多载体数据流中的特定信息识别研究*

郑德权, 胡 熠, 于 浩, 赵铁军, 王青松

(哈尔滨工业大学 计算机科学与技术学院,黑龙江 哈尔滨 150001)

Research of Specific Information Recognition in Multi-Carrier Data Streams

ZHENG De-Quan, HU Yi, YU Hao, ZHAO Tie-Jun, WANG Qing-Song

(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

Corresponding author: Phn: 86-451-6416225 ext 604, Fax: 86-451-6412449, E-mail: zdq@mtlab.hit.edu.cn

http://www.hit.edu.cn

Received 2002-06-24; Accepted 2003-03-25

Zheng DQ, Hu Y, Yu H, Zhao TJ, Wang QS. Research of specific information recognition in multi-carrier data streams. *Journal of Software*, 2003,14(9):1538~1543.

http://www.jos.org.cn/1000-9825/14/1538.htm

Abstract: A method is presented to identify some pieces of specific information in multi-carrier data streams by feature words and based on PinYin matching. An effective knowledge approximation method is used to judge the relation between feature words and context by statistics theory. The part of speech transfer-value as system knowledge can be obtained by inductive learning of training corpus. When data streams are evaluated, the evaluation value can be gained according to the system knowledge by matching all feature words and based on their PinYin, which examines the comparability with context regular of part of speech between all feature words in data streams and themselves in training corpus. Further more, if the evaluation value exceeds the threshold, the data streams will be shielded. Experimental results show that the effect of the experiment system based on this method is efficient for identifying ill information and monitoring & controlling their spreading by multi-carrier data streams.

Key words: information identification; knowledge approximation; part of speech transition; inductive learning

摘 要: 提出了一种识别多载体数据流中包含的特定信息的新方法.该方法按照特征词及其拼音匹配规则,基于统计自然语言理论,通过自动的归纳学习,将从语料库中获得的词性间的转移值作为系统知识,利用有效的知识逼近策略判断真实数据流中的特征词与其上下文的关系,并得到特征词在真实文本中的评测值,以此来考查真实数据流中出现的全部特征词与在语料中所学到的特征词上下文搭配规则上的相似程度.如果整个数据流的评测值超过阈

* Supported by the National High-Tech Research and Development Plan of China under Grant No.2001AA114101 (国家高技术研究发展计划(863))

ZHENG De-Quan was born in 1968. He is a Ph.D. candidate. His research interests are natural language understanding, machine translation. **HU Yi** was born in 1978. He is a master student. His research interests are machine translation, natural language processing. **YU Hao** was born in 1971. He received his Ph.D. degree in 1998 and is an associate professor. His current research areas are natural language understanding, network information processing. **ZHAO Tie-Jun** was born in 1962. He is a professor and doctoral supervisor. His current research areas are computational linguistics, machine translation. **WANG Qing-Song** was born in 1973. He is a master student. His research interests are natural language processing.

值,该数据流将被屏蔽.实验结果表明,根据该方法开发的识别及监控多载体数据流中不良信息的实验系统取得很好的效果.

关键词: 信息识别;知识逼近;词性转移;归纳学习
中图法分类号: TP18 文献标识码: A

With the rapid development of network and communication technology, available information is becoming abundant in multi-carrier data streams, e.g. Web pages, short message, e-mail etc. However, while they offer tremendous convenience and help for people, a large amount of ill information, e.g. reactionary, eroticism, rubbish information,is spread as one likes to influence people's normal work and life seriously, even some rumors, illegal information threatens national security, destroy money market, disturbance of the peace because of its speed and scope. For example, the stock prices of a Japanese company fell sharply to 75 yen from 110 yen because of the influence of rumors spread in network by enemy, the company incurred substantial losses. Some pieces of eroticism information spread by multi-carrier poisoned many young boys and girls, and so on. Therefore, some ill data streams must be monitored and controlled in multi-carrier data streams.

1 Background

Information filtering (IF) technique/information identification method was used to select or reject incoming dynamic data stream in network. Traditional method searched user requests from a number of out-of-order data streams and reviewed the comparability between data streams and user profile according to a fixed and specific requests put forward by user. For the more, this data stream will be decided to accept or filter.

At present, IF or information identification methods take effect for monitoring and controlling specific information or filtering data streams which are based on mechanical matching of keyword, VSM(vector space model)[1,2], concept[3], machine learning[6] and semantic information[4], etc. However, the effect of identification and filtering is not good, which is due to restrict of natural language processing (NLP) themselves. In this paper, a new method is presented, which identifies specific contents spread by multi-carrier data streams and an experiment system is developed for identifying ill data streams.

2 Designment Approach

2.1 Design ideas summary

The reason that the effect of monitoring and filtering is not good by existing IF system is because these methods look on feature words appearing in data streams as isolated and they did not identified on the context environment of chapters or sentences. On the other hand, IF system could not judge whether some pieces of information are normal or ill in data streams, even if they could match feature words.

The basic design ideas are based on statistics model in this paper. By analyzing a number of specific information in multi-carrier data streams nowadays, this paper presents a new method to identify some pieces of specific information by feature words for monitoring and controlling their spread, which uses an effective knowledge approximation to learn the relation between feature words and context. The part of speech (POS) transfer-value as system knowledge can be obtained by inductive learning of training corpus. When data streams are evaluated, the evaluation value can be gained according to the system knowledge by matching feature words, which examines the comparability with context regular of POS between the feature words in data streams and themselves in training corpus. If the evaluation value exceeds the threshold, the data streams will be shielded.

2.2 Process flowchart

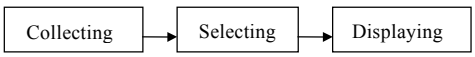

Fig.1 Traditional model for IF

Figure 1 is a traditional model[5] for IF or Information identification, it was described as this, collecting data in advance → selecting useful Information → displaying some results, Fig.2 is processing process for identifying specific information designed in this paper.

2.3 Several important definitions

There are some definitions about data streams as follows.

Definition 1. **Characteristic files**. This paper respectively extracted typical feature words from containing specific information and normal information in multi-carrier data streams to built positive characteristic file and negative characteristic file and, every positive feature word in positive characteristic file need gain evaluation value if it appears in multi-carrier data streams. Feature words were lined in a hash table and indexed according to pronunciation of first Chinese character for quickening speed of machine learning and processing. Feature words that the pronunciation of first Chinese character is alike were organized in a chain table.

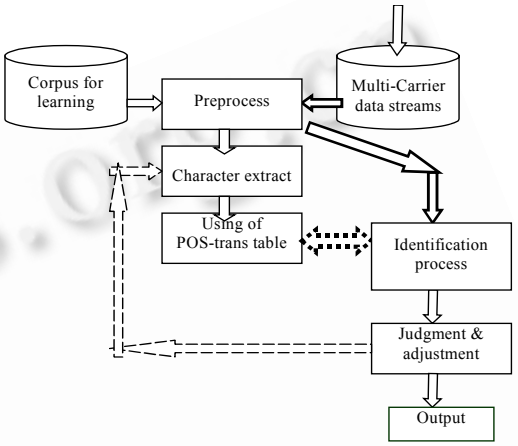

Fig.2 Flow char of processing specific information

Definition 2. **POS transfer-value**. We defined the POS of feature words was *KeyWord* signed in positive character file. We centered on feature words when data streams is processed, and from right to left, defined POS transfer-label as $POS_{l1},POS_{l2},\ldots$, and from left to right, defined POS transfer-label as $POS_{r}1,POS_{r2},\ldots$. The POS transfer-figure is shown as Fig.3.

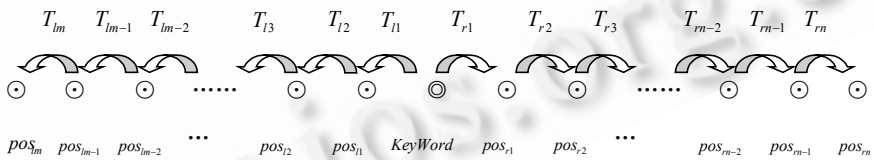

Fig.3 POS transfer-figure

According to lingual intuition, we suppose the POS transfer-value T_i is less if it is farther from *KeyWord* and we think its relation with feature words is loose. The feature word will be regarded as processing center to define two base B_l and B_r that satisfy formula 1 for right and left of feature word, they can be calculated easily, further, the transfer-value of every couple co-occurrence POS can be calculated, which is given in formula 2.

$$B_k+\frac{1}{2}B_k+\left(\frac{1}{2}\right)^2 B_k+\left(\frac{1}{2}\right)^3 B_k+\ldots+\left(\frac{1}{2}\right)^{m-1} B_k=1\ (k=l,r). \tag{1}$$

$$T_{li}=\left(\frac{1}{2}\right)^{i-1} B_l\ \ (i=1,2,\ldots,m);\quad T_{rj}=\left(\frac{1}{2}\right)^{j-1} B_r\ \ (j=1,2,\ldots,n). \tag{2}$$

Definition 3. **Average transfer-value of POS**. The average transfer-value $\overline{T_k}$ of a pair of POS (POS_i,POS_j) is a ratio between gaining the sum of all POS transfer-value and appearing all times when learned training corpus.

Definition 4. **POS transfer-form**. Every feature word is defined a POS transfer-form in positive character file when all training corpora were learned. POS transfer-form is defined as shown in expression (1).

$$K\_Info(POS_{i1},POS_{i2},\overline{T_{k1}})(POS_{i1},POS_{i2},\overline{T_{k2}})\ldots/(POS_{j1},POS_{j2},\overline{T_{t1}})(POS_{i1},POS_{i2},\overline{T_{t2}}). \quad (1)$$

K_Info denotes feature word, POS_{ij} denotes POS and $\overline{T_{ki}}$ denotes average transfer-value of POS. There are n+3 (n=52) POS in this paper and denoted with sequence number 0~54. Sequence numbers of POS 0~51 are defined in Machine Translation System 2000 (MTS2000) by machine translation lab of Harbin Institute of Technology. Sequence number 52 denotes POS of positive feature words, sequence number 53 denotes POS of negative feature words and sequence number 54 denotes unknown POS.

Expression (1) may be described as follows, for feature words *K_Info*, the average transfer-value from POS POS_{i1} to POS_{i2} is $\overline{T_{k1}}$, it is $\overline{T_{t1}}$ from POS_{j1} to POS_{j2}, and so on. The every couple POS and their average transfer-value in two sides of sign "/" shown left POS transfer-form and right POS transfer-form of feature words *K_info*.

Definition 5. **Evaluation report-table.** This report-table's data structure is similar to characteristic file's, it is a hash table indexed on filename of data stream and it only lists filename, evaluation value of real data streams got in identification engine. Feature words and threshold value will be adjusted. Data stream file can be browsed and edited by indexing the report-table.

3 Implementation Process

3.1 Preprocess

(1) Extract every sentence that contains positive feature words from multi-carrier data streams and, transform their format to text-format;

(2) Tag positive and negative feature words first, then, segment Chinese words and tag other POS;

(3) Remove empty word (e.g. preposition, conj, auxiliary word) and retain notional word for quickening system processing speed.

3.2 Building POS transfer-form

Regard a sentence as a processing unit and find its positive feature words *KeyWord*. Calculate POS average transfer-value T_i of every adjacent word in the same sentence. Calculation starting point is the positive feature word tagged in Preprocess module. Every feature word will gain it's POS transfer-form as follow:

(1) Calculate POS transfer-value T_i of different adjacent POS from right to left and write down their co-occurrence times S_i, then, so does from left to right;

(2) For feature word *KeyWord*, append POS_i, POS_j and POS transfer-value T_i to POS transfer-form if them do not exist, otherwise, the new transfer-value T_i will be added up with quondam T_i;

(3) While training corpus have been learned, calculate average transfer-value $\overline{T_k}$ of every couple POS by that every T_i divided every S_i, then, replace T_i with $\overline{T_k}$ in POS transfer-form and regard $\overline{T_k}$ as the average transfer-value of every couple POS from POS_i to POS_j;

(4) Write down the occurrence times of every positive feature word.

3.3 Real data streams evaluation

(1) Match positive feature words in a sentence;

(2) Regard positive feature words as center, fetch average transfer-value $\overline{T_k}$ of every couple POS from corresponding POS transfer-form *K_info* in both left direction and right direction to add up them that regard as an evaluation value of the sentence, at the same time, add up their co-occurrence times. If positive feature word is first or last POS, the left or right evaluation value is 0;

(3) If a sentence contains negative feature word, the evaluation value of the sentence will reduce an experience -value;

(4) The sum that adds up evaluation value of all sentences divided by the co-occurrence times of all couple POS that contain positive feature words is the evaluation value of real data stream. If the evaluation value of real data stream equals or exceeds threshold value α, the multi-carrier data streams will be confirm to contains some pieces of specific information.

3.4 Adaptive adjustment for POS transfer-form and threshold value

Processing real data streams in multi-carrier, feature words, POS transfer-form and threshold value need be adjusted constantly in order to make sure their veracity and objectivity, because the data streams are dynamic and changed continuously. However, only when some pieces of specific information are accumulated to certain quantity, need they be adjusted for raising the efficiency[7], because the proportion of specific information is lower than normal in multi-carrier data streams.

(1) Select new multi-carrier data streams files that affirm to contain specific information;

(2) Count the occurrence times of every positive feature word appearing in positive characteristic file to equal to the occurrence times last time. These data streams files will be regard as new corpus to build a new POS transfer-form;

(3) Add up all average transfer-value of POS in new POS transfer-form and quondam respectively, new POS transfer-form will replace quondam;

(4) Replace threshold value α with α^* got by formula 3.

$$a' = a \times \frac{sum\ of\ all\ average\ transfer\text{-}value\ in\ new\ POS\ transfer\text{-}form}{sum\ of\ all\ average\ transfer\text{-}value\ in\ quondam\ POS\ transfer\text{-}form} \tag{3}$$

4 Experiment Result and Analysis

We employed 1 200 files including positive example (700) and negative example (500) to do open test. Positive examples contain some pieces of ill information, negative do not contain those, even if do, these contents are quoted naturally when they are reported or some critical viewpoints are contained among multi-carrier data streams. Using our experiment system, the recall ratio is beyond 90% and the precision ratio is beyond 80%.

The experiment result showed our approach is promising in comparison with other methods. It gives an integrative consideration to identify some pieces of specific information that appear in changed forms and relevant contexts. However, the research in this paper still stands in surface of natural language processing, we will meet more challenge when face more open environment because of opening of Chinese themselves, and that, we only combined adjacent POS, assigning weight to every position in POS chain is more mechanical and did not introduced relation information between Chinese words. On the other hand, Chinese words segmentation and POS tagging need solve some difficult about ambiguous words still.

5 Future Work

Identification engine is necessary for monitoring and controlling some pieces of specific information spread in Web pages, short message, E-mail, etc. Therefore, we need finish some works ad follows.

(1) We should perfect our training corpus and feature words continuously.

(2) Apply some techniques based on NLU in more exact identification shift monitoring or identification for one type of specific information to another type.

(3) Identification techniques based on natural language semantic processing way should be studied deeply because natural language is complex, is changed continuously and can be expressed by different mode.

In conclusion, Using NLU techniques will help to veracity and objectivity of monitoring and controlling specific information. This is one of key point of intelligent Information security, is emphasis and difficulty to actualize intelligent network information processing for the future.

References:

[1] Richard Hunter J. Performance considerations for information filtering systems using database technology [Ph.D. Thesis]. Florida Institute of Technology, 1998.

[2] Huang XJ, Xia YJ, Wu LD. A text filtering system based on vector space model. In: Cao YQ, ed. Proceedings of the Conference of the 20th Anniversary of CIPSC. Beijing: Tsinghua University Press, 2001. 215~218 (in Chinese with English abstract).

[3] Ali H A. Concept based retrieval and information filtering [Ph.D. Thesis]. University of Nebraska-Lincoln, 2001.

[4] Niu Wei-Xia, Zhang Yong-Kui. Latent semantic indexing is applied in information filtering. Computer Engineering and Application, 2001,37(9):57~62 (in Chinese with English abstract).

[5] Hanani U, Shapira B, Shoval P. Information filtering: Overview of issues, research and systems. User Modeling and User-Adapted Interaction, 2001,11(3):203~259.

[6] Zhang BT, Seo YW. Personalized web-document filtering using reinforcement learning. Applied Artificial Intelligence, 2001,15(7):665~685.

[7] Wu LD. Large-Scale Chinese Text Processing. Shanghai: Fudan University Press, 1997 (in Chinese).

附中文参考文献:

[2] 黄萱菁,夏迎炬,吴立德.基于向量空间模型的文本过滤系统.见:辉煌 20 年——中国中文信息学会 20 周年学术会议论文集.2001.215~218.

[4] 牛伟霞,张永奎.潜在语义索引方法在信息过滤中的应用.计算机工程与应用,2001,37(9):57~62.

[7] 吴立德.大规模中文文本处理.上海:复旦大学出版社,1997.