Applications of Text Clustering Based on Semantic Body for Chinese Spam Filtering

Qiu-yu ZHANG
School of Computer and Communication, Lanzhou University of Technology, Lanzhou, China
Technology & Research Center of Gansu Manufacturing Informatization Engineering, Lanzhou, China
Email: zhangqy66@gmail.com

Peng WANG and Hui-juan YANG
School of Computer and Communication, Lanzhou University of Technology, Lanzhou, China
Email: {wpff66, Rongshu0926}@163.com

Abstract—**The effect of spam filtering method based on statistics is not good enough in filtering the new-type spam with synonymous substitution and camouflage, because the method based on statistics ignores the semantic relation between words in the text, and only judges from the word itself. So, a method of spam filtering based on the semantic body is proposed in this paper. The method adopts lexical chain based on HowNet and TFIDF method based on statistics to extract e-mail features, and handle spam with text clustering method. The result of the experiment shows that the new method proposed in this pager provides a good effect in filtering new-type spam.**

Index Terms—**semantic body, lexical chain, semantic similarity, text clustering, spam filter**

I. Introduction

With the rapid development of Internet application in China, e-mail has become a communicational tool, which plays an increasingly important role in our daily work and life, especially in recent years the rapid development of Chinese e-commerce and the mobile Internet have promoted the increase of text-based enterprise mail and phone mail numbers. At the same time, spam has also developed faster with richer diversity and more forms. China's Anti-Spam Survey Report in the fourth quarter of 2010 showed that the average number of spam received by Chinese Internet users per week was 13.5, covering 38.2% of the e-mails received every week [1].

At present, content-based anti-spam technology mainly adopts the keyword-based and semantically irrelevant spam filtering methods, such as Bayesian, case-based method, text classification method and so on [2-4]. And text clustering [5-6] is one of the processing methods. Traditional text clustering-based methods of spam processing are based on word frequency. These statistical methods ignore the semantic relations between words. However, the new-type spam disguises as normal mails by using synonyms and near-synonyms, so the traditional methods hardly distinguish spam and normal mails.

Therefore, this paper proposes a spam filtering method based on semantic body. Firstly, this method introduces lexical chain based on HowNet [7] into the spam feature processing. Then the text clustering method is used to obtain more accurate clustering result. So this method is a good solution to the problem of synonyms and near-synonyms.

II. Feature Extraction of E-mails

The basic unit of Chinese e-mails is word. This is usually referred to as feature item or feature. The feature item has the following characteristics: it is able to express text content clearly, to distinguish target files and other documents, and to be separated easily. So feature extraction is very important. Traditional text clustering method gets feature items by using statistic-based feature assessment. However, these methods cannot solve problems of synonyms and near-synonyms in new-type spam, which causes email filtering inaccuracy. So, many researchers combine semantic and statistical method. On the basis of the weight of words calculated by traditional TFIDF method, literature [8] adjusts the weight of words and relevant words in synonym set, and those words are combining weighted according to their similarity. However, the synonyms in this literature come from WordNet, and word similarity comes from HowNet. There are big differences between the two resources, so the combination of them is still questionable.

To solve the problem above, the lexical chain based on HowNet and TFIDF method will be combined in this paper to realize the feature extraction.

A. Lexical Chain Processing

A lexical chain is mainly used to solve the problems of synonyms and near-synonyms. The lexical chain [9] is firstly proposed by Hirst in 1991. It's a set composed of a series of semantically related or similar vocabularies. These vocabularies polymerize together on a topic. A lexical chain and text structure have a corresponding relation, which provides important clues about text structure and theme. This paper will construct lexical chains of noun, verb and adjective sets which have passed participle and stop word processing. Specific

doi:10.4304/jcp.7.11.2612-2616

procedure is as follows:

(1) Take the word sets W_1, W_2, W_3… W_t (t is the number of words in the word set) candidate word sets, and take W_1 to construct the initial lexical chain L;

(2) If the semantic similarity between W_1 and W_2 is greater than regulation threshold s, insert W_2 into L, otherwise insert W_2 into a new chain;

(3) Repeat process (2), until all candidate words complete the process of calculation.

Experiment shows, when threshold “s” is 0.7, two words are similar.

Completing the above steps, each mail will construct many lexical chain sets $\{L_1, L_2, L_3 \ldots L_n\}$.

B. TFIDF Method

This paper adopts the improved TF (Term- Frequency) & IDF (Inverse Document Frequency) [10-11] method to calculate the weight of each feature word, and reorders feature words according to feature weights.

TF * IDF general formula is:

$$P(m_{ij}) = tf_i \times \lg(\frac{N}{N_j} + 0.5) \qquad (1)$$

In this formula, N is the number of all mails, N_j is the number of mails containing word W_j, and tf_i is the frequency of m_{ij} in text D_i.

Words similarity reflects the extent of the interchangeability of the two words in different context without changing syntax semantic structure. Experiment proves when the threshold of words similarity is 0.7, the two words can be interchanged. For example: "ZhongYangChuLiQi" and "DianNao", "ShiChang" and "ChaoShi", "TianRan" and "YeSheng", can be interchanged. So when calculating TFIDF, the words in a lexical chain will be regarded as one word, and the word with the highest frequency will be set as the standard of the lexical chain. tf_i of this standard is the sum for all words' frequency in one lexical chain in in D_i. N_j is the number of mails which contain any words of a lexical chain.

After the processing, each of the lexical chain will be instead of one word. A key lexical chain can be get, $L_i=\{(W_{i1}, P(W_{i1})), (W_{i2}, P(W_{i2})), (W_{i3}, P(W_{i3})),\ldots\}$, in which every word is in descending order according to its $P(W_{ij})$ and only N words are the mail features. The word sets in the paper is called Semantic Body.

Definition 1: Semantic Body is the word set, after the processing using HowNet and statistical method, considering semantic relations between words in mails, which can reflect a mail's content features.

III. The Text Clustering Algorithm Based on the HowNet

Text clustering typically has three steps:

(1) Text description, namely text feature extraction or selection;

(2) Text similarity measurement method definition;

(3) Text clustering. Text similarity computing is an important process of the text clustering. General text similarity calculation is measured by vector cosine, but this paper calculates using text similarity based on the HowNet semantic similarity [12].

A. Text Mail Similarity Measurement

In the text clustering, the similarity reflects the credible degree of different texts which are divided into different categories [13]. The calculation of text similarity takes the word as a unit. As Literature [14] points out, if two articles have at least 2-3 pairs of related words, these two articles basically have the same theme. Based on this idea, the text similarity algorithm can be calculated as follows:

Set text i and text j as $D_i = \{W_{i1}, W_{i2}, W_{i3}, \ldots, W_{in}\}$ and $D_j = \{W_{j1}, W_{j2}, W_{j3}, \ldots, W_{jm}\}$:

(1) Take the W_{i1} in D_i;

(2) Calculate the similarity between W_{i1} and the words in the text D_j, and choose the word with the largest similarity as its matching word, and remove it from D_j;

(3) If the similarity between the best matching words is greater than 0.8, it means the two words are very similar, then adds 1 to the similar word pairs;

(4) Repeat step (1) to step (3), until finish the scanning of the words in D_i or D_j;

(5) If the number of similar word pairs exceeds the specified value, it shows two mails are similar, and they belong to the same class. Otherwise, they belong to different class.

After the above calculation, the similarity of two mails could be obtained.

B. Class Feature Extraction

When the texts whose similarity reaches the threshold are in the same class, the class feature needs to be extracted.

Set a certain category of mail text as $C_i = \{D_1, D_2, D_3, \ldots D_n\}$. The method for the feature extraction is as follows:

(1) Extract all the words in C_i, and compose a collection, then figure out the frequency of each word, represented as $D^* = \{(W_1, P_1), (W_2, P_2), (W_3, P_3)\ldots\}$. P_i is the word frequency of W_i in all the feature words collection.

(2) Calculate the similarity between every two words in the collection, and find out the words whose similarity is greater than 0.8 to form a lexical chain. Select the word with the largest frequency in the lexical chain as lexical chain standard, and modify its frequency as the sum of all the word frequency in lexical chain, then add it to the subset d_i of the class.

(3) Put all the words in d_i in descending order according to the frequency, and then take only M words as the feature set of the class.

After the processing, mail text set of a class then obtains its features, namely, semantic body of a class.

C. The Similarity between Mail Text and Classes

After feature extraction, we get semantic body. Then, the semantic body and class features will be do text similarity computation. If the result is greater than 0.8, it is considered to belong to this category. If the result is less than 0.8, it is considered that this mail belongs to a new category. The way of this paper depends on the similarity of HowNet.

D. Text Clustering Algorithm

On the basis of the semantic distance clustering algorithm, we designed the text clustering algorithm based on the similarity calculation of semantic features using the previous mail similarity and similarity between mail and the class.

Similarity computation algorithm requires the following four properties [15]:

(1) Reflexivity: words, sentences, are similar to themselves.

(2) Monotonicity: similarity should increase or decrease continuously.

(3) Symmetry: If *A* is similar to *B*, *B* is also similar to *A*.

(4) Transitivity: If *A* is similar to *B*, *B* is similar to *C*, and then *A* is also similar to *C*.

In this clustering algorithm, the text is the unit of similarity calculation, and the similarity is based on the text similarity of the HowNet. Obviously, the similarity between the content of a mail and itself is 1, which meets the reflexivity; the range of the text similarity is [0, 1]. The similarity monotone increases or decreases, so it is consistent with monotonicity. The mail *A* is similar to *B*, and the mail *B* is similar to *A*, they have the same similarity, so it is consistent with the symmetry. Under a certain similarity threshold, the mail *A* and mail *B* are similar, the mail *B* and mail *C* are similar. So the mail *A* and mail *C* are similar. It's consistent with transitivity. Besides, the sets which meet the reflexivity, symmetry and transitivity are disjoint sets. These sets have a faster clustering.

The steps of mail clustering:

(1) After the feature processing of section II, set *N* texts:

$$\begin{aligned} D_1 &= \{W_{11}, W_{12}, W_{13}, ...\} \\ D_2 &= \{W_{21}, W_{22}, W_{23}, ...\} \\ D_3 &= \{W_{31}, W_{32}, W_{33}, ...\} \\ &... \\ D_n &= \{W_{n1}, W_{n2}, W_{n3}, ...\} \end{aligned} \tag{2}$$

On the initial conditions, each text exists as a separate class.

(2) Have a text similarity calculation between the Semantic Body of D_i and D_j, and then put the texts who meet the measurement conditions into a class.

(3) Have a feature extraction for the classes which meet the conditions, and then get the semantic body of this class;

(4) Go back to (2) and have a re-iteration. If the number of classes does not change after two iterations, then stop the iteration, the clustering is completed.

The merger of classes is similar to the merger between classes and texts.

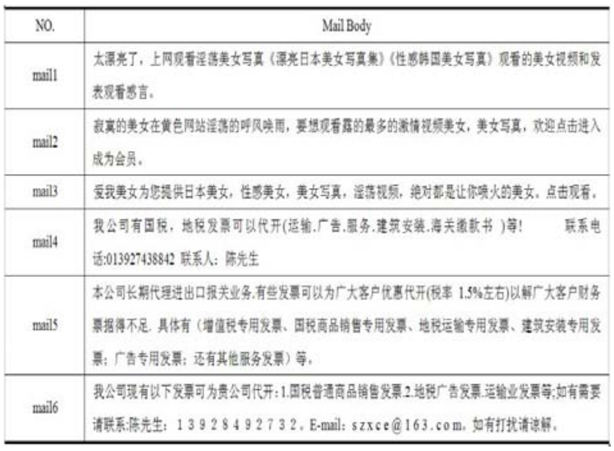
IV. Experiment Results and Analysis

A. Simple Example Demonstration

In this paper, a simple test is used to illustrate spam filtering based on semantic body. As the number of selected emails is small, TFIDF cannot get good results. The number of feature words selected from semantic body and pairs of similar words obtained in computing the text similarity is relatively small. In Figure1, this paper presents three pornographic mails and three agent bill business mails.

Words segmentation is done to the 6 mails in Figure 1 on the basis of ICTCLAS. ICTCLAS is Chinese Lexical Analysis System based on multilayer HMM model, which is developed by the Institute of Computing Technology.

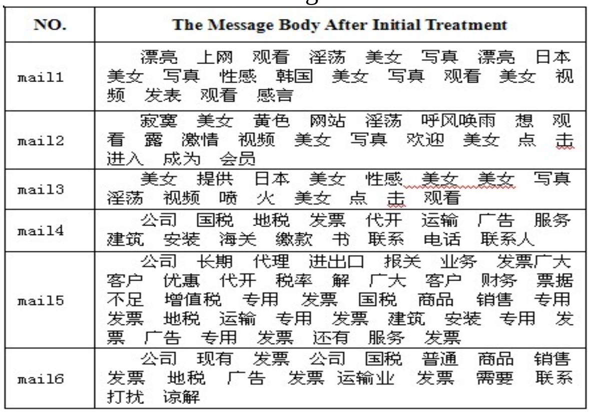
The system has the Chinese word segmentation, POS tagging and other functions. Then, the paper removes stop words and keeps nouns, verbs, adjectives.



NO.	Mail Body
mail1	太漂亮了，上网观看淫荡美女写真《漂亮日本美女写真集》《性感韩国美女写真》观看的美女视频和发表观看感言。
mail2	寂寞的美女在黄色网站淫荡的呼风唤雨，要想观看露的最多的激情视频美女，美女写真，欢迎点击进入成为会员。
mail3	爱我美女为您提供日本美女，性感美女，美女写真，淫荡视频，绝对都是让你喷火的美女，点击观看。
mail4	我公司有国税，地税发票可以代开(运输.广告.服务.建筑安装.海关缴款书)等! 联系电话:013927438842 联系人：陈先生
mail5	本公司长期代理进出口报关业务.有些发票可以为广大客户优惠代开(税率 1.5%左右)以解广大客户财务票据得不足. 具体有（增值税专用发票、国税商品销售专用发票、地税运输专用发票、建筑安装专用发票；广告专用发票；还有其他服务发票）等。
mail6	我公司现有以下发票可为贵公司代开：1.国税普通商品销售发票2.地税广告发票.运输业发票等;如有需要请联系:陈先生：１３９２８４９２７３２。E-mail：ｓｚｘｃｅ@１６３.ｃｏｍ。如有打扰请谅解。

Figure 1. The original mail body

The results are shown in Figure 2:

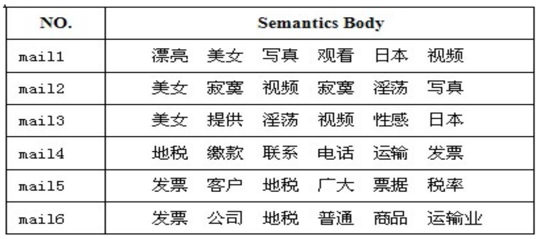


NO.	The Message Body After Initial Treatment
mail1	漂亮 上网 观看 淫荡 美女 写真 漂亮 日本 美女 写真 性感 韩国 美女 写真 观看 美女 视频 发表 观看 感言
mail2	寂寞 美女 黄色 网站 淫荡 呼风唤雨 想 观看 露 激情 视频 美女 写真 欢迎 美女 点 击 进入 成为 会员
mail3	美女 提供 日本 美女 性感 美女 美女 写真 淫荡 视频 喷 火 美女 点 击 观看
mail4	公司 国税 地税 发票 代开 运输 广告 服务 建筑 安装 海关 缴款 书 联系 电话 联系人
mail5	公司 长期 代理 进出口 报关 业务 发票广大 客户 优惠 代开 税率 解 广大 客户 财务 票据 不足 增值税 专用 发票 国税 商品 销售 专用 发票 地税 运输 专用 发票 建筑 安装 专用 发票 广告 专用 发票 还有 服务 发票
mail6	公司 现有 发票 公司 国税 普通 商品 销售 发票 地税 广告 发票 运输业 发票 需要 联系 打扰 谅解

Figure 2. The message body after segmentation, removing stop words

By using lexical chains, synonyms and near-synonyms are merged. For example, the similarity value between "HanGuo" and "RiBen" is 1 , the same as "ZengZhiShui" and "DiShui", so these words can form lexical chains, and take the word with the largest value of TF as lexical chain representative.

After removing synonyms and near-synonyms through lexical chain, key words of each mail selected by using TFIDF compose the mail's semantics body. The results are shown in Figure 3.



NO.	Semantics Body
mail1	漂亮 美女 写真 观看 日本 视频
mail2	美女 寂寞 视频 寂寞 淫荡 写真
mail3	美女 提供 淫荡 视频 性感 日本
mail4	地税 缴款 联系 电话 运输 发票
mail5	发票 客户 地税 广大 票据 税率
mail6	发票 公司 地税 普通 商品 运输业

Figure 3. Semantic Body of email

Finally, the result of calculation by using the text clustering method based on the HowNet is shown in Table I.

TABLE I
TEXT CLUSTERING

Pornographic	mail1	mail2	mail3
Service proxy class notes	mail4	mail5	mail6

B. *Application Demo*

In this paper, training set is composed of 600 spams which are from Chinese_rules.cf. 100 e-mails are collected as processing samples.

Finally, the experiment results are compared with results of Bayesian and SVM. Experiment platform is IBM ThinkPad SL400 with Intel Core 2 Duo T567 1.8GHZ, 2G DDR2 memory. In this experiment, 3 parameters are used for comparison:

Recall:

$$R = \frac{N_A}{N_S} \qquad (3)$$

Recall is the rate of spam detection. This indicator reflects the ability of spam filtering system in detecting spam. The higher the recall is, the less the slipping spam are.

Precision:

$$P = \frac{N_A}{N_A + N_B} \qquad (4)$$

Precision is the rate of Spam identification. Precision reflects the ability of spam filtering system in identifying spams accurately. The higher the precision is, the smaller the possibility of legitimate mail misidentification is.

F Value:

$$F = \frac{2PR}{R + P} \qquad (5)$$

Actually, F value is the harmonic mean of recall and precision. It integrates recall and precision into one indicator.

Among formulas, N_A is the number of the spam filtered out, N_S means the actual number of spam, N_B means the number of the normal mails misidentified as spam. Experiments lead to the conclusions shown in Table II.

It can be seen from Figure 4 and Figure 5 that this method has improvements in all aspects.

TABLE II
THE RESULT COMPARISON

Methods **Spam**	**Recall (%)**	**Precision (%)**	**F (%)**
Bayesian	66.3	90.1	76.4
SVM	81.7	88.4	84.9
This Method	85.5	90.7	88.2

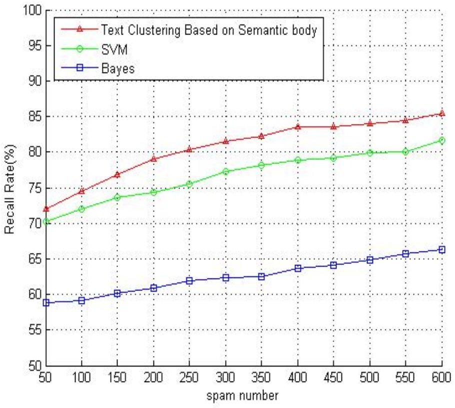


Figure 4. Comparison of recall under three methods

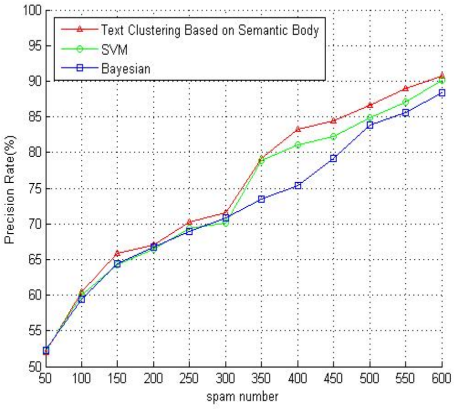


Figure 5. Comparison of precision under three methods

The experiments prove that although the proposed method in dealing with synonyms and near-synonyms achieves good results, there are still some restrictions due to HowNet. For example, the similarity between "GuoShui" and "DiShui" in the mail4 should be great, but the HowNet does not contain "GuoShui". So in the processing, it can only be merged artificially.

There is room for improvement for the text clustering method in this paper. Therefore, the research in next step will focus on how to improve the accuracy of clustering, which is significant to the improvement of the filtering effect.

ACKNOWLEDGEMENTS

This work is supported by the Master Instructor Research Fund of Education Department of Gansu Province (0703-07).

REFERENCES

[1] "Investigation Report of China Anti-Spam in the fourth quarter of 2010," anti-Spam.cn/pdf/2010_04 _report. PDF. September 2011.

[2] ZHANG Yan-qiu, and WANG Wei, "E-mail classification by SVM optimized with genetic algorithm, "Journal *of Computer Applications*, vol.29, pp.2755-3757, 2009.

[3] ZHANG Qiu-yu, SUN Jing-tao, and YAN Xiao-wen, "Research of Spam Filtering System Based on Latent Semantic Analysis and MD5," J*ournal of University of Electronic Science and Technology of China*, vol.36, pp. 1223-1227, 2007.

[4] W. Haiyan, Z. Runsheng, and W. Yi, "An anti-Spam filtering system based on the naïve Bayesian classifier and distributed checksum clearinghouse," *Proceedings of the Third International Symposium on Intelligent Information Technology Application*. IEEE, pp.128-131, 2009.

[5] LANG Jia-yun, and HU Xue-gang, "Clustering-Based Email Filtering Method with Hazy Category," *Computer System & Application*, vol.19, pp. 147-150, 2010.

[6] R. Kathleen McKeown, Regina Barzilay, David Evans, Vasileios Hatzivassiloglou, Judith L. Klavans, Ani Nenkova, and Cal Sable, Barry, et al, "Tracking and summarizing news on a daily basis with Columbia's News blaster," *Proceedings of the second international conference on Human Language Technology Research*, pp. 280-285, 2002.

[7] LIU Ming, WANG Xiao-long, and LIU Yuan-chao, "Research of Key-Phrase Extraction Based on Lexical Chain", *Chinese Journal of Computers*, vol.33, pp.1264-1254, 2010.

[8] XU Jian-min, LIU Qing-jiang, FU Ting-ting and DAI xu, "Improved Feature Selection Method Based on Similarity of Synonymous," *Journal of Hebei University (Natural Science Edition)*, vol.30, pp. 97-101, 2010.

[9] E. Gonenc, and C.Ilyas, "Using lexical chains for keyword extraction," *Information Processing and Management*, vol.43, pp.1705-1714, 2007.

[10] SHI Cong-ying, XU Chao-jun, and YANG Xiao-jiang, "Study of TFIDF algorithm," *Journal of Computer Application*, vol.29, pp.167-170, 2009.

[11] G. SALTON, and T.Y. CLEMENT, "On the construction of effective vocabularies for information retrieval," *Proceedings of the 1973 Meeting on Programming Languages and Information Retrieval*. ACM New York, pp.48-60, 1973.

[12] LIU Qun, and LI Su-jian, "Word Similarity Computing Based on How-net," *Third Chinese Lexical Semantics Workshop, Taibei. Chinese*, 2002.

[13] XU jun-ning, "Research on Document Clustering Based on Semantic Similarity of HowNet", *Xian University*, 2010.

[14] LIN li, "Text Clustering Research Based on Semantic Distance," *Xiamen University*, April 2007.

[15] DAI Wenhua, "Text Classification and Clustering based on Genetic Algorithm," *Science Press*, August 2008.

Qiu-yu ZHANG: Professor and master tutor. Director of software engineering center, vice dean of technology & research center of Gansu manufacturing informatization engineering, director of "software engineering" characteristic research direction and academic group of Lanzhou University of Technology. His research interests include: image processing and pattern recognition, multimedia information processing, information security, software engineering etc.

Peng WANG: Graduate student. He was born in TaiYuan Shanxi province in 1986. His research interests are in Chinese text classification, Anti-Spam etc.

Hui-juan YANG: Graduate student. She was born in Yulin Shanxi province in 1984. Her study should be Chinese text classification, Anti-Spam etc.