2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology

A Wavelet-based Model to Recognize High-quality Topics on Web Forum

You Chen, Xue-Qi Cheng, Yu-Lan Huang
Institute of Computing Technology, Chinese Academy of Sciences
{chenyou, cxq, huangyulan}@software.ict.ac.cn

Abstract

Web forum has become an important resource on the Web due to its rich information contributed by millions of Internet users every day. Meanwhile, thousands of junk or valueless messages exist in web forum. Recognizing high-quality topics should be fundamental tasks in Search Engine and Web Mining systems. However, it is not a trivial problem to quantify high-quality topics on web forum. Users face a daunting challenge in identifying a small subset of topics worthy of their attention. In this paper, we present several characteristics to measure high-quality topic, based on these characteristics, we propose a novel model to recognize high-quality topics on web forum. Our model consists of three steps. First, time series signals which contain distinctive characteristics between high-quality topics and non-high-quality topics are extracted from topics. Second, features are obtained from signals by using Wavelet Packet Transform (WPT). Third, upon the features, high-quality topics are recognized by using Back-Propagation Neural Network. Conducting experiments on Tencent Message Boards which have 2,710,994 messages and 189,962 authors ranging from Jan 1, 2005 to Nov 12, 2007, we demonstrate the efficiency of our model, showing that the average accuracy rate of high-quality topic recognition is 95% and nearly 50,000 topics can be recognized in one second.

1. Introduction

Web forum is a popular platform where thousands of people present their views on various topics. Among these topics, some are valuable while others are not.

As the scale of web forum is becoming larger and larger, it is rather a hard job for people and systems to recognize high-quality topics from these tons of information in a timely manner.

High-quality topic recognition on web forum is different from traditional topic detection, which is one of the major tasks of TDT [1]. Traditional topic detection [2] [3] [4] uses term-based algorithm to discover topics in a corpus of news archive. The corpus consists of temporally ordered news stories, which can be viewed as a stream-like structure. However, messages on web forum have a tree-like structure and relationship between messages is more complicated than that between documents in news archive.

High-quality topic recognition is different from hot topic detection, which only cares about quantity [5] [6], regardless of its quality.

Weimer et al. [7] proposed a system to find high-quality messages on web forum. Their system learned from human ratings by applying SVM classification based on features such as Surface, Lexical, Syntactic, Forum specific and similarity features. The system based on these features can achieve high accuracy. However, the target of the system is to find high-quality messages. The system only considered document content on web forum. It did not consider the relationships between these messages, and neglected the rich structure of web forum.

The target of our model is to find high-quality topics. A topic consists of many start messages and their corresponding replies. It contains link structure, and it is related to users and their post time. It is important to consider these features to find high-quality topics on web forum.

Web forum has large quantity of messages which have close relationships with their contexts. Language of messages on web forum is irregular. Key properties of messages on web forum are: low publication threshold and a lack of editorial control. The quality of messages varies drastically from excellent to abuse and spam on web forum. As the availability of such content increases, the task of identifying high-quality content on web forum becomes increasingly important.

In this paper, we present several characteristics to measure high-quality topic, based on these characteristics, a wavelet-based model is proposed to recognize high-quality topic on web forum.

In order to recognize high-quality topics from a given set of boards during a given time period, four major findings are contributed in this paper:

978-0-7695-3496-1/08 $25.00 © 2008 IEEE
DOI 10.1109/WIIAT.2008.17

(1) Characteristics which are used to measure high-quality topic are presented. There is no precise definition about high-quality topic. In this paper, we intend to provide a base description of high-quality topic.

(2) Time series signals are extracted from topics. There are several distinctive characteristics between high-quality topics and non-high-quality topics. We extracted the distinctive characteristics from signals.

(3) Features are obtained from signals by using Wavelet Packet Transform (WPT). Through WPT, a signal is decomposed into many wavelet coefficients. Coefficients have more distinctive information than the original signal. We can extract statistic features from these coefficients.

(4) High-quality topics are recognized by utilizing Back-Propagation Neural Network on features. Features are considered as input of the Neural Network, and the output of the Neural Network is a judgment which can tell you whether a topic is a high-quality one.

2. Topic on web forum

2.1. Structure of topic on web forum

Topics on web forum have a tree-like structure which is depicted in Figure 1.

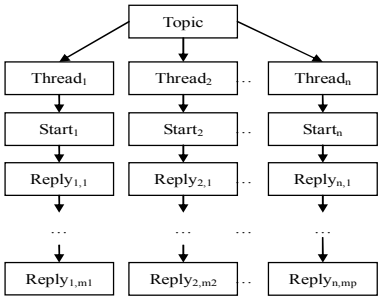

Figure 1: Structure of topic on web forum.

To have a clear comprehension of topics on web forum, some definitions for web forum terms are as follows:

(1) Message: A message is the article that an author writes on a certain subject. It can be classified as two kinds, one is start message, and the other is reply message [6].

(2) Message Board: A message board is designed for a certain domain on web forum [6].

(3) Thread: A thread consists of a start message and its reply messages [6].

(4) Topic: A topic is defined as a seminal event or activity that happens on several message boards during a given time, along with all directly related events and activities[8][9]. A topic is composed of several threads.

In Figure1, $Start_i (1 \le i \le n, n \ge 1)$ represents start message, and $\mathrm{Re}\,ply_{i,j} (1 \le i \le n, j \ge 1, n \ge 1)$ represents reply message which corresponds to $Start_i$. Each thread has a start message and related reply messages. A topic is composed of $n (n \ge 1)$ threads which may be from several message boards.

2.2. Characteristics of a high-quality topic

Till now, there is no precise definition of high-quality topic. Intuitively, there are some qualitative characteristics to judge a high-quality topic in human mind. People can tag high-quality topics manually according to these characteristics. They are as follows:

(1) Complete activity or event. At least one complete activity or event must be discussed in a high-quality topic.

(2) Quantity and uniqueness of information. A topic contains large quantitive or particular contents is worthy of attention.

(3) Clear description. Clear description of a topic can result in a deep discussion between authors. From these discussions, people can obtain a lot of useful information, and they are likely to express their views in a clear way.

(4) Valuable content. A topic with valuable content will attract numerous authors to participate in it and to contribute their valuable views on it.

3. Recognition model

3.1. Model architecture

We designed the architecture of the high-quality topic recognition model as Figure 2.

The model consists of several engines. When a user query is sent to Indexing Engine, all threads and thread IDs corresponding to the query will be extracted from web forum by Crawling Engine. The user query can be a set of message boards or a time span. The relationships between start messages and its corresponding reply messages are stored in Indexing Engine.

Indexing Engine sends the corresponding threads and thread IDs to Clustering Engine. As to these threads, Clustering Engine will group several clusters. Topics are generated from these clusters, and each topic is assigned a topic ID. A topic ID and its corresponding thread IDs are stored in Indexing Engine.

For each topic ID, Authority Engine will calculate an average authority. Signal Engine will extract three

time series signals for each topic, and the three time series signals contain important information of a topic.

Wavelet Engine is conducted on the three signals. The core of the Wavelet Engine is WPT. Features can be obtained from these coefficients by Feature Engine. We get energy feature for each coefficient, and then merge these features into feature vector.

When feature vector goes through Classify Engine, it will be classified by engine. Only high-quality topic IDs will be sent to Index Engine by Classify Engine. According to these topic IDs, Index Engine will send corresponding threads to the Query Engine, and Query Engine will exhibit the high-quality topics to users.

For the Clustering Engine, we have conducted several research works before, and there are a lot of excellent clustering algorithms to solve it [10][11][12]. It is not the key point of the model. Crawling Engine, Indexing Engine and Query Engine are general and they are easy to implement. In this paper, we focus on Signal Engine, Wavelet Engine, Feature Engine and Classify Engine.

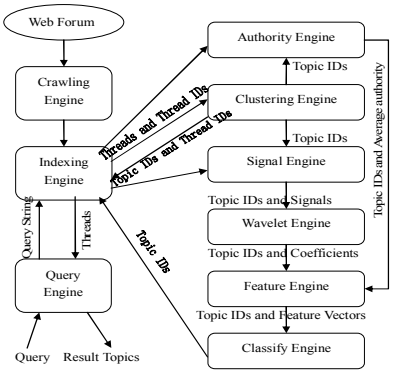

Figure 2: Architecture of the high-quality topic recognition model

3.2. Feature extraction

We have investigated many tagged high-quality topics and non-high-quality topics, and find out differences between them in terms of replies, threads, authors and message size. Generally, a high-quality topic has a lot of replies, and numerous authors participate in it. Many authors of the topic have deep discussions on views of the topic.

We divided topic life cycle into many time intervals, and conducted investigation on these intervals. There are distinctive characteristics at these time intervals in terms of replies, authors and messages size. As to a topic, first, we computed number of authors, number of replies and size of messages at each time interval. And then, we joint these time intervals to form three time series signals $S_{author}, S_{reply}, S_{size}$. S_{author} is about number of authors at each time interval, S_{reply} is about number of replies, and S_{size} is about size of messages.

High-quality topic always involves high authority authors. Average authority of a topic is also an important factor to recognize high-quality topic. In our model, features is generated from two resources: one is from wavelet coefficients which are the results of WPT on above three signals, and the other is from authority which is computed by extended HITS [13] algorithm.

3.2.1 Wavelet features

Wavelet features are obtained by using WPT on $S_{author}, S_{reply}, S_{size}$. The signals extracted from topics are not directly suitable to extract features. They tend to be complex, noisy and multi-sensory. In Figure3, there are two topics, and we present three original signals $S_{author}, S_{reply}, S_{size}$ for each topic. The two topics were classified as two classes. One class is high-quality topic, and the other is non-high-quality topic. $S_{author}, S_{reply}, S_{size}$ correspond to number of authors, number of replies and size of messages respectively.

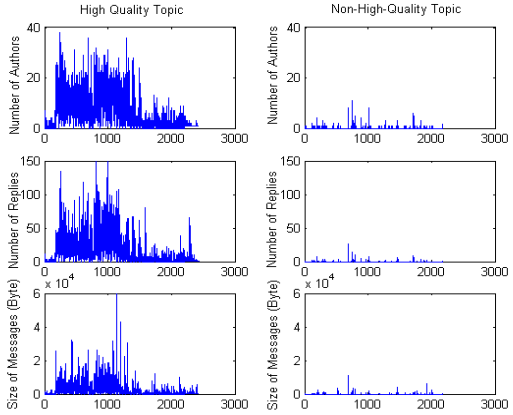

Figure 3: Original signals of high-quality topic and non-high-quality topic on number of authors, number of replies and size of messages.

The left of Figure3 is three original signals of high-quality topic, and the right is about non-high-quality topic. The x-axis units are time intervals. In Figure3, a time interval is one hour. There are 2,429 and 2,179 time intervals in high-quality topic and non-high-quality topic respectively. For each time interval, there is a value corresponding to y-axis. The value is independent from other time intervals, and it is calculated at a specific time interval.

There are many differences between high-quality topic and non-high-quality topic in terms of the three signals in Figure 3, but it is not directly suitable to extract features from these signals. It is necessary to

transform these original signals into different frequency component and then study each component with a resolution matched to its scale.

Wavelet transform is capable of providing time and frequency localizations simultaneously, while Fourier transforms could only provide frequency representations. The computation of wavelet transform can be very efficient. Fast wavelet transform only needs $O(N)$ multiplications, and its space complexity is also linear. Another important aspect of wavelet transform is their ability to reduce temporal correlation so that the correlations of wavelet coefficients are much smaller than the correlation of the corresponding temporal process.

Daubechies' wavelets [14] give remarkable results in discrete signal analysis and synthesis. Various experiments and studies have show that Daubechies' wavelets are better than other wavelet forms for dealing with general discrete signals. In this paper, we use WPT to decompose signals, and Daubechies' wavelet is selected as wavelet packet function.

WPT is an extension of Discrete Wavelet Transform and can be obtained by a generalization of the fast pyramidal algorithm. Each detail coefficient vector is decomposed into two parts using the same approach as in approximation vector splitting. The complete binary tree is produced as shown in Figure4.

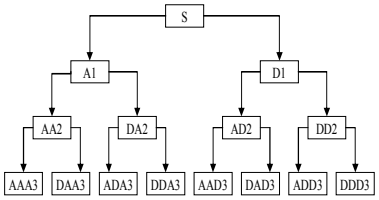

Figure 4: Wavelet packet decomposition tree, "A" presents approximation coefficients, "D" presents detail coefficients and number 1, 2, 3 presents decomposition level. After decomposition, S is assigned as $AAA3 + DAA3 + ADA3 + DDA3 + AAD3 + DAD3 + ADD3 + DDD3$

We stared with $h(n)$ and $g(n)$, the two impulsive responses of low-pass and high-pass analysis filter. Corresponding to the scaling function and the wavelet function, respectively. The sequence of function $W_n(x), n = 0,1,2,\cdots$ is defined as:

$$W_{2n}(x) = \sqrt{2}\sum_{k=0}^{2N-1} h(k) W_n(2x-k) \quad (1)$$

$$W_{2n+1}(x) = \sqrt{2}\sum_{k=0}^{2N-1} g(k) W_n(2x-k) \quad (2)$$

Where $W_0(x) = \Phi(x)$ is the scaling function and $W_1(x) = \Psi(x)$ is the wavelet function. In other words, the tree indexed family of analyzing functions can be reached by:

$$W_{j,n,k}(x) = 2^{-j/2} W_n(2^{-j}x - k) \quad (3)$$

k can be interpreted as a time-localization parameter, j as a scale parameter and n as an oscillation parameters. $W_{j,n,k}$ analyzes the fluctuations of the signal roughly around the position $2^j k$, at the scale 2^j and at various frequencies for the different admissible values of the last parameter n [15].

"db4" has 4-vanishing moment. The intuition of vanishing moments of wavelets is the oscillatory nature which can be thought to the characterization of differences or details between a datum with the data in its neighborhood. With higher vanishing moments, if data can be represented by low-degree polynomials, their wavelet coefficients are equal to zero.

In this paper, we use "db4" as wavelet packet function to transform signals. There are two stages of transform, one is decomposition, and the other is reconstruction. First, original signals are decomposed into three levels. After wavelet packet decomposition, we will get eight wavelet packet coefficients on level three. The eight coefficients are $X_{30}, X_{31}, X_{32}, X_{33}, X_{34}, X_{35}, X_{36}, X_{37}$ which are ordered from low frequency to high frequency. Second, each coefficient will be reconstructed, $X_{30}, X_{31}, X_{32}, X_{33}, X_{34}, X_{35}, X_{36}, X_{37}$ are reconstructed as $S_{30}, S_{31}, S_{32}, S_{33}, S_{34}, S_{35}, S_{36}, S_{37}$ i.e. X_{30} will be reconstructed as S_{30}, and X_{31} will be reconstructed as S_{31}. Through decomposition and reconstruction, original signal S can be represented as:

$$S = S_{30} + S_{31} + S_{32} + S_{33} + S_{34} + S_{35} + S_{36} + S_{37} \quad (4)$$

For each reconstructed coefficient $S_{3i}(i = 0,1,2,\cdots,7)$, its corresponding energy $E_{3j}(j = 0,1,2,\cdots,7)$ is calculated as

$$E_{3j} = \sum_{k=1}^{n} | x_{jk} |^2 \quad (5)$$

Where $x_{jk} = S_{3j}[k], (j = 0,1,2,\cdots,7, k = 1,2,\cdots,n)$, n is the frequency of signal and $S_{3j}[k]$ is the amplitude of S_{3j} at k^{th} point. Normalization has been done on energy $E_{3j}(j = 0,1,2,\cdots,7)$, and then feature vector F_{sub} is formed as

$$F_{sub} = [f_1, f_2, f_3, \cdots, f_8] = [\frac{E_{30}}{E}, \frac{E_{31}}{E}, \frac{E_{32}}{E}, \cdots, \frac{E_{37}}{E}] \quad (6)$$

Where $E = (\sum_{j=0}^{7} | E_{3j} |^2)^{1/2}$

For each topic, three signals $S_{author}, S_{reply}, S_{size}$ are extracted. We use wavelet packet to transform the three signals, and then extracted three feature vectors corresponding to them. We merge these three sub feature vectors into a big feature vector as:

$$F_{wavelet} = F_{sub1} \cup F_{sub2} \cup F_{sub3} \quad (7)$$

Where F_{sub1} corresponds to S_{author}, F_{sub2} corresponds to S_{reply} and F_{sub3} corresponds to S_{size}.

3.2.2 Authority features

A topic involving many high authority authors is always very influential so that many people will follow it. We define a high authority author as one that has posted many substantial messages that led to a few of deep discussion. Thus we can use the relationship between messages. This reply relationship is like the link relationship between web pages. So an extended HITS algorithm [13] is selected to analyze author authority. We calculated authors' authority scores recursively as follow:

$$a(i) = \sum_{j \to i} w(i,j) * h(j), \quad h(j) = \sum_{j \to i} w(i,j) * a(i) \tag{8}$$

Where $a(i), h(i)$ are author $i's$ authority and hub score respectively, and $w(i,j)$ is the weight of the reply relationship of poster j to poster i. For each topic, we computed its average authority score as follow

$$Authority\ Score_{average} = \frac{\sum_{i \in author_{topic}} a(i)}{Number_{author}} \tag{9}$$

Where $Number_{author}$ is the number of authors whose messages are relevant to its topic domain. $author_{topic}$ is the set of relevant authors in the topic.

We add feature of average authority score into $F_{wavelet}$ to form the last feature vector F_{last} which will be used as the input of the classifier. F_{last} is as follow:

$$F_{last} = F_{wavelet} \bigcup Authority\ Score_{average} \tag{10}$$

High-quality topics are recognized by utilizing BPNN [16] on feature vector F_{last}. BPNN is a classic classifier proved to be efficient in many areas. In this paper, we will use one layer of hidden nodes in this paper to classify high-quality topics. First, we use training dataset of topics which are represented by F_{last} and the tagged classes to train a network. And then, when a topic which is represented by F_{last} goes through the BPNN network, a specific class will be given by the output node.

3.3. Overall flow of the model

The detail of the model is depicted in Figure5, which is a key part of Figure2.

High-quality topic recognition model is mainly composed of two stages: clustering and classification.

First, topics are generated by using clustering algorithm on several message boards in a given time span. Second, time series signals and average author authority are obtained from the topics. Signals are decomposed into wavelet packet coefficients by WPT. According to these wavelet coefficients and average authority, we generated a feature vector F_{last}.

Based on the feature vector, a classifier BPNN can be trained.

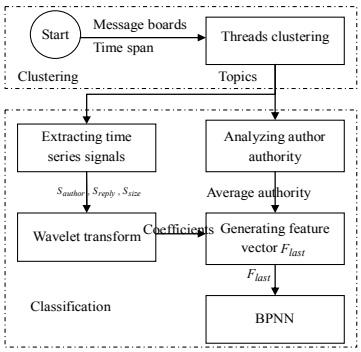

Figure 5: Overall flow of the high-quality topic recognition model

4. Experiments and performance analysis

In this section, we conduct several experiments to illustrate the performances of our model. Experiments consist of three stages. First, a complete description about our experimental dataset will be given. Second, we will extract feature vectors from topics using WPT and HITS algorithm. Third, we will evaluate the impact of our model on classification speed and accuracy. To evaluate the impact of our model on classification accuracy, three standard metrics are used [17]:

Accuracy: The percentage of correctly classified instances over the total number of instances.

Precision: the number of class members classified correctly over the total number of instances classified as class members.

Recall: the number of class members classified correctly over the total number of class members.

The experiments were run on a Windows machine having configurations Dual-Core AMD Opteron(tm) Processor 2214HE, 2.21GHz, 8.0GB RAM.

4.1. Experimental data

Our Experimental data is provided by Tencent, Inc, and the data is from Tencent web forum [18] which is a primary public web forum in the world. The web forum has numerous members (nearly 230 million) and visitors, and it covers various fields.

To prepare the labeled training and test data, we used k-means algorithm to group 112 topics from the experimental data. And then, we selected one hundred topics which are verified by assessors from the clustering results.

We employed 200 people to judge quality of the 100 topics. For each topic, they decided whether it is a high-quality one or not. If people think a topic is a high-quality one, they can vote for it. We choose topics which gained more than 120 votes as high-quality ones. Among these 100 topics, there are 20 high-quality topics and 80 non-high-quality topics.

Statistics of the experimental data are depicted in Table 1. Its contents range form Jan 1, 2005 to Nov 12, 2007, and it has 162,781 threads, 2,548,213messages and 189,962 authors. The size of the messages is 713,373,492Bytes.

Table 1: Statistics of Experimental Data

Name	Value
Number of Message Boards	29
Start Time (yyyy/mm/dd)	2005-01-01
End Time (yyyy/mm/dd)	2007-11-12
Number of Threads	162,781
Number of Reply Messages	2,548,213
Number of Authors	189,962
Size of Messages(Byte)	713,373,492

4.2. Experiments set-up

As to the one hundred labeled topics, we used WPT and HITS algorithm to transform them into feature vectors. On one hand, average authority of each topic is calculated by HITS algorithm; on the other hand, three hundred time series signals were mined from one hundred labeled topics. In these signals, one hundred are about number of authors, another one hundred are about replies and the remaining one hundred are about message size.

The signals were transformed into features by using WPT. With the transformed feature vectors, we constructed our experimental data set which is composed of feature vectors and classes. There are two classes in our experiments, one is high-quality topic, and the other is non-high-quality topic.

4.3. Feature vector extraction

A topic can be represented by three time series signals S_{author}, S_{reply}, S_{size}, which are depicted in Figure3.

In Figure3, there are six signals for two topics. We conducted decomposition and reconstruction on these six signals by using WPT. We select 'db4' as wavelet packet function and a signal is transformed to a complete binary tree which has three levels. The results of decomposition and reconstruction are depicted in Figure6, Figure7 and Figure8. For each signal, there are eight coefficients (3,0),(3,1),⋯,(3,7) at level three of wavelet packet tree.

In Figure6, the left is about high-quality topic, and the right is about non-high-quality topic.

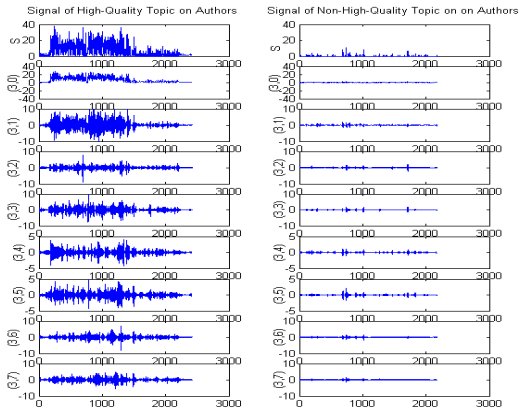

Figure 6: S_{author} signals and their corresponding reconstructed coefficients at level three of wavelet packet tree. The signals are about number of authors involved in a topic during several time intervals. The first row is the original signals of topics. There are eight coefficients for each original signal. The positions of coefficients are from second row to last row.

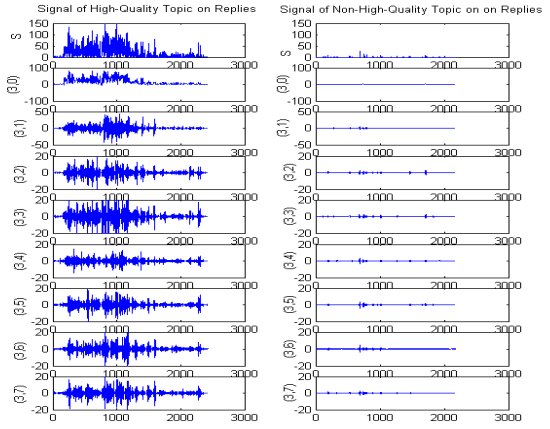

Figure 7: S_{reply} signals and their corresponding reconstructed coefficients at level three of wavelet packet tree.

From Figure6, we can see that in a high-quality topic, there are numerous authors, and they have active discussions. Discussions involved in a high-quality topic always continue for a long time. All these merits can result in a high-quality topic in a real life. However, in non-high-quality topics, few people attend, and they seldom conduct an active discuss.

Even they promote an intense discussion; it will not last for a long time, as you see in Figure6. These phenomena can be also found in Figure7 and Figure8.

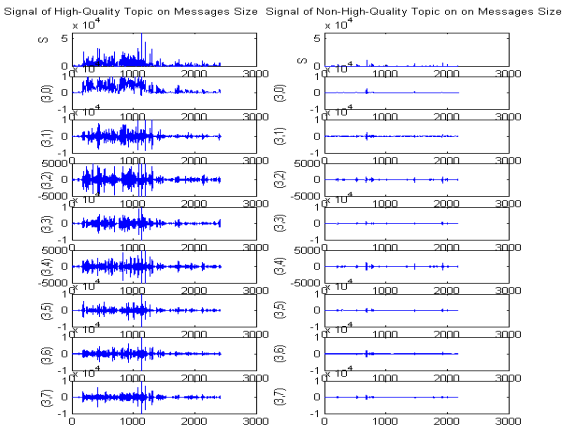

Figure 8: S_{size} signals and their corresponding reconstructed coefficients at level three of wavelet packet tree.

We extracted energy features from the reconstructed coefficients, and formed three feature vectors $F_{sub1}, F_{sub2}, F_{sub3}$ for each topic. F_{sub1} corresponds to the energy features of S_{author}, F_{sub2} corresponds to S_{reply} and F_{sub3} corresponds to S_{size}. The results of the extracted energy feature vectors are showed in Figure9, Figure10 and Figure11.

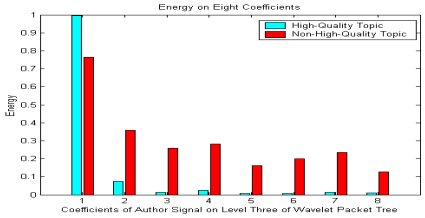

Figure 9: Energy feature vector F_{sub1}, it is about energies of eight coefficients at level three of wavelet packet tree. The energy is extracted from the coefficients of authors signal.

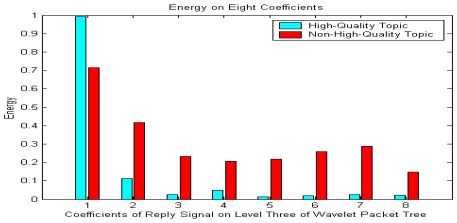

Figure 10: Energy feature vector F_{sub2}, it is about energies of eight coefficients at level three of wavelet packet tree. The energy is extracted from the coefficients of replies signal.

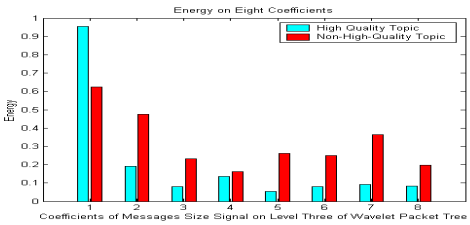

Figure 11: Energy feature vector F_{sub3}, it is about energies of eight coefficients at level three of wavelet packet tree. The energy is extracted from the coefficients of size signal.

In Figure9, Figure10 and Figure11, the energies of eight coefficients (3,0),(3,1),…,(3,7) at level three of wavelet packet tree are defined as $[\frac{E_{30}}{E}, \frac{E_{31}}{E}, \frac{E_{32}}{E}, \ldots, \frac{E_{37}}{E}]$ which can be found in section 3.2.

From the feature vector figures, we can see that the first coefficient's energy of high-quality topic is larger than that of non-high-quality topic. However, the remaining coefficients' energies of high-quality topic are smaller than those of non-high-quality topic. As to a high-quality topic, the energy of its first coefficient is close to value "1", which is a total value of eight coefficients' energies. The remaining seven coefficients' energies are very small. But to non-high-quality topic, the first coefficient's energy occupies nearly 60%~70% of the total energy. From the second to eighth coefficient, the energy of non-high-quality topic is larger than that of high-quality topic. It is obviously to identify high-quality topic apart from non-high-quality topic using energy features. Upon these features and average authority of a topic which was computed by HITS algorithm, we formed a feature vector to conduct an evaluation of classification in next section.

4.4. Evaluation of high-quality topic classification

Our experimental data consists of one hundred samples, twenty are high-quality topics, and the remaining eighty are non-high-quality topics. Each sample is composed of twenty-six features; twenty-four are energy features, one for average authority, and the last are class feature. There are two classes in our data, one is high-quality topic, and the other is non-high-quality topic.

In order to examine the classification accuracy, recall and precision within the experimental data, we randomly divided the experimental data set into three equal partitions and performed three-fold cross validation. For each fold, we ran experiments to measure the classification accuracy, recall and precision of our model.

The comparisons of accuracy, recall and precision are showed in Figure12. As in the figure, fold1 has high precision (average 98%) and low recall (average 68%). it suggests that few non-high-quality topics are classified as high-quality topics. However, there still some high-quality topics are not detected. Average recall rate (nearly 86%) of three-fold cross validation at different learning rate of BPNN was depicted in Figure13. The highest recall rate in Figure13 is nearly 94%, which is obtained by selecting learning rate as 0.04. When we use 0.04 as learning rate of BPNN, we can obtain high performances on accuracy (95%), recall (94%) and precision (96%). Fold3 has high recall (average 100%) and low precision (average 81.3%), it demonstrates that all high-quality topics in test data are classified correctly, but some non-high-quality topics are also classified as high-quality topic.

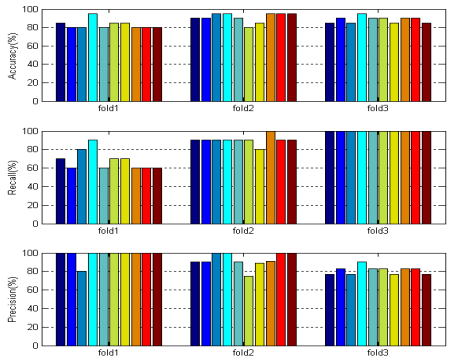

Figure 12: The accuracy, recall and precision changes according to the different fold and the learning rate of BPNN. There are ten learning rates (0.01,0.02,⋯,0.10)

Average accuracy, recall and precision of each fold can only partly reflect performances of our model, actual performances of our model can be reflected by the average classification accuracy, recall and precision of the three-fold cross validation. In Figure13, we can see that the overall performances of our model is superior, and the highest average classification accuracy, recall and precision of the three-fold cross validation are 95%, 94% and 96% respectively. They can be obtained by selecting 0.04 as learning rate of BPNN. It shows that our model can recognize nearly all high-quality topics from experimental data, and there are few non-high-quality topics classified as high-quality topics. It also demonstrates that the extracted energy features are able to identify high-quality topic and non-high-quality topic.

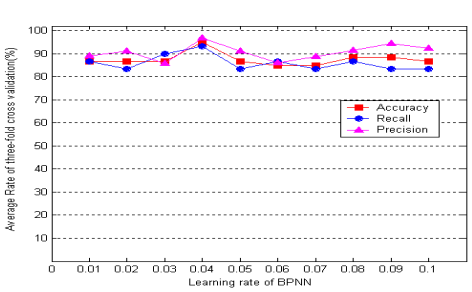

Figure 13: The average classification accuracy, recall and precision changes of the 3-fold cross validation, according to the learning rate of BPNN.

The time spent on our model has two parts, one is time of features extraction, and the other is time of classification. Our experimental data whose size is 730M can be represented by feature vectors whose size is 30K in less than one minute in the model. It is a quick process for feature vectors extraction. And then, classier can be built on these feature vectors. The classifier built on these feature vectors is lightweight and has a quick test process, nearly 50,000 classifications per second in our experiment.

5. Conclusions and future directions

Unlike most of the previous works, we aim to recognize high-quality topics on web forum. It is a practical problem in web search and web mining. A topic on web forum has a tree-like structure, which is different from a stream-like structure in traditional topic detection. Messages on web forum are irregular and the relationships between them are complicated. It is a challenge to retrieve high-quality topics from tons of information in a timely manner.

We propose a novel model for recognizing high-quality topics on web forum. It was done in four steps: First, topics were clustered from web forum. Second, time series signals were extracted from those topics. Third, feature vectors were extracted by using WPT and HITS algorithm. Fourth, high-quality topics were classified by BPNN using extracted feature vectors. We had conducted several experiments on evaluation of our model, and the experiments demonstrated superior performances of classification accuracy and speed.

Extracted feature vector is a key factor to influence the performances of classifier. It was verified that feature vector extracted by wavelet transform and HITS algorithm had led to a high efficiency and accuracy classifier. In this paper, we only extract energy features from wavelet packet coefficients. However, many other features for example, entropy existing in the coefficients can also be extracted to recognize high-quality topic. There are several

redundant features in the extracted feature vector; feature selection can be conducted to remove them. We plan to improve the performances of our model at these aspects.

Although the characteristics of high-quality topic mentioned in this paper can be considered as a basic expression of high-quality topic, we should perfect the characteristics in future. We also plan to improve the performance of our model in real system.

6. Acknowledgments

This study is supported by the 863 National Basic Research Programs of China under Grant No.2007AA01Z438 & 2006AA01Z452 and the Tencent (Shenzhen) Inc Fund project named topic detection and tracking though messages. We thank the anonymous reviewers for valuable suggestions.

7. References

[1] TDT. The 2003 topic detection and tracking (TDT2003) task definition and evaluation. http://www.nist.gov/speech/tests/tdt/tdt2003/index.htm

[2] Frederick Walls, Hubert Jin, Sreenivasa Sista, Richard Schwartz. Topic Detection in Broadcast News, 1999.

[3] Khoo Khyou Bun, Mitsuru Ishizuka. Topic Extraction from News Archive Using TF*PDF Algorithm. Proceedings of the 3rd International Conference on Web Information Systems Engineering, 2002.

[4] Bingjun Sun, Prasenjit Mitra, C. Lee Giles, John Yen. Topic segmentation with shared topic detection and alignment of multiple documents. Proc. ACM SIGIR, 2007.

[5] Kuan-Yu Chen, Luesak Luesukprasert, and Seng-cho T. Chou. Hot Topic Extraction Based on Timeline Analysis and Multidimensional Sentence Modeling. IEEE Transactions on Knowledge and Data Engineering, Vol 19, No8, 2007.

[6] Lan You, Yongping Du. BBS Based Hot Topic Retrieval Using Back-Propagation Neural Network. International Joint Conference on Natural Language Processing, 2007.

[7] Markus Weimer, Iryna Gurevych. Predicting the Perceived Quality of Web Forum Posts. In: Proceedings of the Conference on Recent Advances in Natural Language Processing, 2007.

[8] TDT: Annotation Manual Version 1.2, http://www.nist.gov/speech/tests/tdt/, Aug. 2004.

[9] The 2004 Topic Detection and Tracking (TDT '04) Task Definition and Evaluation Plan, http://www.nist.gov/speech/tests/tdt/, 2004.

[10] Rui Xu, Donald Wunsch II. Survey of Clustering Algorithms. IEEE Transactions On Neural Networks, Vol. 16, No. 3, 2005

[11] Benjamin Rosenfeld, Ronen Feldman. Clustering for Unsupervised Relation Identification. Proc. ACM CIKM, 2007.

[12] H. Zeng, Q. He, Z. Chen, W. Ma, and J. Ma. Learning to cluster Web search results. Proc. ACM SIGIR, 2004.

[13] Kleinberg, J. Authoritative Sources in a Hyperlinked Environment. Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms. ACM Press, 1998.

[14] Ingrid Daubechies. Ten Lectures on Wavelets. CBMS-NSF Regional Conference Series in Applied Mathematics, 1992.

[15] M. Chendeb, M. Khalil and J. Duchêne. Wavelet Based Method for Detection: Application in Proprioceptive Rehabilitation, in Proc 26th International Conf IEEE EMBS, 2004.

[16] Rumelhart, D. E., Hinton, G. E., &Williams, R. J. Learning internal representation by back propagating errors. Nature, 332, 533-536, 1986

[17] Nigel Williams, Sebastian Zander, Grenville Armitage. A Preliminary Performance Comparison of Five Machine Learning Algorithms for Practical IP Traffic Flow Classification. Proc. ACM SIGCOMM, vol.135, iss.5, 2006

[18] Tencent web forum. http://bbs.qq.com/