best of our knowledge. In this work, we intend to test some data mining techniques for mining blog files from the blogosphere, with a focus on mining the blogs of personal diary type. By mining such blogs, it may be possible for researchers to know what people are really thinking about certain things and how they express their feeling at the personal and private level. It may also be possible to develop smart computer-based techniques for blog surveillance to detect future potential troubling problems [2].

People tend to write their true opinions and feeling about something on their blogs. Such opinions and feeling are usually not expressed in surveys or polls. If such opinions and feeling are understood and dealt with early, certain potential problems could be fixed or avoided. For example, Jeff Weise, the 16-year-old Minnesota high school gunman who killed nine people before committing suicide, left a strange and scary writing on the internet [11]. He also maintained a blog at livejournal.com, last updated on January 27, 2005, less than two months before his shooting on March 21, 2005.

It is somehow difficult for the U.S. government to know the *real* opinions of ordinary people in many other countries. Opinions obtained from surveys or from newspapers are usually not the true ones, as the agencies and media are most likely controlled by the governments. But many people can write their personal opinions and feeling in their blogs and these blogs can be mined for useful information.

The major difference between blogs and the standard web pages is that the blogs are dated and most of them allow readers to place comments on each blog document. The user comments play an important role in the blogosphere, as it creates communication channels between the blog authors and the readers. Another important feature of the blogs is that blog authors can place individual blogs into different categories, according to some predefined categories, although the definitions of the categories may be different for different authors.

An important technique in data mining is to cluster relevant documents into different clusters according to their relevance. Algorithms for clustering web documents have been studied by many researchers [14,15]. However, as we pointed out, there is a significant difference between the standard web pages and the blog web pages. The special features of the blog files demands special treatments for better clustering results.

The main purpose of this paper is to examine the importance of utilizing comments in clustering blog documents (blog entries). Comments can be placed by the readers to reflect their views on particular blog entries. They can also be placed by the blog authors to explain their writing, or to answer some questions raised in the readers' comments.

This paper is organized as follows. Section 2 describes some basics about the blog documents. Section 3 describes the data preparation and clustering methods as well as the clustering metrics. Section 4 reports our experimental results using a collection of real-life blog files. We summarize this paper and discuss some interesting possibilities for further research in Section 5.

2. Blog Documents

We use vector-space model to encode the blog web pages. Specifically, each blog page can be viewed as a column vector in

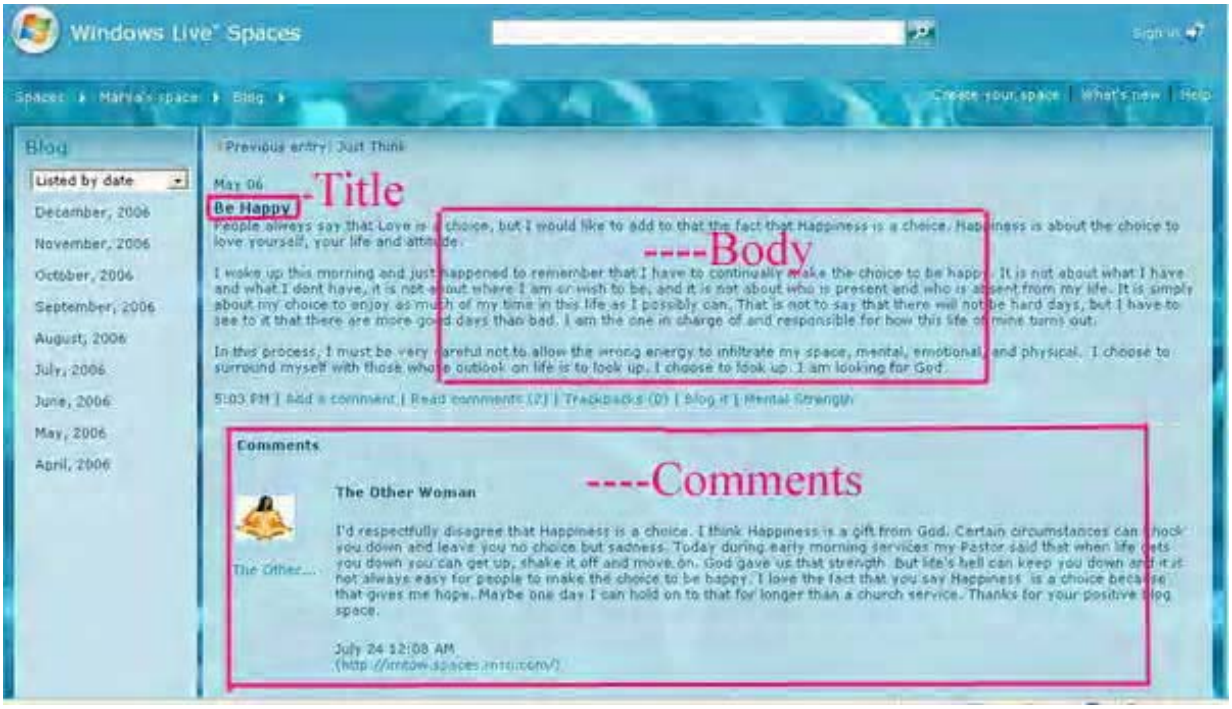

Figure 1. A sample blog page with title, body, and comments

a data matrix, and each word used can be considered as one row of the matrix. Each entry in the word-page matrix is the

frequency of occurrence of a particular word in a particular page. Different weight scheme can be used to emphasize or deemphasize certain words. The vector-space model and the word-page matrix representation are common in information retrieval [13].

We actually consider a blog page as three parts. The first part is the blog title, which is usually short. The second part is the blog body, i.e., the content of the blog page. The third part is the comments of the authors and/or the readers. Figure 1 is a snapshot of a sample blog page, with title, body, and comment clearly marked.

We hypothesize that the use of title and comment words in the dataset will enhance the discrimination of the blog pages and result in more accurate clustering solutions. Since the words in the comments reflect the specific views and questions and answers of the authors and the readers, they may hold more weights in discriminating individual blog pages.

3. Data Preparation and Clustering

In this section, we describe the methods used to prepare the data for processing, the *k*-means clustering algorithm, as well as the metrics to assess the quality of the clusters obtained.

3.1 Data Preprocessing

We selected three categories of blog files. One contains blogs related to gun control, another to church, and the third one to Alzheimer's disease. They were downloaded from Windows Live Spaces http://spaces.live.com by searching with the key words "Gun Control", "Church", and "Alzheimer's Disease". Each entry is an HTML file and has at least one comment. Each category has 70 files for a total of 210 blog files.

In the preprocessing phase, we transform the HTML files to a collection of numerical attribute vectors that the clustering algorithm can operate directly on. We detagged the HTML files and converted them into text files, each containing the title, body, and comments in separate parts. Then we stem the words to remove the common word endings. Next we delete stop words and count the number of occurrences of each word in the title, body, and comment parts of the document. Now we can represent each document by three vectors, namely v_t, v_b, and v_c, for the title, body, and comments part respectively. The vector for the whole document is a weighted sum of all three vectors:

$$v = w_t v_t + w_b v_b + w_c v_c$$

where w_t, w_b, and w_c are referred to as the title weight, body weight, and comment weight, respectively.

The word-page matrix A is composed of a set of such document vectors, each vector becomes a column in A. Thus $A = (v_1 \ldots v_m)$, where each column v_j corresponds to a blog document and each row to a particular word. v_{ij} is the weighted occurrences of the word i in the document v_j. To balance the influence of small size and large size documents, we scale each document vector v_j to have its Euclidean norm equal to 1.

3.2 Feature Selection

Feature selection is a widely used technique to build simpler and more comprehensible models, improve data mining performance, and help to prepare, clean, and understand data. It is very useful in text clustering to reduce data matrix dimension, decrease computational time, remove possible noise and improve clustering accuracy. As comments may reveal the views and questions and answers of the authors and the readers, we hypothesize that the terms selected in comments may be a good supplement to the terms selected in blog body and may help improve the clustering accuracy.

There are some unsupervised feature selection methods in literature. Notable among them are Document Frequency (DF) [15], Entropy Based Ranking (EN) [9], and Mean TFIDF (TI) [15]. Tang et al. showed in [15] that TI outperforms DF in some text clustering cases. Thus we choose TI as the feature selection method.

The term (word) frequency (*tf*) in the given document is the number of times a given term appears in the document. This count is usually normalized to prevent a bias towards longer documents:

$$tf = \frac{d_i}{\sum_{k=1}^{m} d_k}$$

where d_i is the number of occurrences of the term i, and the denominator is the number of occurrences of all terms.

The inverse document frequency (*idf*) is a measure of the general importance of the term, which is defined as:

$$idf = \log \frac{d}{t_i}$$

Where d is the total number of documents and t_i is the number of documents that contain term i.

Thus we define *tfidf* = *tf* * *idf*. A high weight in *tfidf* is reached by a high term frequency and a low document frequency of the term in the whole collection of documents. The weight hence tends to filter out common terms.

TI is the mean value of *tfidf* over all the documents for each term [15]. We can use TI to measure the quality of the term. The higher the TI value, the better the term to be ranked.

3.3 Clustering

The *k*-means algorithm [10] (with its many variants) is a popular clustering method for text and web collections [17,18]. It gained its popularity due to its simplicity and intuition. The algorithm is an iteration procedure and requires that the number of clusters, k, be given a priori. Suppose that the k initial cluster centers are given, the algorithm iterates as follows:

(1) It computes the Euclidean distance from each of the documents to each cluster center. A document is assigned to the cluster with the smallest distance.

(2) Each cluster center is recomputed to be the mean of its constituent documents.

(3) Repeat steps (1) and (2) until the convergence is reached.

The criterion function for the convergence can be computed, e.g., as

$$f_r = \frac{1}{m}\sum_{i=1}^{m} Edist^2(v_i, c_j^{(r)}),$$

where r is the step of the iterations. The function $Edist(v_i, c_j)$ computes the Euclidean distance from the document v_i to a cluster center c_j. Given a convergence criterion ε, the *k*-means algorithm stops when $|f_{r+1} - f_r| < \varepsilon$. Note that f_r is a monotonically decreasing function with a lower bound. So its limit exists [5].

3.4 Clustering Metrics

We evaluated the quality of the clusters by computing the two metrics entropy and purity [19]. The measure "entropy" gauges the distribution of each class of documents within each cluster. The measure "purity" computes the extent to which each cluster contains documents from primarily one class. Suppose there are q classes and the clustering algorithm returns k clusters, the entropy E of a cluster S_r of size n_r is computed as

$$E(S_r) = -\frac{1}{\log_2 q}\sum_{i=1}^{q}\frac{n_r^i}{n_r}\log_2\frac{n_r^i}{n_r}$$

where n_r^i is the number of documents in the ith class that are assigned to the rth cluster. The entropy of the entire clustering solution is computed as:

$$\text{Entropy} = \sum_{r=1}^{k}\frac{n_r}{n}E(S_r)$$

In general, the smaller the entropy value, the better the clustering solution. Analogously, the purity of the cluster S_r can be defined as

$$P(S_r) = \frac{1}{n_r}\max_i(n_r^i)$$

and the purity value of the entire clustering solution is computed as

$$\text{Purity} = \sum_{r=1}^{k}\frac{n_r}{n}P(S_r)$$

It is clear by definition that the larger the purity value, the better the clustering solution.

4. Experimental Results

We conducted some experiments to cluster the above mentioned blog set. We wanted to test the contribution of each part of a blog to the clustering accuracy.

4.1 Influence of Weight

To test the influence of the relative weights of the title, body, and comment, on the accuracy of the clustering solutions, we varied the weight triple (w_t, w_b, w_c). We ran the standard *k*-means algorithm with an arbitrary starting guess of three clusters. After it converged, we computed the entropy and purity values. (See Table 1).

It can be seen that the clustering solutions were not very good if we only used one of the title, body, or comment parts. The entropy for clustering only the title is as high as 0.9512. The accuracy of clustering the blog body is better than clustering the title or comment alone. However, we can see its entropy is still higher than 0.8. Using all of the three parts improves the clustering solution a lot. When the weight for the title, body, and comment are all ones, the entropy drops to 0.7504 and the purity increases to 0.6143.

In Table 1, we also show the results of increasing the comment weight or increasing the title weight. When the comment weight is 3, 4, or 5, the clustering results is greatly improved. In particular, when we increased the comment weight to 3, we obtained the best clustering solution.

Table 1. Influence of weight of the title, body, and comment on the clustering accuracy

Title Weight	Body Weight	Comment Weight	Entropy	Purity
1	0	0	0.9512	0.4143
0	1	0	0.8259	0.5619
0	0	1	0.9244	0.4762
1	1	1	0.7504	0.6143
1	1	2	0.7698	0.6238
1	**1**	**3**	**0.5109**	**0.8190**
1	1	4	0.5470	0.8000
1	1	5	0.5457	0.8000
1	1	0	0.8619	0.5714
2	1	0	0.8619	0.5714
3	1	0	0.8457	0.5810
4	1	0	0.8498	0.5810
5	1	0	0.8781	0.5571

The experimental results confirmed our hypothesis that the blog comments should be utilized effectively in data mining algorithms when dealing with the blog files.

For this particular dataset, increasing the title weight does not seem to improve the clustering solution as appreciably as increasing the comment weight does. One reason might be that there are usually too few words in the title. However, we can still see a slight improvement of the purity and entropy measures in the clustering solutions, especially when the title weight is increased from 2 to 3.

4.2 Feature Selection

In this section we use feature selection to filter out unimportant words (terms) and cluster on the word-page matrix composed by only the important words (terms).

Table 2. Influence of feature selection percentage on clustering accuracy

w_t	w_b	w_c		100%	50%	30%	20%	10%
1	1	0	Entropy	0.8619	0.8619	0.8619	0.8619	0.8619
			Purity	0.5714	0.5714	0.5714	0.5714	0.5714
1	1	1	Entropy	0.7504	0.7504	0.6814	0.6814	0.7550
			Purity	0.6143	0.6143	0.6143	0.6143	0.6095

Table 2 shows the clustering accuracy with the percentage of the features selected.

If we use only the title and the body of a blog for clustering, then reducing the percentage of the features used will not change the clustering accuracy. If we apply feature selection to all the blog content including the comments, then we can see with certain percentage of features selected (30% and 20% in this experiment) the entropy value can be reduced. This experiment also shows that making good use of the words (terms) in the comments can help increase the clustering accuracy.

5. Summary

We proposed to make use of the blog comments in clustering blog documents. It is our understanding that the blog comments, leaving either by the authors or by the readers, contain higher level discriminating property in classifying blog documents than the body of the blog entries themselves. By adding more weight to the blog comments, our experiments using real-life blog files show that we can get better clustering solutions with the *k*-means algorithm.

Blogs are relatively new and more studies are needed to understand the blogosphere. This paper represents a first step in utilizing a particular feature of the blogs, the comments, to enhance the effectiveness of a clustering algorithm in classifying blog pages.

Further research work along this line can consider the timing effect of the blogs. Because blog pages are dated, entries published at different time may have different relevance to other blogs. In better clustering blog documents, or in finding blog communities, it may be important to make use of the time information [8]. The utilization of predefined category information may also improve the classification of blog files.

Since blogs are published by amateur authors, most of them are non-professional writers; the blog language tends to be informal and incomplete, and contains errors and typos. It is possible to use singular value decomposition (SVD) technique (latent semantic indexing or LSI) to remove the noise of the databases [4]. The blog texts are normally unedited and the structures are nonuniform, it is crucial that robust preprocessing techniques be used before the dataset is mined. Because of the wide demography of the blog authors, the use of SVD to deal with synonym and polysemy can be advantageous.

We are also interested in experimenting other data mining algorithms with blog datasets. We are in the process of building larger blog datasets with blogs from different sources.

6. ACKNOWLEDGMENTS

The research work of J. Zhang was supported in part by the National Science Foundation under grants CCR-0092532 and CCF-0527967, in part by the Kentucky Science and Engineering Foundation under grant KSEF-148-502-05-132, and in part by Alzheimer's Association under grant NIRG-06-25460.

7. REFERENCES

[1] Aschenbrenner, A., and Miksch, S. *Blog Mining in a Corporate Environment, Te*chnical Report ASGAARD-TR-2005-11, Smart Agent Technologies, 2005.

[2] Berry, M. W., and Browne, M. Email surveillance using non-negative matrix factorization, *Computational & Mathematical Organization Theory*, 11 (2005), 249-264.

[3] Blood, R. *The Weblog Handbook: Practical Advice on Creating and Maintaining Your Blog*, Perseus Publishing, Cambridge, MA, 2002.

[4] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. Indexing by latent semantic analysis, *Journal of the Society of Information Science*, 41(1990), 391-407.

[5] Dhillon, I. S., and Modha, D. S. Concept decompositions for large sparse text data using clustering. *Machine Learning*, 42, 1 (2001), 143-175.

[6] Herring, S. C., Scheidt, L. A., Bonus, S., and Wright, E. Bridging the gap: A genre analysis of weblogs. In *Proceedings of the 37th Hawaii International Conference on System Sciences*, 2004.

[7] Hoyt C. Mining the blogosphere, *the HUB Magazine*, January 10, 2006, http://hubmagazine.com/?p=76, last accessed on October 30, 2006.

[8] Kumar, R., Novak, J., Raghavan, P., and Tomkins, A. On the bursty evolution of Blogsphere. In *WWW2003*, (Budapest, Hungary, 2003).

[9] Liu, H., Li, J., and Wong, L. A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. *Genome Informatics*, 13(2002), 51-60.

[10] MacQueen, J. B. Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Symposium on Mathematics*, Statistics, and Probability, University of California Press, 1967, 281-297.

[11] Malkin, M. All about the Minnesota school shooter, March 23, 2005, http://michellemalkin.com/archives/001837.htm, last accessed on November 1, 2006.

[12] Nicolov, N., Salvetti, F., Liberman, M., and Martin, J.H. Computational approaches to analyzing weblogs. In *Papers from 2006 AAAI Spring Symposium*, 2006.

[13] Salton, G., and McGill, M.J. *Introduction to Modern Retrieval*, McGraw-Hill, New York, NY, 1983.

[14] Sifry, D. Sifry's alerts, at http://www.sifry.com/alerts/archives/000436.html, accessed on October 31, 2006.

[15] Tang, B., Shepherd, M., Milios, E., and Heywood, M. Comparing and combing dimension reduction techniques for efficient test clustering, In *Proceedings of the Workshop on Feature Selection for Data Mining*, SIAM Data Mining, 2005.

[16] Torio, J. *Blogs, A Global Conversation*, Master's Thesis, Syracuse University, 2005

[17] Xu, S., and Zhang, J. A parallel hybrid web document clustering algorithm and its performance study, *Journal of Supercomputing*, 30(2004), 117-131.

[18] Zamir, O., and Etzioni, O. Web document clustering: A feasibility demonstration. In *SIGIR'98*, (Melbourne, Australia, 1998).

[19] Zhao, Y., and Karypis, G. *Criterion Function for Document Clustering Experiments and Analysis*, Technical Report #01-40, Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN, 2001.