Proceedings of the 4th WSEAS Int. Conf. on Information Security, Communications and Computers, Tenerife, Spain, December 16-18, 2005 (pp258-264)

Dynamic Monitoring Model for BBS Content Security*

Jing Yu, Yanping Zhao
School of Management & Economics
Beijing Institute of Technology
Beijing
P.R.China
paopaofish@bit.edu.cn, zhaoyp@bit.edu.cn

Abstract: - Bulletin Board Service (BBS) community should be a harmonious public speaking location full of harmless opinions. This paper analyzes the characteristics of article content in BBS, especially for Chinese Language, and proposes a novel dynamic monitoring model for BBS content security. The model is based on a fast multi-pattern matching algorithm, and uses a combined approach of Text Filtering, Social Network Analysis (SNA) and spot investigating method. The system can discover illegal article efficiently, and identify danger persons and the behavior pattern dynamically.

Key-Words: - BBS; Content security; Multi-pattern match; SNA

1 Introduction

Theory researches and practice applications about content security filtering mainly aim at ordinary WWW texts or Email. However, the existing methods and technologies are not very satisfactory in view of BBS texts. The reasons as follows: Firstly, the article content of BBS is displayed incisively and vividly, and use large numbers of shortening symbols, fabricating terms which make the traditional filtering systems cannot distinguish. Secondly, BBS has extremely quick public respond and spread speed, so that it has difficulty in monitoring BBS timely and dynamically. Finally, since BBS is comparatively easy for users register new accounts at any moment, the analysts are overwhelmed with the incomplete or misleading information from many sources, and have difficulty in discovering their relations. There is a need to help the analysts to identify key danger persons who maybe have several virtual accounts, and track and analyze relation data of them.

This paper proposes an innovative dynamic monitoring model for BBS content security, which uses a combined approach of Text Filtering, Social Network Analysis (SNA) and spot investigating method, especially efficient for Chinese natural language.

This paper is organized as follows, in section 2, some related works; section 3 proposes the framework of dynamic monitoring model for BBS content security; section 4, mainly designs two core modules, one supports Content Filter based on a novel fast Multi-Pattern matching algorithm, the other supports Link Analyzer; section 5, some initial test results; section 6, conclusions and future work.

2 Related Work

Roughly there are three kinds of content security filtering technologies. The first is key word(s) matching algorithms. The second is statistical method based on Vector Space Model (VSM). The last one is based on knowledge base. The most popular one is simple pattern (key word) matching based on Single-Pattern or Multi-Pattern. The speed of pattern matching algorithm is emphasized in order to meet the need of broadband information filtering on network. However, in the practical application of filtering system, two existing problems need to be solved. They are how to make the rules for filtering content and which kind of pattern matching algorithm is more efficient to Chinese language texts. The second filtering technology is statistical method based on VSM, which is comparatively used more in text classification research than in others. The accuracy is enhanced greatly through training a large amount

* This research is supported by the National Natural Science Foundation of China (70471064), and Research Foundation of Beijing Institute of Technology (BIT-UBF-200308G10).

of clearly classified corpus and constructing feature vector (key words) to represent each class. But this method consumes more time in computing similarity between feature vectors and that of unknown documents [1]. The two technologies above-mentioned have a common disadvantage of non-semantic. The content filtering technology based on knowledge base is becoming a research hotspot all over the world. This research is more meaningful to Chinese text filtering since it is based on character sets not words. Jin Yao-hong uses Hierarchical Network of Concepts theory (HNC)[2] and matching method based on complicated Chinese semantic knowledge rules. Its accuracy is more than 90%, but its efficiency is comparatively rather low.

We can see speed and accuracy are both important to the broadband network content filtering. So it is necessary to improve the existing filtering algorithm to adapt to Chinese text efficiently. Some experts worked on mixtured filtering method research [3][4]. Others apply SNA technology to BBS research. For example, Ville H. Tuulos and et al combine topic model with social networks for chat data topic mining [5]. Kou Zhong-bao and et al discover the "small world" phenomenon of BBS [6]. But there hasn't been research about SNA applied to BBS content security monitoring.

3 Framework of Dynamic Monitoring for BBS Content Security

3.1 BBS Data Source

The data used in this paper was captured from Tianya Club BBS, which is very popular in P.R.China (www.tianyaclub.com). The total number of registered IDs is about 3,000,000 and the number of live users is often over 30,000. So Tianya Club provides a very good data resource for our research.

In general, every BBS is more or less composed of some discussion groups (community). Each group has a series of articles organized according to different topics. The first article is initial article. Others are reply articles that comments on the initial one. Each article contains account ID, title, group information, date and time, contents. The account ID of initial article is the username of announcer, while that of reply article is the username of replier. Social network G emerges because of the reply relation among these account IDs. Each account ID who has taken part in topic discussion is a node v in G. The set of v is V. A link is established between two nodes when one account ID replies to the other. Each link e is an arc from reply ID to initial ID. The set of e is E. A social network in BBS is a small world [6].

3.2 Dynamic Monitoring Model for BBS Content Security

Fig.1 presents the framework of dynamic monitoring model for BBS content security.

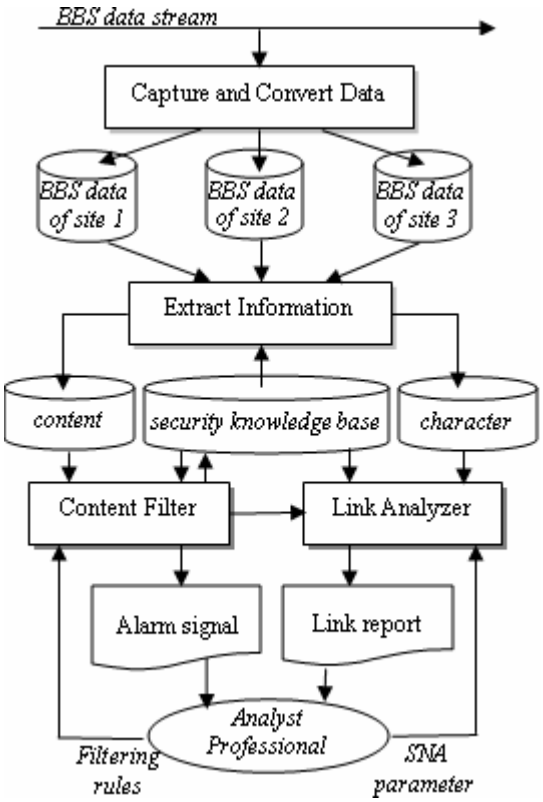

Fig. 1 Framework of dynamic monitoring model for BBS content security.

The system is aimed at monitoring terrorists or danger persons' speech among all LAN users who access the BBS server, such as Tianya Club. The whole framework includes four modules.

(1) "Capture and Convert Data" module
This module is in charge of capturing, processing and converting BBS network data on the gateway of LAN with fast data packet capture and protocol analysis technology.

Based on the characteristic of Chinese language, the "plasmodium", "synonym", "meaningless" phenomenon exists in the BBS content. We give definition 1 and explanation here and the others in the following sections.

Definition 1: "plasmodium" is a Chinese language phenomenon in BBS. There are a few of deliberate interferential characters, such as "*", written among Chinese characters by danger persons to escape the filtering system.

It is based on noticeable character of Chinese language different from English language is that one English word is usually equivalent to two or three Chinese characters. For example, the combination of character "中" and "国" produces "中国" which means "China". This kind of phenomenon does not influence readers' understanding, such as "中*国" ("中国" actually). But the negative influence of "*" is that "中*国" cannot be matched by the pattern of "中国". So the data need to be cleaned out these deliberate interferential characters after converting. The cleaned BBS data is stored as .txt files in the appointed site directories.

(2) "Extract Information" module
Useful information contained in .txt files above-mentioned is extracted and stored into database. The data is divided into two kinds of information we will deal with. One kind is "title" and "content" that are the dealing objects of Content Filter module. The other is character information, such as "account ID", "data", "time" etc. that are the dealing objects of Link Analysis module.

(3) "Content Filter" and "Link Analyzer" modules
"Content Filter" can monitor article content and identify whether any filtering rule is matched. If there is a BBS article containing illegal content, the record of this speaker is marked with different level. Then "Link Analyzer" makes a judgment that who is the author of the illegal article (danger persons) and records his characters. Along with the expansion of danger communities matched with filtering rules, "Link Analyzer" can help to analyze the status of these account IDs characters according to some SNA parameter values. Moreover, Visualization Software can help simulate community structure and discover persons' relations, community distribution, information spread channels, behavior patterns and influence patterns. The following section introduces these two core modules in details.

4 Two Core Modules

4.1 The novel fast multi-pattern matching algorithm

Speed is very important in security filtering application. A system with low efficiency cannot perform well in monitoring. Most popular fast matching algorithms are based on Multi-pattern. The most notable algorithms include Aho-Corasick (short as AC) [12] and Commentz-Walter (short as CW) [11] algorithms. A string is defined as a pattern. The advantage of Multi-pattern matching algorithm is that all the occurrences of any given strings can be found only in one process of scanning text. This kind of algorithms has two steps. Firstly, for a given set of strings, a finite state automaton (or we call it pattern tree) is constructed. And then, the text steam is scanned from end to end with this pattern tree. Actually, pattern matching is a process of state transition. Each state is represented by a number. "0" represents start state. There is another special state, output state. When an output state is received, the corresponding string is emitted. We denote a set of strings by S, and intend to detect all the occurrences of any strings in S in a text stream T. Pattern matching is that given a current state and the next input character, the machine checks to see if the character causes a failure transition. If not, then it makes a transition to the corresponding state according to the character. In case of a failure transition the machine must reconsider the character causing the failure for the next transition and the same process is repeated recursively until the given character leads to a non-failure output state. Most multi-pattern algorithms use the shift method of Single-pattern matching algorithm BM [13], for example, CW. When a pattern character is not matched by the next input character, several text characters would be skipped according to a shift value. In this way, the matching process becomes more efficient.

We made comparison to some article contents in Chinese language and in English language with CW algorithm. We discovered that the traditional key words matching algorithm is not suitable for Chinese language. The reasons as follows:

(1) From fig.2 (b), each English word is composed of 26 letters from 'a' to 'z'. And the length of each word is about 5 letters. But from fig.2 (a) ordinary Chinese characters are more than 3,000 at least to make the basic units of a word, and each word is made of about two characters. The result of this difference is that Chinese pattern tree (fig. 2(a)) is far wider than English pattern tree (fig.2 (b)). Much

matching time is consumed in matching the first character in Chinese pattern tree.

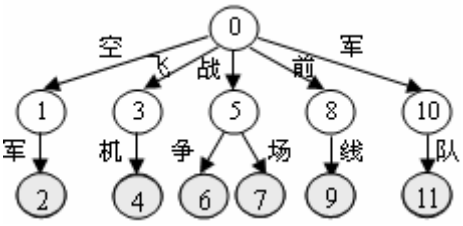

Fig.2 (a) Chinese pattern tree.

Where we use s_1 to represent string1: 空军(airman) in the left branch, s_2: 飞机(airplane), s_3: 军队(army), s_4: 战争(battle), s_5: 前线(battlefront), s_6: 战场(battleground). States: 2, 4, 6, 7, 9, 11 represent their output states.

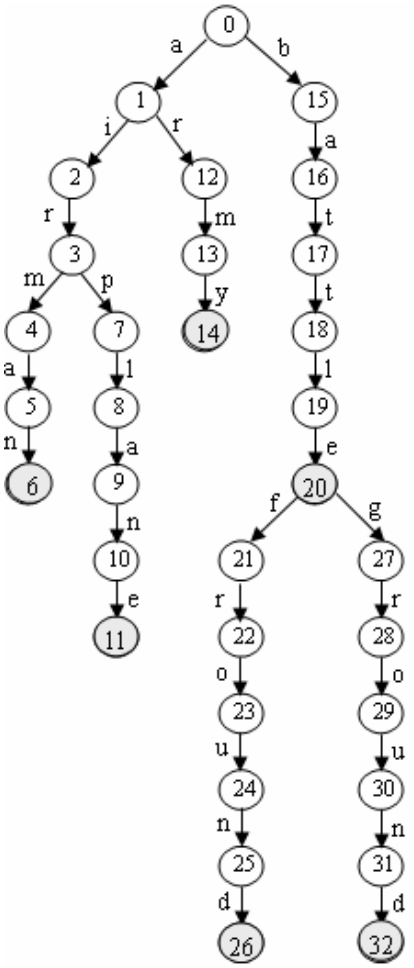

Fig.2 (b) English pattern tree.

Where we use s_1 to represent string1: airman, in the left branch, s_2: airplane, s_3: army, s_4: battle, s_5: battlefront, s_6: battleground. States: 6, 11, 14, 20, 26, 32 represent their output states.

(2) Since an English word (about 5 letters) is longer than a Chinese word (about 2 characters). It results that Chinese pattern tree (fig.2 (a)) has fewer depths, not as much deep as English pattern tree (fig.2 (b)). And since the shift value is determined by minimum pattern length, Chinese words are too short and the system cannot save running time in shift stage.

Our solution is that all Chinese characters contained in string patterns are organized according to each computed unique value, read into a Hash table. At each node of Hash table, child Hash table is set dynamically through computing the latter character of this node in one string. This kind of finite automaton is based on Hash. (See fig.3)

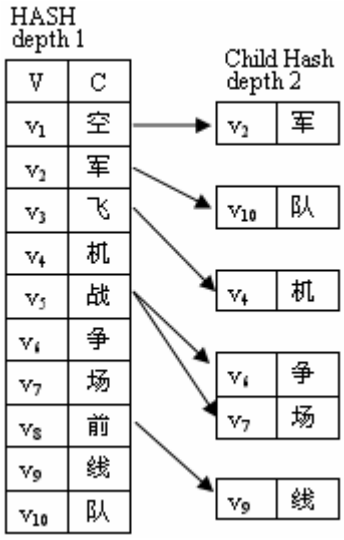

Fig.3 New string pattern tree based on fast multi-pattern matching algorithm.

If a string is composed of three characters, depth 3 child hash table can be set dynamically.

4.2 Filtering rules based on pattern

The next problem is how to make filtering rules. We use statistical method based on VSM. The model maps each document to a set of normalized terms as a vector, which is expressed as $(T_1, W_1; T_2, W_2; \ldots, T_n, W_n)$, where T_i is a feature term, W_i is the weight for T_i. The traditional TF-IDF [14]weights:

$$\omega(t_{ik}) = \frac{tf_i(t_k) \times \log(\frac{N}{N_i} + 0.05)}{\sqrt{\sum_{i=1}^{n} tf_i^2(t_k) \log^2(\frac{N}{N_i} + 0.05)}} \quad (1)$$

The filtering rules we made should have representation and discrimination. That is feature terms must represent the content of the target and can discriminate a class of documents from others. The traditional TF-IDF only can compute the weight of a term in one class, but cannot discriminate the

difference of its weight in several classes. Therefore, the modified TF-IDF weights are defined as follows:

$$\omega(t_{ik}) = \frac{tf_i(t_k) \times \log(\frac{N}{N_i} + 0.05)}{\sqrt{\sum_{i=1}^{n} tf_i^2(t_k) \log^2(\frac{N}{N_i} + 0.05)}} \times DI_{DA} \times (1 - DI_{DC}) \quad (2)$$

DI_{DA} (Distribution Information among Classes) and DI_{DC} (Distribution Information inside a Class) are two impact factors. DI_{DA} represents the distribution difference of feature terms among kinds of classes. DI_{DC} represents the distribution difference of feature terms in the same class. DI_{DA} and DI_{DC} are defined as follows:

$$DI_{DA} = \frac{\sqrt{\sum_{i=1}^{m} (tf_i(t_k) - \overline{tf(t_k)})^2 \Big/ (m-1)}}{\overline{tf(t_k)}} \quad (3)$$

$$\overline{tf(t_k)} = \frac{1}{m} \sum_{i=1}^{m} tf_i(t_k)$$

Where $tf_i\ (t_k)$ represents the frequency of feature term t_k in class i, m represents the number of classes. When a feature term t_k only appears in a class, DI_{DA}=1 and it discriminates class strongly. When the frequency of t_k in each class is equal, DI_{DA}=0 and it can hardly discriminates class.

$$DI_{DC} = \frac{\sqrt{\sum_{j=1}^{n} (tf_j(t_k) - \overline{tf'(t_k)})^2 \Big/ (n-1)}}{\overline{tf'(t_k)}} \quad (4)$$

$$\overline{tf'(t_k)} = \frac{1}{n} \sum_{j=1}^{n} tf_j(t_k)$$

Where $tf_j\ (t_k)$ represents the frequency of feature term t_k in document j, n represents the number of documents in one class. When a feature term t_k only appears in a document of a class, DI_{DA}=1 and it can hardly discriminates class. This document is likely to be an exception of this class. When the frequency of t_k in each document of a class is equal, DI_{DA}=0, it discriminates class strongly.

The system computes the weights of feature terms according to the modified TF-IDF and selects those terms with their weights bigger than a threshold. And make the selected terms as patterns in the pattern tree. This kind of pattern also can be added and modified manually. Besides, we have special treatment for synonym and meaningless words. Their definitions are as follows:

Definition 2, synonym means the meaning of one word is similar to another, or a dialect, or an irregular word to represent a meaningful word. We set one delegate word for a synonym list. The system evaluates the weight of each word in this list equal to that of delegate. For example, a delegate "弟弟妹妹" (brothers and sisters) always replaces its shortening symbols and fabricating terms "ddmm" in the list.

Definition 3, meaningless words, the meaning of these words has not enough discrimination among classes. Some of them need to be identified and picked out. Administrator can add or modify them into a stop word list.

4.3 Link Analyzer

Social network analysis theory has many meaningful parameters [7][8], which can explain some social problems in community environment. We use these parameters to discover important information from BBS character relation data. The parameters as *Degree, Closeness, Betweenness* can disclose the status of danger persons in BBS, and parameters as *Characteristic path length, Cluster coefficient* disclose the dense degree of community and information spread speed within community.

Degree represents the number of edges connected with node *v*. *Indegree* represents the number of reply from other nodes to node *v*. *Outdegree* represents the number of reply from node *v* to other node. The three parameters evaluate the difference of core status among danger account IDs. They are defined:

$$d(v) = \frac{\deg(v)}{\deg_{\max}} \quad (5)$$

$$d_{in}(v) = \frac{in\deg(v)}{in\deg_{\max}} \quad (6)$$

$$d_{out}(v) = \frac{out\deg(v)}{out\deg_{\max}} \quad (7)$$

It holds $0 \le d(v), din(v), dout(v) \le 1$, *degmax*, *indegmax* and *outdegmax* are the maximum value of *Degree*, *Indegree* and *Outdegree* respectively.

Closeness weighs the adjacency degree between node *v* and other nodes $u \in V$. It is defined:

$$cl(v) = \frac{n-1}{\Sigma_{u \in V} d(v,u)} \quad (8)$$

Where n represents the number of nodes, $d(v, u)$ represents the shortest path length from node v to node $u \in$ V. It holds $0 \le cl(v) \le 1$.

Betweenness represents the depending degree on node v from node $u \in$ V to other nodes $z \in$ V. It is defined:

$$bw(v) = \frac{1}{(n-1)(n-2)} \sum_{\substack{u,z \in V; n(u,z) \neq 0 \\ u \neq z, v \neq u, v \neq z}} \frac{n(u,z;v)}{n(u,z)} \quad (9)$$

Where $n(u, z)$ is the number of geodesics from u to z and $n(u, z; v)$ is the number of geodesics from u to z passing through v. It holds $0 \le bw(v) \le 1$.

Characteristic path length represents the average shortest path length of $n(n-1)/2$ pair of nodes. It is defined:

$$L = \frac{\sum_{v,u} d(v,u)}{n(n-1)/2} \quad (10)$$

Cluster coefficient is the average of all nodes C_v. C_v is defined:

$$C_v = \frac{|\mathrm{E}(\Gamma_v)|}{k_v(k_v-1)/2} \quad (11)$$

Where $k_v = \deg(v)$, $|\mathrm{E}(\Gamma_v)|$ represents the edges that the neighbors of node v connect each other.

We can obtain the dynamic characteristics of dangerous social network, and discover new behaviour patterns through observing and comparing the change of parameter values in different time. We use the equation sum expression as follows:

$$\Delta p = \sum_{i=1}^{m} (p_i(t_2) - p_i(t_1))^2 \quad (12)$$

Where m represents the number of parameters, t_1 and t_2 represent two different times, $p_i(t)$ represents the value of parameter i at time t. If the value of Δp is bigger than threshold, there is a notable pattern variance. The system need to monitor and analyze again.

5 Initial Testing Results

The follow figures and tables are some test results. In the figures, The 33 account IDs are filtered out because the content of their articles is matched by the filtering rules of military affairs class. They discussed around this topic. We use spot investigating methods to track some vital link clues. This method contains three steps as follows:

Step One: General Investigation

We consider all account IDs as a whole. In fig.4 there are several cliques in this community. The central key persons of cliques are obvious. ID1 and ID7 have bigger nodes and deeper color. And the articles from ID18, ID19, ID20 and ID21 cannot arouse other persons' interest. We can identify each ID character's status through computing SNA parameter values. (See tab.1)

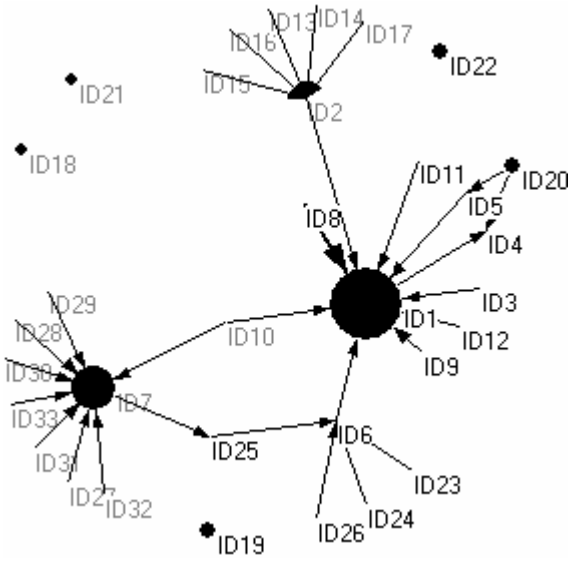

Fig.4 General investigation.

Tab. 1 Top three account IDs with maximum parameter values.

Indegree	Account	ID1	ID7	ID2
	Value	0.3125	0.2500	0.1563
Outdegree	Account	ID1	ID10	ID20
	Value	0.0625	0.0625	0.0625
Degree	Account	ID1	ID7	ID2
	Value	0.1875	0.1406	0.0938
Closeness	Account	ID1	ID10	ID6
	Value	0.4642	0.3906	0.3845
Betweenness	Account	ID7	ID1	ID25
	Value	0.0302	0.0262	0.0252

This table represents the top three key account IDs according to the maximum parameter values. We

find ID1 and ID7 have higher centrality and are considered as leaders.

Step Two: Partition Investigation

We need to pay more attention to the two leaders ID1 and ID7. The partition group we cared also includes the account IDs which have relation to either ID1 or ID7. (See fig.5)

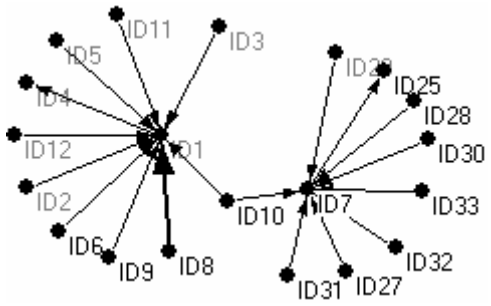

Fig.5. Partition investigation.

We can find an important role that may be ignored in step 1. He is ID10. He not only replied ID1, but also replied ID7. Although he has not high centrality as ID1 and ID7, he may be the information source.

Step Three: Individual Investigation

We should emphasis on investigating the key persons and their repliers. We can find more neighborhood information and clues.

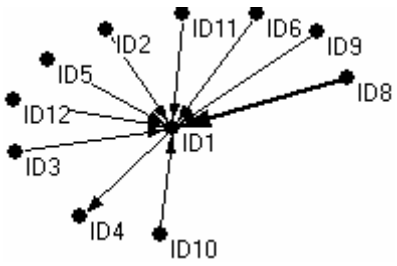

Fig.6. Individual investigation.

In fig.6, nine account IDs replied ID1 except ID4. ID1 replied ID4 unexpectedly. We suppose that another clique will come up around ID4.

6 Conclusions and Future Work

This paper introduces the BBS content security framework roughly. The dynamic monitoring model uses a combined approach of Text Filtering, Social Network Analysis (SNA) and spot investigating method. In the content filtering, a novel fast multi-pattern matching algorithm is proposed and more efficient for Chinese language especially. In order to improve the effect of content filtering, we design to add a semantic function in the content filter module based on knowledge rules. Besides, more BBS data mining work need to do with SNA theory.

References:

[1] He Jing, Liu Hai-yan, Real-time Content Filtering Based on Vector Space Model, *Computer Engineering*, Vol.30, No.15, 2004, 8, pp. 26-27.

[2] Jin Yao-hong, Design and Implementation of Semantic-based Text Filtering System, *Computer Engineering and Application*, No.17, 2003, pp. 22-25.

[3] SU Gui-yang, LI Jian-hua, MA Ying-hua, et al, Improving the precision of the keyword-matching pornographic text filtering method using a hybrid model, *Journal of Zhejiang University SCIENCE*, Vol.5, No.9, 2004, pp.1106-1113.

[4] Liu Chang-yu, Tang Chang-jie, Yu Zhong-hua, Bayes Discriminator for BBS Document Based on Latent Semantic Analysis, *Chinese Journal of Computers*, Vol.27, No.4, 2004, pp. 566-572.

[5] Ville H. Tuulos, Henry Tirri, Combining topic models and social networks for chat data mining, *Proceedings-IEEE/WIC/ACM International Conference on Web Intelligence, WI 2004*, 2004, pp. 206-213.

[6] Kou Zhong-bao, Zhang Chang-shui, Reply networks on a bulletin board system, *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, Vol.67, No.3, March, 2003, pp. 36117-1-6.

[7] Petter Holme, Characteristics of Small World Networks, *http://www.tp.umu.se/_kim/Network/holme1.pdf*, 20th, April, 2001.

[8] D. J. Watts, S. H. Strogatz, Collective Dynamics of `Small-World' Networks, *Nature,* 1998, pp. 393-440.

[9] Commentz-Walter B, A string-matching algorithm fast on the average, *Proc 6th International Colloquium on Automata Languages and Programming*, 1979, pp.118-132.

[10] Aho AV, Corasick MJ, Efficient string matching: an aid to bibliographic search, *Communications of the ACM*, No.18, 1975, pp.333 -340.

[11] Boyer RS, Moore J S, A fast string-searching algorithm, *Communications of the ACM*, No.20, 1977, pp.762 - 772.

[12] Salton,G. (Ed.), Automatic Text Processing. Addison Wesley, Massachusetts, 1989.