FAIR: A Fuzzy ART Network Based Scheme for Retrieving Useful Information from Blogs

Long-Sheng Chen* and Zue-Cheng Lin

***Abstract*—Blogs could be viewed as the 4th crucial Internet application, after E-mail, Instant Message, and Bulletin Board System (BBS). The business world has experienced significant influence by the blogosphere. A hot topic in the blogosphere may affect a product's life period. Moreover, an exposure of an inside story in the blogosphere may influence a company's reputation. Nowadays, a lot of companies attempt to discover useful knowledge from that huge amount of blogs for business purposes. Therefore, the major objective of this study is to propose a Fuzzy Adaptive Resonance Theory (ART) network based Information Retrieval (FAIR) scheme by combining Fuzzy ART neural network, Latent Semantic Indexing (LSI), and association rules (AR) discovery to extract knowledge from blogs. In the proposed FAIR, Fuzzy ART network firstly has been employed to segment bloggers. Next, for each customer segment, we use LSI technique to retrieve important keywords. Then, in order to make the extracted keywords understandable, association rules mining is presented to organize these extracted keywords to form concepts. Finally, a real case of cosmetics products evaluation has been provided to demonstrate the effectiveness of the proposed FAIR scheme.**

***Index Terms*—Association Rule Discovery, Blogs, Fuzzy Adaptive Resonance Theory Neural Network, Information Retrieval, Latent Semantic Indexing.**

I. Introduction

Blogs are one of the fastest growing sections of the Internet and are emerging as an important communication mechanism that is used by an increasing number of people [11]. A blog can be considered as a journal that can continuously allow the users update their own words and post their work online through software. Bloggers (blogs users) often make a record of their lives and express their opinions, feelings, and emotions through writing blogs [12]. One of the most important features in blogs is the ability for any reader to write a comment on a blog entry. This ability has facilitated the interaction between bloggers and their readers [13]. In blogs, lots of cyber communities have also emerged [14] and become a new way to discuss specific issues, such as infectious diseases [15], cancers [16], online campaigning [17], research topics [18], tourism promotion [19] and so on. Nowadays, the business world has experienced significant influence by the blogosphere. A hot topic in the blogosphere may affect a product's life period. Moreover, an exposure of an inside story in the blogosphere may influence a company's reputation [28]. Lots of companies realized blogs might be a whole new channel of promotion and begin to study how to discover useful knowledge for business purposes.

Following this trend, researchers have paid more and more attentions to study some issues regarding blogs. Todoroki *et al*. [18] propose to utilize a blog as an electronic research notebook, since a blog system provides user-friendly interface compatible with web browsers, easy-to-use authoring tools and full-text retrieval. Chau and Xu [13] present a semi-automated approach to facilitate the monitoring, study, and research on blogs of online hate groups. Lin and Huang [19] indicate that blogs can significantly influence browsers and indirectly promote tourism. Du and Wanger [23] seek to explore blogs' success factors from a technology perspective. Asano [24] investigated whether a 'fiction novel' on blogs describing a girl undergoing epilepsy surgery can potentially facilitate familiarity to epilepsy surgery among the general Internet users in Japan. Most of related studies have been conducted to measure the influence of the blogosphere However, relatively few papers discuss extracting knowledge from blogs, such as usage mining [25] and structure mining [13]. And, it just has relatively few works to discuss blog content mining.

Current blog content mining focuses on extracting useful information from blog entry collections, and determining certain trends in the blogosphere [29]. Latent Semantic Analysis (LSA) is proposed for mining content from blogs. Besides, to broaden the usefulness of the blog search engine, Probabilistic Latent Semantic Analysis (PLSA) is applied to detect the keywords from various blog entries with respect to certain topics. This simple algorithm presents the blogosphere in terms of topics with measurable keywords, and hence it tracks the popular conversations and topics in the blogosphere. However, the keywords extracted by those Information Retrieval (IR) techniques are not very easy to be understood.

This work proposed a Fuzzy Adaptive Resonance Theory (ART) network [2] based Information Retrieval (FAIR) scheme by integrating Fuzzy ART, Latent Semantic Indexing (LSI) [6], and association rules discovery [9], [10] to extract knowledge from blogs contents. In FAIR, Fuzzy ART network first has been employed to segment bloggers. Next, for each customer segment, we use LSI technique to retrieve important keywords. Then, in order to make the extracted keywords understandable, association rules discovery is presented to organize these extracted keywords to form

Manuscript received January 7, 2009. This work was supported in part by the National Science Council of Taiwan (Grant No. NSC 96-2416-H-324 -003 -MY2) .

Long-Sheng Chen is an assistant professor of the Department of Information Management, Chaoyang University of Technology, Taichung 41349, Taiwan (e-mail: lschen@cyut.edu.tw).

Zue-Cheng Lin is a graduate student of the Department of Information Management, Chaoyang University of Technology, Taichung 41349, Taiwan (e-mail: s9614638@cyut.edu.tw).

concepts. Finally, a real case of cosmetics products has been provided to demonstrate the effectiveness of the proposed FAIR scheme.

II. Latent Semantic Indexing

Most of bloggers' comments are written in text format. Information retrieval (IR) techniques can extract useful knowledge from textual data. That's also the reason why we employed IR techniques to mine useful knowledge in this study. The vector space model is a traditional information retrieval model that represents documents and queries as vectors in a multi-dimensional space. When indexing terms are extracted from a document collection, each document is represented as a vector of term frequencies. Similarity comparisons among documents and/or between documents and queries are made by the similarity between two vectors [26]. Among lots of IR techniques, the text retrieval method using latent semantic indexing (LSI) [6] technique with truncated singular value decomposition (SVD) is a well-known approach [21]. It has been intensively studied in recent years. LSI has been applied to a wide variety of learning tasks, such as search and retrieval, classification and filtering [22].

LSI is also one of vector space approaches for modeling documents, and it was reported that this technique can bring out the "latent" semantics in a collection of documents [6]. In addition, LSI can handle "Polysemy" and "Synonymy" problems in text mining related area. The SVD reduces the noise contained in the original representation of the term–document matrix and improves the information retrieval accuracy [21]. In other words, LSI which is an automatic method that transforms the original textual data to a smaller semantic space by taking advantage of some of the implicit higher-order structure in associations of words with text objects [7], [6]. It has been reported that SVD can be applied to education, solving linear least-squares problems, and data compression [8].The transformation is computed by applying truncated SVD to the term-by-document matrix. After SVD, terms which are used in similar contexts will be merged together.

III. Proposed FAIR scheme

This section will introduce the implemental procedure of proposed FAIR scheme.

A. The Procedure of Proposed Scheme

In this section, we will introduce the procedure of our proposed FAIR scheme. As shown in Fig. 1, FAIR scheme can be divided into 5 steps. They are

1) Data collection
2) Data preprocess
3) Clustering
4) Information retrieval
5) Association rules discovery

In step 1, we collect some bloggers' comments regarding cosmetics products. Then, these collected text examples should be preprocessed to construct a document-term matrix. Some jobs such as word segmentation can be done in step 2. In step 3, we employ Fuzzy ART neural network to segment bloggers. The main purpose of this step is to gather similar customers into together. Then, we can extract more specific information for every single one customer segment. In step 4, LSI is used to extract important keywords of each customer segment. In this step, we can imply the interested topics or issues in different customer segment by using those acquired keywords. However, if too many keywords are extracted, it's not easily to be understood. Therefore, we introduce association rules discovery in step 5. The association rules which provide the information between/among keywords can help us to transform keywords into concepts. Finally, we can understand what customers think about cosmetics products. The extracted information can be provided to enterprises for commercial decision making. The detailed information about these five steps can be found in subsections B & C.

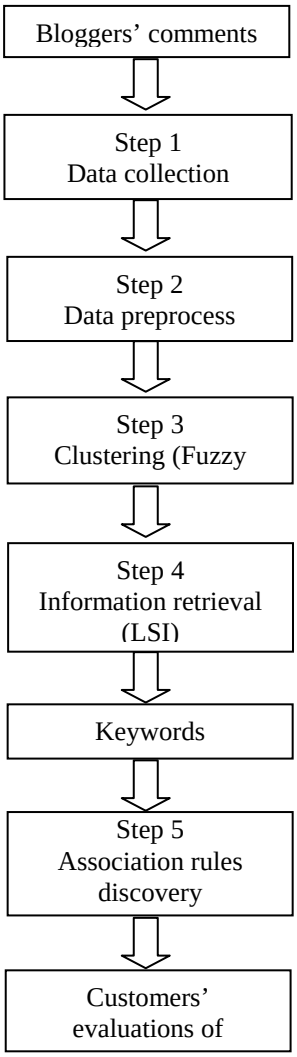

Fig. 1 The procedure of the proposed FAIR scheme

B. Data Collection and Data Preprocess

This study aims to discover knowledge regarding cosmetics products evaluation from blogs content. Before implementing data mining tasks, these collected should be preprocessed. A brief process of data preprocess has been shown in Fig. 2. To segment words, the CKIP Chinese word segmentation system of Academia Sinica (Taiwan) is

employed to segment words. The detailed information of CKIP can be found in [1]. After that, a document-term matrix can be constructed for further data mining.

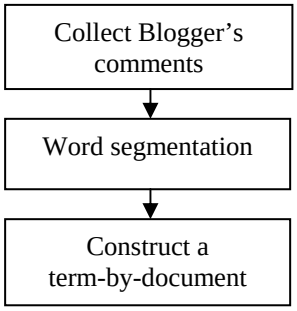

Fig. 2 The steps of data preprocess

C. Fuzzy ART Neural Network

Next Fuzzy ART neural network was utilized to segment customers (bloggers). Fuzzy ART is a famous method of clustering. Instead of constructing clusters by a given number of clusters, it assigns examples onto the same cluster by comparing their similarity. The major difference between Fuzzy ART and other unsupervised neural networks is the so called vigilance parameter (ρ) [4], [5]. The Fuzzy ART network allows the user to control the degree of similarity of patterns placed on the same cluster.

Fuzzy ART network clusters instances depends on two distance criteria, match (S) and choice function (T). For input vector I and category j, the match function is defined as (1).

$$S_j(I) = \frac{|I \wedge w_j|}{|I|} \tag{1}$$

where w_j is an analog-valued weight vector associated with cluster *j*. Symbol $\wedge$ represents the fuzzy AND operator. For example, the result of $(a \wedge b)_i$ must be $min(a_i, b_i)$. The norm $|...|$ is defined by $|a| = \sum_i |a_i|$.

The choice function is defined as (2).

$$T_j(I) = \frac{|I \wedge w_j|}{a + |w|} \tag{2}$$

where α is a small constant. Increasing α biases the search more towards clusters with large w_j . Each input vector is assigned to the category that maximizes $T_j(I)$ while satisfying $S_j(I) \geq \rho$, where the vigilance ρ, is a constant, $0 \leq \rho \leq 1$.

Fuzzy ART has three parameters, the choice parameter (α), the learning parameter (β), and the vigilance parameter (ρ) needed to be tuned. According to suggestion of Burke and Kamal (1995)[3], the choice parameter ($\alpha > 0$) is suggested to be close to zero. The learning parameter (β) which defines the degree to which the weight vector, w_j , is updated with respect to an input vector claimed by node J. In the fast-learning mode, Carpenter *et al*. [2] suggest that $\beta = 1$. In the fast commit-slow recode mode, $\beta = 1$ for first-time commitments (fast learning/commitment) and $\beta < 1$ (slow recode) otherwise. This study uses the fast commit-slow recode option here.

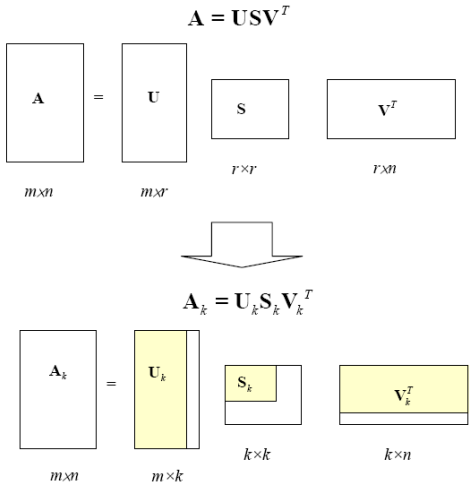

Fig. 3 The singular value decomposition [6]

D. Information Retrieval

After clustering, we employ LSI to realize the discussed topics in every cluster. Fig. 3 briefly introduces the concept of SVD. Let A be an $m \times n$ matrix of rank r whose rows represent documents and columns denote terms (variables). Let the singular values of A (the Eigen values of $A \cdot A^T$) be $\sigma_1 \geq \sigma_2 \geq ..... \geq \sigma_r$. The singular value decomposition of A expresses A as the product of three matrices $A = USV^T$, where $S = diag(\sigma_1,...,\sigma_r)$ is an $r \times r$ matrix, $U = (u_1,...,u_r)$ is an $m \times r$ matrix whose columns are orthonormal, and $V^T = (v_1,...,v_r)^T$ is an $r \times n$ matrix. LSI works by omitting all but the k largest singular values in the above decomposition, for some suitable k (k is the dimension of the low-dimensional space). It should be small enough to enable fast retrieval and large enough to adequately capture the structure of the corpus. Let $S_k = diag(\sigma_1,...,\sigma_k)$, $U_k = (u_1,...,u_k)$ and $V_k = (v_1,...,v_k)$. Then $A_k = U_k S_k V_k^T$ is a matrix of rank k, which is the approximation of A. The rows of $V_k S_k$ above are then used to represent the documents. In other words, the row vectors of A are projected to the k-dimensional space spanned by the row vectors of U_k; we sometimes call this space the LSI space of A.

E. Association Rules Discovery

When you submit your final version, after your paper has been accepted, prepare it in two-column format, including figures and tables. Association rule mining searches [9], [10] for interesting relationships among items in a given data set [27]. The details of association rule mining [10], [27] are as follows:

Let I= $\{i_1, i_2, \ldots, i_m\}$ be a set of literals, called items.
Let D be a set of transactions, where each transaction T is a set of items such that T $\subseteq$ I. Each transaction has a unique identifier, called its TID.

Let X be a set of items in I. A TID T is said to contain itemset X, if and only if $X \subseteq T$. An itemset is any subset of the set of all items, I.

An association rule is an implication of the form $X \Rightarrow Y$, where $X \subset I$, $Y \subset I$ and $X \cap Y = \varnothing$. The rule $X \Rightarrow Y$ has the implication that the occurrence of itemset X in a TID T infers itemset Y also occurs.

The standard measures to assess association rules are the support and confidence. The rule $X \Rightarrow Y$ holds in the transaction D with confidence c if c% of transactions in D that contain X also contain Y. The rule $X \Rightarrow Y$ has support s in the transaction set D if s% of the transactions in D contain both X and Y. An itemset containing k items is referred to as a k-itemset. A large (frequent) itemset is an itemset whose support is above a threshold. The set of large k-itemsets is commonly denoted by L^k. To find L^k, a set of candidate k-itemsets denoted by C^k is generated by joining L^{k-1} with L^{k-1}.

IV. Implementations

A. Use Data Collection and Data Preprocess

We collect 150 comments from the following famous blogs in Taiwan. After removing some useless articles such as those who only contains meaningless icons or something not readable, 100 bloggers' comments are left for further analysis.

Fig. 4 provides an example of a blogger's comment.

http://www.urcosme.com/index.htm
http://www.wretch.cc/blog
http://www.wretch.cc/blog/kiki185371/10029858
http://www.wretch.cc/blog/Alady/24080782
http://www.wretch.cc/blog/csmeow/4101936
http://blog.webs-tv.net
http://tw.blog.yahoo.com
http://www.pixnet.net/blg/

Fig. 5 is a part of results of CKIP system. 283 keywords are obtained after word segmentation processing of CKIP. In order to reduce the size of dimensionality, we remove those with very low occurrence frequency. Finally, 144 keywords are left to construct document-term matrix. Table. 1 shows a part of constructed document-term matrix.

Fig. 4 An example of a blogger's comment toward cosmetics products

Fig. 5 A part of results in CKIP system

Table. 1 The document-term metrix

	Acridness	Bright	Irritation	Refreshing	Whitening	Price	.
#1	1	0	0	1	1	0	.
#2	1	1	0	0	0	1	.
#3	0	0	1	0	1	0	.
#4	0	1	1	1	1	0	.
#5	1	0	0	0	0	1	.
..	..	..	..	..	..	..	.
	..	..	..	..	..	..	.

B. The Results of Customer Segmentation

Before implanting Fuzzy ART network, we need to determine the vigilance parameter (ρ) to control the degree of similarity of patterns placed on the same cluster. The optimal setting of vigilance parameter is obtained by trial and error. From Fig. 6, we found the number of clusters will converge to 10 when we decrease ρ from 0.9 to 0.5. Therefore, $\rho = 0.55$ is determined and then we can obtain 10 clusters. By the way, we set $\alpha = 0.00001$, $\beta = 1$, respectively. Table 2 summarizes the built 10 clusters and their number of containing objects.

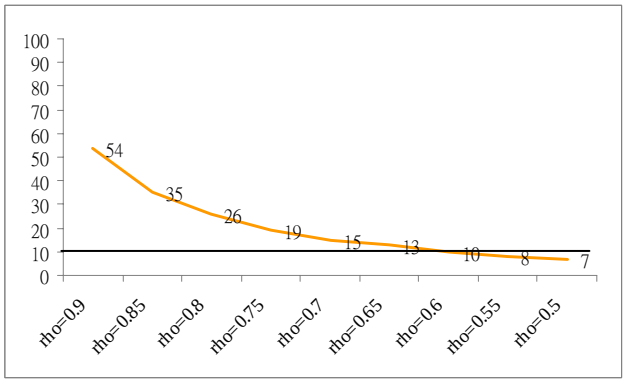

Fig. 6 The relation between the number of clusters and the defined similarity (vigilance parameter)

Table. 2 The constructed clusters of Fuzzy ART

Cluster / No. / Data size	1	2	3	4	5	6	7	8	9	10
No of objects	18	20	12	6	11	6	5	6	14	2
Cumulative (%)	18	20	50	56	67	73	78	84	25	100

Table 3 shows the keywords in each cluster. However, we can find lots of them overlaps. From Table 3, it's not easy to identify the unique characteristics of every cluster. Therefore, we introduce LSI in next step.

C. Information Retrieval

The results of LSI can be found in Table. 4. Compared with Table. 3, it's easy to understand that the extracted keywords by LSI are more unique than those of only using occurrence frequency. The number of extracted keywords is also smaller. From Table 4, companies can use these keywords to understand what customers are talking about. For example, when using cosmetics products, the important topics of customers in cluster are "Long term", "Bright", and "Whitening." It's easy to know that the factors "long-term whitening efficacy" and "they can make me look brightly" are very important for this customer segmentation. In this case, only three keywords are acquired and they are very easy to interpret. However, if we extract too many keywords, the interpretation will be a tough task. Take cluster 1 for instance, several keywords such as "Efficacy, Refreshing, Good, Feeling, Moisture, Dry, Moment, Effective…" are obtained by implementing LSI. In this condition, to organize these keywords to form a concept or a topic isn't very easy for human beings. That's also the reason why we propose association rules mining technique in next step.

Table. 3 Results of Fuzzy ART: the extracted keywords ranked by their frequency

Cluster#1		Cluster#2		Cluster#3		Cluster#4	
Efficacy	50%	Efficacy	50%	Whitening	75%	Face	67%
Feeling	39%	Whitening	40%	Efficacy	58%	Useful	50%
Good	33%	Alcohol	30%	Alcohol	50%	Whitening	50%
Whitening	33%			Heavy	33%	Efficacy	50%
Alcohol	33%			Feeling	33%	A pock	50%
						Months	33%
						Obvious	33%
Cluster#5		Cluster#6		Cluster#7		Cluster#8	
Efficacy	73%	Not bad	50%	Lotion	60%	Useful	100%
Useful	55%	Useful	50%	Obvious	60%	Good	67%
Whitening	55%	Efficacy	50%	Alcohol	60%	Alcohol	67%
Smell	36%	Long	33%	Useful	40%	Face	50%
A pock	36%	Lotion	33%	White	40%	Small	33%
		Good	33%	Acridness	40%	Lotion	33%
		Whitening	33%	Whitening	40%	Quick	33%
				Feeling	40%	Whitening	33%
				Dark	40%	Refreshing	33%
				Face	40%	Comfortable	33%
				Whitening	40%	Feeling	33%

Cluster#9		Cluster#10	
Efficacy	43%	Useful	100%
Alcohol	43%	Works	100%
Useful	36%	Whitening	100%
Heavy	36%	Requirement	100%
Feeling	36%	Moment	100%
Face	36%	A pock	100%
		Feeling	100%
		Water	50%
		Efficacy	50%
		White	50%

D. Association Rules Discovery

For the purpose of being easy to use, CBA software version 2.0 is employed in this study. A part of results of CBA are summarized in Table. 5. From this table, 4 concepts can be organized by the discovered association rules. Take cluster 1 for example, rules 1 to 4 form concept 1 which indicates long-term efficacy in whitening, but not in smell of alcohol; rules 5 to 8 form concept 2 which focus on the effectiveness of cosmetics products; rule 9 indicate the same conclusion with concept 1 (avoid alcohol smell); rules 10 to 12 form concept 4 which focus on the usefulness of cosmetics in curing whelks.

Table. 4 Extracted Keywords using LSI

Cluster#2	Cluster #1	Cluster #9	Cluster #3	Cluster #5
Water, Cold, Lots, Good, Long term, Unevenness, Useful	Efficacy, Refreshing, Good, Feeling, Moisture, Dry, Moment Effective	Works, Obvious, Habit, Strong, Dark	Smell, Friends, Oily, Condition, Long, Cheap	Mask, Worthy, Easy, Inferior to, Waste, To reduce inflammation
20%	**38%**	**52%**	**64%**	**75%**
Cluster #4	Cluster #6	Cluster #8	Cluster #7	Cluster #10
Long term, Bright, Whitening	Dry, Enough, Refreshing, Bottle	Whelk, A pock, Comfort-able, Price	Eye, Bright, Irritation, Expensive	Pricking, Brilliant
81%	**87%**	**93%**	**98%**	**100%**

Table. 5 Using the discovered association rules to form concepts: An example of cluster #1.

Item	Association rules	Conf, Sup	Concept
1	Moment → Long term	5.556%,1	#1
2	Whitening → Long term	5.556%,1	
3	Useful → Long term	5.556%,1	
4	Alcohol → Long term	5.556%,1	
5	Useful → Effective	5.556%,1	#2
6	Good → Effective	5.556%,1	
7	Whitening → Effective	5.556%,1	
8	Efficacy → Effective	5.556%,1	
9	Smell → Alcohol	5.556%,1	#3
10	To reduce inflammation → Whelk	5.556%,1	#4
11	Feeling → Whelk	5.556%,1	
12	Efficacy → Whelk	5.556%,1	

V. Conclussion

A. Figures and Tables

The main purpose of this work is to find what costumers are interested in and to provide the useful knowledge to companies for personalized or direct marketing. This study

proposed a FAIR scheme for extracting knowledge from blogs. A real case of cosmetics products is provided to demonstrate the effectiveness of our method. Experimental results indicate that the proposed FAIR scheme indeed can extract useful knowledge on blogs. In addition, this study integrates association rules mining to organize the extracted keywords into realizable concepts. These association rules can make the finding more understandable.

In addition to polysemy and synonym problems, there are lots of blogger-created netspeaks including funny words and interesting icons in blogs. It is a very tough task to define or understand them clearly. For further works, researchers should pay much attention to understand what their true meanings are.

VI. Acknowledgement

The authors would like to thank the National Science Council of Taiwan, R.O.C. for supporting this research in part under Contract No. NSC 96-2416-H-324 -003 -MY2.

References

[1] W.-Y. Ma and K.-J. Chen, "Introduction to CKIP Chinese word segmentation system for the first international Chinese word segmentation bakeoff," *The 2nd SIGHAN Workshop on Chinese Language Processing*, 2003, pp. 168-171.

[2] G.A. Carpenter, S. Grossberg, and D.B. Rosen, "Fuzzy ART: fast stable learning and categorization of analog patterns by an adaptive resonance system, " *Neural Networks*, vol. 4, 1991, pp.759-771.

[3] L. Burke and S. Kamal, "Neural networks and the part family/machine group formation problem in cellular manufacturing: a framework using fuzzy art," *Journal of Manufacturing Systems*, vol. 14, no. 3, 1995, pp.148-159.

[4] L.-S. Chen and C.-T. Su, "Using granular computing model to induce scheduling knowledge in dynamic manufacturing environments," *International Journal of Computer Integrated Manufacturing.*, vol. 21, no. 5, 2008, pp. 569-583.

[5] C.-T. Su, L.-S. Chen, and Y. Yih, "Knowledge acquisition through information granulation for imbalanced data," *Expert System with Applications*, vol. 31, no. 3, 2006, pp. 531-541.

[6] S. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas and R. A. Harshman, "Indexing by latent semantic analysis," *Journal of the Society for Information Science*, vol. 41, no. 6, 1990, pp. 391-407.

[7] M.W. Berry, S. T. Dumais, and G. W. O'Brien, "Using linear algebra for intelligent information retrieval," *SIAM Review*, vol. 37, 1995, pp. 573-595.

[8] A. G. Akritas and G. I. Malaschonok, , "Applications of singular-value decomposition (SVD)," *Mathematics and Computers in Simulation*, vol. 67, 2004, pp. 15-31.

[9] R. Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules between Sets of Items in Large Databases," *ACM SIGMOD International Conference on Management of Data*, vol. 22, no. 22, 1993, pp. 207-216.

[10] R. Agrawal and R. Srikant, "Fast Algorithm for Mining Association Rules," *20th International Conference on Very Large Data Bases*, 1994, pp. 487-499

[11] E. Cohen and B. Krishnamurthy, "A short walk in the Blogistan," *Computer Networks*, vol. 50, no. 5, 2006, pp. 615-630.

[12] B. A. Nardi, D. J. Schiano, M. Gumbrecht, L. Swartz, "Why we blog? " *Communications of the ACM*, vol. 47, no. 12, 2004, pp. 41-46.

[13] M. Chau and J. Xu, "Mining communities and their relationships in blogs: a study of online hate groups," *International Journal of Human - Computer Studies*, vol. 65, no. 1, 2007, pp. 57-70.

[14] R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, "Trawling the web for emerging cyber-communities," *Computer Networks*, vol. 31, no. 11-16, 1999, pp. 1481-1493.

[15] M. Larkin, "Blogs: new way to communicate about infectious diseases," *The Lancet Infectious Diseases*, vol. 5, no. 12, 2005, p. 748.

[16] I. Oransky, "Cancer blogs," *Lancet Oncology*, vol. 6, no. 11, 2005, pp. 838-839.

[17] K. D. Trammell, "Blog offensive: an exploratory analysis of attacks published on campaign blog posts from a political public relations perspective," *Public Relations Review*, vol. 32, no. 4, 2006, pp. 402-406.

[18] S. Todoroki, T. Konishi, S. Inoue, "Blog-based research notebook: personal informatics workbench for high-throughput experimentation," *Applied Surface Science*, vol. 252, no. 7, 2006, pp. 2640-2645.

[19] Y.-S. Lin and J.-Y. Huang, "Internet blogs as a tourism marketing medium: A case study," *Journal of Business Research*, vol. 59, 2006, pp. 1201–1205.

[20] M. Kobayashi and K. Takeda, "Information Retrieval on the Web," *ACM Computing Surveys*, vol. 32, no. 2, 2000, pp. 144-173.

[21] J. Gao and J. Zhang, "Clustered SVD strategies in latent semantic indexing," *Information Processing and Management*, vol. 41, 2005, pp. 1051–1063.

[22] A. Kontostathis and W. M. Pottenger, "A framework for understanding Latent Semantic Indexing (LSI) performance," *Information Processing and Management*, vol. 42, 2006, pp. 56–73.

[23] H. S. Du and C. Wagner, "Weblog success: Exploring the role of technology," *International Journal of Human-Computer Studies*, vol. 64, 2006, pp. 789–798.

[24] E. Asano, "A public outreach in epilepsy surgery using a serial novel on BLOG: A preliminary report," *Brain & Development*, vol. 29, 2007, pp. 102–104.

[25] F. M. Facca and P. L. Lanzi, "Mining interesting knowledge from weblogs: a survey," *Data & Knowledge Engineering*, vol. 53, 2005, pp. 225-241.

[26] X. Tai, F. Ren, and K. Kita, "An information retrieval model based on vector space method by supervised learning," *Information Processing and Management*, vol. 38, 2002, pp. 749–764.

[27] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Mogan Kaufmann Publishers, 2001.

[28] Y. Chen, F. S. Tsai, and K. L. Chan, "Blog search and mining in the business domain", *Proceedings of 2007 ACM SIGKDD Workshop on Domain Driven Data Mining (DDDM2007)*, August 12, 2007, San Jose, California, USA.

[29] Y. Chen, F. S. Tsai, and K. L. Chan, "Machine learning techniques for business blog search and mining," *Expert Systems with Applications*, vol. 35, 2008, pp.581–590.