

適合線上資源收集的中文語料自動儲存分類技術

陳俊良², 李明哲^{1,2}, 黃乾綱², *簡立峰¹

1. 台北南港中央研究院資訊科學研究所

2. 台灣大學資訊工程學研究所

*E-mail: lfchien@iis.sinica.edu.tw

摘要

隨著互聯網資源的大量增加, 適合收集利用線上資源的中文語料自動儲存分類技術日漸重要. 為此本文將提出一以 PAT-tree 為基礎自動中文語料分類技術. PAT-tree 提供適合互聯網環境的工作結構(Working Structure), 這個結構考慮中文特性以及語料庫的語言模型應用, 所發展的分類特徵以及語言模型都可以直接利用 PAT-Tree 索引.

Chinese Corpus Storage and Classification from Online Resources

Chun-Liang Chen², Min-Jer Lee^{1,2},
Chien-Kang Huang², Lee-Feng Chien¹

1. Institute of Information Science, Academia Sinica, NangKang, Taipei

2. Dept. of CS and IE, National Taiwan University, Taipei

Abstract As unlimited volume of corpus resources is now available over the Internet, techniques which can be easily adapted with the change of online resources for automatic corpus collection and classification from the Internet become very important. The purpose of this paper is to introduce our research effort on Internet-based Chinese corpus classification. A PAT-tree-based approach is developed and has shown its efficiency in utilizing Internet resources for Chinese natural language analysis.

1. Introduction

As Internet has become a major media for information dissemination, unlimited volume of corpus resources is now available over the Internet. To allow the unlimited volume of daily exploding online resources that can be effectively organized into corresponding subject domains and utilized in different natural language processing applications, techniques which can be easily adapted with the change of online resources for automatic corpus collection and classification therefore become very important. Considering there are not obvious works dealing with the above research issue especially on Chinese applications, the purpose of this paper is trying to introduce our research effort on developing Internet-based Chinese corpus storage and classification techniques.

2. PAT-tree-based Approach for Online Corpus Storage and Classification

Online resources are usually large and dynamic, it requires an efficient working structure for storage and access. In addition to providing fast search and easy update, such a structure would be efficient in extracting some statistical data like frequency values, associations, possible contexts of any data sequences. The sequences could be character strings, word strings, POS strings or even phone strings, depending on what kinds of abstract information to be extracted from the corpus.

2.1 The PAT-tree-based Working Structure for Corpus Storage

According to our experiments [1,2] and similar analysis [3,4], suffix-tree indexing techniques such as PAT trees are efficient in representing online corpus as a high-order N-gram *language model*, especially for when the corpus is large and dynamic.

PAT tree is an efficient indexing structure successfully used in the area of information retrieval [5]. Using this indexing structure for indexing full-text content of document databases, all possible data strings including their frequency counts in the database can be retrieved and updated in a very efficient way, but not every string with arbitrary length need to be stored. The time complexity of finding arbitrary length data segment and its frequency counts from the PAT tree is very efficient.

Considering the inherent difficulty of word segmentation and unknown proper noun extraction in Chinese, the indexing unit for Chinese corpus could be character rather than word for English in the PAT trees. In this way each distinct character string in the corpus which occurs within sentence fragment will be conceptually recorded. This is based on an assumption that only the context within a certain logical segment is significant for natural language analysis. We use punctuation marks such as “.” and “,” as delimiters to determine a segment boundary. For example, the different suffixes generated for the string “ $\alpha\beta\gamma\nu, \omega\varepsilon$ ” at the segment level are “ $\alpha\beta\gamma\nu$ ”, “ $\beta\gamma\nu$ ”, “ $\gamma\nu$ ”, “ ν ”, “ $\omega\varepsilon$ ”, and “ ε ” (each alphabet indicates a Chinese character). Using the PAT tree data structure, it is easy to perform prefix search. For each data stream recorded in the PAT tree, the existence of its composed sub-strings can be easily detected if the searching string is exactly a prefix of a suffix string. For example, if we check whether “ $\beta\gamma$ ” is a sub-string of the above string “ $\alpha\beta\gamma\nu, \omega\varepsilon$ ” using the PAT tree, we find that the answer is yes because “ $\beta\gamma$ ” can be found as a prefix of the suffix string “ $\beta\gamma\nu$ ”.

During the implementation, each distinct suffix string will be represented as a node in the PAT tree and has only a pointer to its position in the document to save space. In the mean time, each node consists of three parameters: the comparison bit, the number of external nodes and the frequency count for the purpose of searching and information updating. The number of external nodes indicates the number of different suffix strings in the sub-trees, and the frequency count indicates the frequency of occurrence of the corresponding string in the data stream. Moreover, the comparison bit indicates the first different bit of the strings recorded in the sub-trees. The comparison bit is primarily used in each node as an indication of which bit of the searching string is to be used for

branching. By storing the frequency of the recorded string and the number of external nodes (those nodes which have comparison bits greater than their parents') of each node, we are able to determine the frequencies of every strings and even occurrences of every string pairs stored in the PAT tree.

In fact, such an indexing structure is especially useful in constructing a variable n-gram Chinese language model. The PAT tree actually provides indices to all possible segments of characters with an arbitrary length N, where N can be arbitrary or just the maximum sentence length, together with the frequency counts for these segments in the corpus. For each character string it is easily to extract the occurrences of all neighboring strings. Since the content of online resources can be taken to train domain-specific language models, the language models can be easily adapted with the update of the online resources and the corresponding PAT tree indices. With the above advantages, the PAT-tree is then taken as the primary working structure in the proposed approach for both corpus storage and classification.

2.2 PAT-tree-based Corpus Classification

The representative meaning of a term or character string in a document is often subject to the specific domain of the document. To extract more precise linguistic information such as the extraction of domain-specific terms, the documents used as the training corpus need to be classified into corresponding domains. Slightly different from conventional document classification, the corpus classification here requires to be *incremental* with the increase of the online resources. In addition, it should allow *multiple classification* for ambiguous documents, because a few documents miss-classified into wrong collections normally make little interference from the point of view of linguistic analysis.

The proposed PAT-tree-based approach is not designed as a specialized algorithm for classification but a flexible working structure for the purposes of both corpus classification and information extraction. There are two major steps to perform: *PAT-tree-based classification* and *PAT-tree updating*. Because words in common dictionaries are too general to be “key” words, it is necessary for the online corpus classification to define a proper feature set which can be easily adapted with the update of the text collections. With a suffix-tree-like data structure it was proven may generate the feature set dynamically[1]. For this reason, each domain-specific text collection and classifying document from the online resource has a PAT tree to represent its feature vector. With the PAT-tree indices, the feature set to be selected can be variable length n-grams or extracted terms [6]. Such a unique feature makes the PAT tree efficient in performing *incremental classification*, because each new document after classification can update the corresponding PAT tree(s) and help for classifying incoming documents immediately. More importantly, the PAT tree indices also serve as the domain-specific variable-n-gram language models of the corresponding text collections. No extra feature vectors are therefore needed for the corpus classification application.

For realizing the performance, several different vector-space-model-based

estimation methods were implemented. The required parameters with TF-IDF-based metrics such as term frequency values and document frequency values were found can be efficiently extracted using the PAT-tree indices. As an experiment to be introduced below, on average the proposed approach can index and classify more than ten news documents per second on a PC.

2.3 Preliminary Experiments

The first experiment was performed to realize the flexibility of the proposed approach for the considering problem rather than to develop a really efficient approach in Chinese document classification at that stage. The initial training documents were obtained from CNA news abstracts of the whole eight sections published from January to May in 1997, and the testing documents were that published in June. Table. 1 lists some of the detailed information of the testing database constructed for corpus classification. It is noted that the documents in the testing database were news abstracts. On average, a document contained about 150 Chinese characters.

In this example, each of the testing documents was incrementally classified and indexed in sequence of chronicle order. The classified documents were immediately added into the corresponding text collection(s) for classifying subsequent documents. A VSM-based estimation metric was implemented. The estimation metric is defined below:

Let d be a document to be classified, c is an examining text collection, $T = (t_1, t_2, \dots, t_n)$ is a set of possible distinct variable n -gram patterns in d where T can be accessed by the PAT tree of d . $Score_c(d)$ is a similarity estimation function defined as follows.

$$Score_c(d) = \sum_i (L^2(t_i) * Tf(t_i; d) / N_c(t_i)) \dots \dots \dots (1)$$

where $L(t_i) = 0$ if $t_i \notin c$
 $L(t_i) = \text{stringlength}$ if $t_i \in c$

In the above equation, $L(t_i)$ is the string length of t_i while $Tf(t_i; d)$ is the term frequency of t_i in document d . $N_c(t_i)$ is the total number of distinct text collections where the pattern t_i ever occurs. If the above formula is applied to bi-gram based estimation, the set T is all possible character bi-grams in d . Similarly, if it is applied to variable- n -gram based estimation, the set T is all possible character strings in d .

The experimental results as illustrated in Table 2 were that obtained with the above TF-IDF-based metric. The obtained results as listed in Tables 1 and 2 show its feasibility in both corpus storage and classification. In terms of corpus storage, it performs efficient especially in handling variable n -gram features. In fact, the proposed PAT-tree-based approach was found can deal with several hundred megabytes of texts in extraction of linguistic information[7] and extract domain-specific terms incrementally [2]. The obtained results actually depend on the grouping of each text collection. It is observed that the more precise the pre-divided text collections, the better the results can be the obtained. The weather section (section 7) is an example with very high precision and

recall rates. In fact, in another test with different text collections the proposed approach can achieve about 90% recall and precision rates.

| Domain | PAT-tree Size (K Bytes) | Training Size | | Testing Size | | Speed (#Doc/Sec on PC) |
|------------|-------------------------|---------------|---------|--------------|---------|------------------------|
| | | Doc# | K Bytes | Doc# | K Bytes | |
| Politics | 27,310 | 12,587 | 4,490 | 1,262 | 382.0 | 10.51 |
| Transport | 15,852 | 10,114 | 2,490 | 1,173 | 227.0 | 15.45 |
| Business | 7,071 | 7,337 | 1,800 | 1,082 | 219.0 | 15.03 |
| Education | 4,894 | 2,942 | 775 | 424 | 87.4 | 11.13 |
| Recreation | 2,571 | 1,626 | 413 | 248 | 49.3 | 13.47 |
| Local | 11,117 | 6,580 | 1,690 | 968 | 201.0 | 14.57 |
| Weather | 867 | 2,896 | 3,130 | 435 | 474.0 | 8.97 |
| Misc. | 5,844 | 989 | 1,690 | 132 | 193.0 | 3.80 |
| Total/Avg. | 75,526 | 45,071 | 16,478 | 5,724 | 1,832.7 | 12.07 |

Table 1. The testing database and some of the obtained processing results.

| | With Bi-gram Features | | | With Variable N-gram Features | | |
|------------|-----------------------|-------------|-----------|-------------------------------|-------------|-----------|
| | Top1 Recall | Top2 Recall | Precision | Top1 Recall | Top2 Recall | Precision |
| Politics | 98.7% | 99.8% | 33.6% | 96.8% | 99.4% | 45.1% |
| Transport | 46.9% | 87.1% | 71.3% | 60.0% | 85.4% | 74.6% |
| Business | 20.8% | 30.7% | 99.6% | 35.8% | 54.9% | 98.0% |
| Education | 24.3% | 71.0% | 77.5% | 50.7% | 78.5% | 72.4% |
| Recreation | 59.3% | 69.8% | 97.4% | 75.0% | 82.7% | 93.0% |
| Local | 14.0% | 47.3% | 80.0% | 42.9% | 73.5% | 80.0% |
| Weather | 98.2% | 100.0% | 100.0% | 100.0% | 100.0% | 99.8% |
| Misc. | 38.6% | 97.7% | 38.9% | 90.2% | 98.5% | 54.1% |
| AVG. | 50.3% | 71.7% | 74.8% | 64.3% | 81.4% | 77.1% |

Table 2. The obtained recall and precision rates on incremental corpus classification.

3. Concluding Remarks

To handle Internet resources, it really needs a specially-designed information spider, which is designed to automatically extract relevant resources such as Web pages and network news from the Internet. Based on the proposed approach, we have constructed a prototype system for real-time Chinese news collection. Up to now, we have obtained more than 2 giga bytes of Chinese netnews which were found mainly in Chinese new agencies and classified into 25 categories. In the mean time, 4 giga bytes of BBS data from educational news groups supported by Ministry of Education in Taiwan were also

collected and classified into 250 categories daily. Besides, we also have developed a Web spider for web resource collection. The proposed corpus classification technique has been shown their efficiency in utilizing Internet resources for Chinese natural language analysis.

References

1. Chien, Lee-Feng (1997) *PAT-Tree-Based Keyword Extraction for Chinese Information Retrieval*, Proceedings of ACM SIGIR'97, Philadelphia, USA, pp. 50-58.
2. Chien, Lee-Feng (1999) *PAT-Tree-Based Adaptive Keyphrase Extraction for Intelligent Chinese Information Retrieval*, to appear on Information Processing and Management , Elsevier Press.
3. Yamamoto, M. and Church, K. (1998) *Using Suffix Arrays to Compute Term Frequency and Document Frequency for All Substrings in a Corpus*, Proceedings of the Sixth Workshop on Very Large Corpora, Montreal, pp. 28-37.
4. Nagao, M. and Mori, S. (1994) *A New Method of n-gram statistics for large number of n and automatic extraction of words and phrases from large text of Japanese*, Proceedings of COLING-94, pp. 611-615.
5. Gonnet, G. H., Baeza-yates, R. et al. (1992) *New Indices for Text: Pat Trees and Pat Arrays*. Information Retrieval Data Structures & Algorithms, pp. 66-82, Prentice Hall.
6. Zamir, Oren and Etzioni Oren (1998) *Web Document Clustering: A Feasibility Demonstration*, Proceedings of ACM SIGIR'98, pp. 46-53.
7. Chen, Chun-Liang, Bai, Bo-Ren, et al. (1998) *PAT-tree-based Language Modeling with Initial Application of Chinese Speech Recognition Output Verification*, Proceedings of the 1998 International Symposium on Chinese Spoken Language Processing (ISCSLP) Best Student Paper Award, Singapore.