Communications and Network, 2013, 5, 81-85
doi:10.4236/cn.2013.51B019 Published Online February 2013 (http://www.scirp.org/journal/cn)

Design and Implementation of a New Chinese Word Segmentation Dictionary for the Personalized Mobile Search*

Zhongmin Wang, Jingna Qi, Yan He
School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an, China

Received 2012

ABSTRACT

Chinese word segmentation is the basis of natural language processing. The dictionary mechanism significantly influences the efficiency of word segmentation and the understanding of the user's intention which is implied in the user's query. As the traditional dictionary mechanisms can't meet the present situation of personalized mobile search, this paper presents a new dictionary mechanism which contains the word classification information. This paper, furthermore, puts forward an approach for improving the traditional word bank structure, and proposes an improved FMM segmentation algorithm. The results show that the new dictionary mechanism has made a significant increase on the query efficiency and met the user's individual requirements better.

Keywords: Chinese Word Segmentation; Dictionary Mechanism; Natural Language Processing; Personalized Search; Word Classification Information

1. Introduction

With the rapid growth of network information resources, the personalized information services have become a hot issue of the contemporary search engine. How to obtain the user's search intention becomes the primary task of the personalized search. At present, many technologies can be used to get users' interests, such as natural language understanding, concept-based retrieval, semantic network technology, etc. In order to overcome some of the shortcomings of these common technologies and meet the need of the users' mobile search, a new dictionary word segmentation mechanism is proposed which can get the word classification information of the user's query directly after word segmentation. And these category messages can be used to infer the user's search intentions.

The traditional dictionary mechanisms for Chinese word segmentation are mainly based on binary-seek-by-word, TRIE indexing tree, binary-seek-by-characters [1]. Among the three methods, the mechanism based on binary-seek-by-characters is the improvement of the previous two dictionary mechanisms, which integrates the simple dictionary body of binary-seek-by-word dictionary mechanism and high-efficiency search processing of TRIE indexing tree dictionary mechanism. Since it still uses the binary-seek-by-word structure, the search range for second word is not reduced, and its efficiency is limited [2]. At present, public dictionary mechanisms have their own strengths, but they are basically improved on the basis of traditional dictionary mechanism. Such as double-character-hash-indexing, multi-character-hash-indexing [1,2,4,5]. They are indeed effective ways to improve the matching efficiency of the system. But directly using the hash method on words makes hash table or index table difficult to build, and the complex structure of this kind of word bank makes it difficult to maintain.

This paper puts forward an efficient dictionary mechanism for word segmentation based on the classification information of word. It can get word classification information of the users' query to provide the basis for the establishment of the user's interest model. On this basis, this paper also provides an approach for improving the FMM algorithm to fit for the new dictionary mechanism which could greatly enhance the efficiency of the word segmentation.

2. A New Dictionary Mechanism with the Word Classification Information

The new dictionary mechanism that contains the word classification information is the improvement of binary-seek-by-characters. It employs a new hash mecha-

*This work is supported by Shaanxi Science & Technology Department project (2011K06-26) & national natural science foundation project (61100166)

nism of the second-character's area code subsection, which can greatly reduce searching scope of the second-character and accelerate the process of the system. The new dictionary mechanism could also provide the word classification information after word segmentation which is the basis for the personalized analysis of the user's intention.

2.1. New Dictionary Architecture

The category information of the new dictionary is designed as **Figure 1**, which consists of entertainment, sports, urban, nature, engineering and other 12 kinds of word category. Each category contains the corresponding branches with 3 levels, and each level is encoded sequentially. This encoding method can quickly identify the classification information of the target word.

As for the design of the new dictionary with word classification information, firstly process the original linear dictionary according to the category information to get independent linear dictionary based on category information. Then change the classified linear dictionary into the new dictionary according to the hash mechanism of the second-character's area code subsection. And classification information is added to each entry in this process.

2.2. Chinese Word Frequency Statistics

As [3,7] shown, the number of two-character words is quite larger than the single words and multi-character words. As the number of words that have the same first character is so large, it needs more time to find the second character using the dichotomizing search mechanism. So reducing the searching scope of the second character can improve the query efficiency greatly.

To avoid the problem of too long hash table of the second character caused by double-character-hash-indexing, this paper proposes a new method—the hash mechanism of the second-character's area code subsection. First of all, get the statistical result of distribution range of the second word in the original word bank. Second, divide second characters into non-uniform subsection. In this process, entries with the same first character are mapped to different subsection according to their second-character's area code (Section number is: 1-20). The new mechanism only increases a small amount of storage space, but greatly reduces the range of the second-character inquiry.

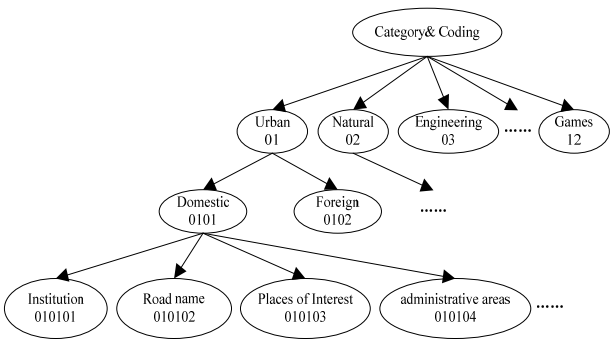

Figure 1. Category information of the dictionary.

2.3. The Construct Procedure of Dictionary

During the construction of dictionary, each character in the first-character Hash table is added with an additional data structure which is used to store the information of area code subsection of the second-character. Then the second-character is hashed into 20 sub-tables according to its area code. Through this mechanism, we can get the pointer of second-character indexing table according to the section number of the second-character's area code. This will greatly reduce the range of inquiry of the second-character. The construct procedure of the dictionary is shown in **Figure 2**.

Table 1. Chinese word freq uency statistics.

The number of characters	1	2	3	4	5
The number of words	9919	*65891*	*26352*	*21699*	*5124*

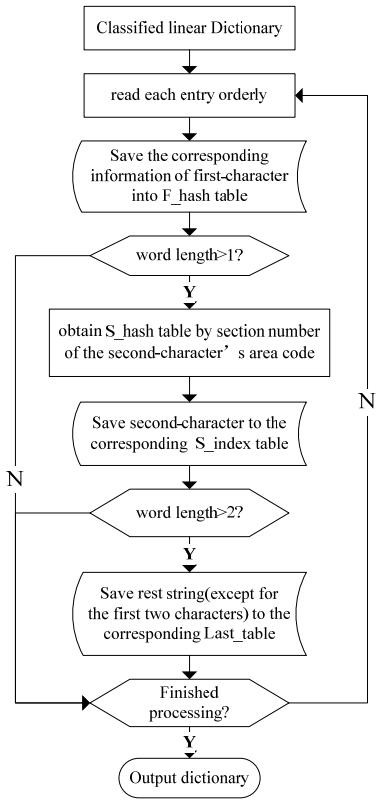

Figure 2. Construct flow of dictionary.

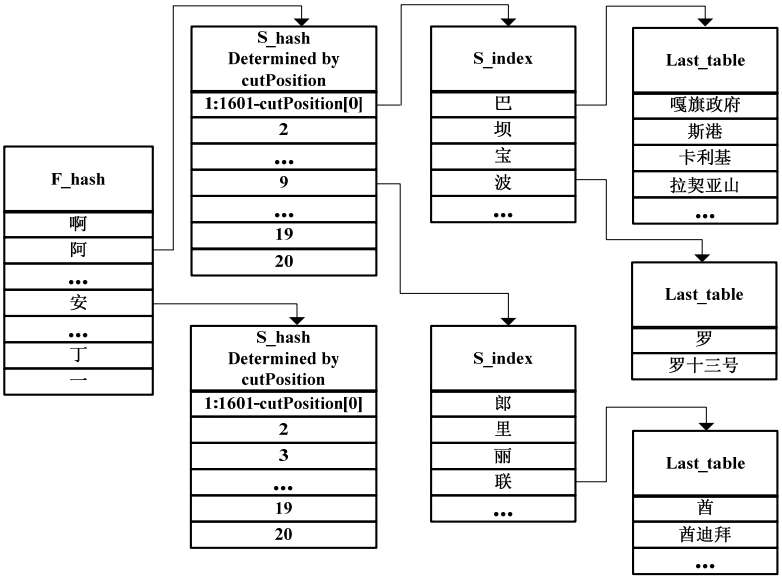

Figure 3. The logical structure of the dictionary.

The dictionary mechanism proposed in this paper also integrates the information of the word frequency. During the segmentation, word frequency can be taken into account. And the words with high frequency have the priority to be selected, which may further improve the efficiency of the system.

2.4. New Dictionary's Logical Structure

The logical structure of the new dictionary is shown in **Figure 3**. As shown in this figure, the new dictionary structure consists of four parts: First-character Hash table, Area code subsection of the second-character Hash table, Second-character indexing table, Dictionary text table.

1) *First-character Hash table: F_hash*

First-character Hash table which adds some attribute values is formed on the basis of present mature mechanism of First-character Hash table. The structure is shown in **Table 2**.

The first characters (word First) are the Chinese characters from GB2312 encoding table. According to the machine code of the Chinese characters, we can quickly locate the first character in the F_hash table. Hash function is as follows:

$$offset = (c_1 - 0\text{xB0}) \times 94 + (c_2 - 0\text{xA1}) \quad (1)$$

c1, c2, represent the high byte and the low byte of the character's machine code. Offset represents character's array subscripts in F_hash.

In this table, is Word represents whether the first character is a word. If the first character is a word (the value of is Word is TRUE), then frequency and coding in the table can be used to record the word frequency and their classification respectively. S_hash is a pointer which points to the table of S_hash. This table belongs to the second-character of words which begin with wordFirst. After obtain the second-character's subsection index, it can navigate to the next unit with the current pointer of s_hash.

2) *Area code subsection of the second-character Hash table: S_hash*

In the process of this stage, all the words in the original word bank are analyzed statistically to get the frequency of each second-character. And then according the frequency, the second-characters are arranged into different areas. The second-character with higher frequency is divided into a small interval, and low frequency is divided into large interval. By this means, words begin with the same character are divided into non-uniform subsection. It is the non-uniform interval that makes the number of words in each subsection basically achieve uniform distribution. The new mechanism can greatly reduce the inquiry range of the second-character, and can avoid the disadvantages of query efficiency limitation caused by non-uniform distribution of the second-character. The structure of S_hash is shown in **Table 3** regionIndex (1-20) represents subsection index of second-character's area code. s_index is a pointer that points to the table of S_index.

3) *Second-character indexing table : S_index*

Table 2. F_hash structural.

wordFirst	isWord	frequency	coding	s_hash

Table 3. S_hash structura.

regionIndex	s_index

This table stores the second-character of words which meet the table of F_hash and S_hash constraints. Each record in the table consists of 5 parts. Its structure is shown in **Table 4** word Sec represents second-character. isWord, frequency, coding are same as the property in First-character Hash table. last_table is a pointer that points to the table of Last_table.

4) *Dictionary text table: Last_table*

Last_table is composed of dynamic arrays which stores all words excluding the first two characters. The structure of this table is shown in **Table 5** lastStr represents the remaining string. Frequency, coding is same as the property in First-character Hash table.

3. Improved Fmm Algorithm

Based on the new dictionary construction, this paper proposed an improved Former Max Matching algorithm (FMM). In word segmentation, instead of using fixed maximum segmentation length [6-8], the improved FMM algorithm using dynamic length determined by the target words which begin with the current first two characters on query sentence segmentation. This method effectively overcomes the ineffectiveness of length limitation caused by FMM.

Compared to FMM, the improved FMM has the following two improvements:

1) According to the new dictionary structure, the second-character W_2 is located like the following way. First, obtain region Index depending on area code of W_2. Second, locate second-character's place in S_hash that corresponding to W_1 according to the value of region Index. Then, we can find the second-character in S_index accordance with region Index. This strategy of second-character subsection can greatly reduce the searching scope of the second-character.

2) After getting the position of the second-character, we can obtain Last_table corresponding to W_1W_2. Then the improved FMM intercepts the current Chinese string according to the length of words in Last_table and compares the intercepted word with the corresponding words in Last_table. During the process, the improved FMM will select the optimal segmentation according to word length and frequency. Because the segmentation length is dynamically determined by the length of words to be matched, the efficiency is improved greatly, and the shortcoming of low efficiency, caused by fixed segmentation length in FMM, is avoided.

Table 4. S_index structural.

wordSec	isWord	frequency	coding	last_table

Table 5. S_index structural.

lastStr	frequency	coding

4. Experimental Results

By comparing the hash mechanism of the second-character's area code subsection with binary-seek-by-characters dictionary mechanism, the new dictionary mechanism is tested from two aspects: time and space complexity. In the experiments, the same improved FMM algorithm is used to process the test text (31K) based on the two dictionaries. The experimental results are given in **Table 6**:

As shown in **Table 6**, the memory space of new dictionary mechanism is 13% more than that of the binary-seek-by-characters dictionary mechanism, but the time efficiency is improved by about 20%. Besides, word classification information can be obtained after word segmentation. For example, the input test sentence is: "计算机网络是计算机技术与通信技术结合的产物", and output result is shown in **Figure 4**.

For the first word of "计算机网络" of the result, it belongs to the classification"031402,031404,032001" which indicates the word belongs to the engineering application of computer science and technology.

Table 6. Time and space consuming comparison of the two dictionary mechanism.

Dictionary mechanism	Space consuming/Byte	Time consuming/ms
Binary-seek-by-characters	3553,068	156
New mechanism	4014,108	125

Figure 4. Segmentation test.

5. Conclusions

The hash mechanism of the second-character's area code subsection, proposed in this paper, can greatly reduce the searching scope of the second character. It can improve the speed of word segmentation, and realize the improvement of time efficiency by consuming small space. With the increase in the number of dictionary entries, the superiority of the new dictionary mechanism will be more significant. Meanwhile, the new dictionary integrates word classification information effectively which can be used for user's interest modeling and provides the basis to meet the personalized needs of mobile users.

REFERENCES

[1] M.-S. Sun and Z.-P. Zuo, "An Experimental Study on Dictionary Mechanism for Chinese Word Segmentation," *Journal of Chinese Information Processing*, Vol. 1, 2000, pp. 1-6.

[2] W. Yang, L.-Y. Ren and R. Tang, "A Dictionary Mechanism for Chinese Word Segmentation Based on the Finite Automata," *2010 International Conference on Asian Language Processing (IALP)*, pp. 39-42.

[3] Z. X. Li, Z. P. Xu, W. Q. Tang and R. X. Tang, "Ambiguity Processing in Word Segmenting," *Computer Engineering and Applications*, Vol. 38, No. 11, 2002, pp. 106-109.

[4] Q. Y. Zhang and S. Chai, "Chinese Word Segmentation Dictionary using Two-level Index," *Computer Engineering and Applications*, Vol. 19, 2009.

[5] Q. H. Li, Y. J. Chen and J. G. Sun, "A New Dictionary Mechanism for Chinese Word Segmentation," *Journal of Chinese Information Processing*, Vol. 17, 2003, pp. 13-18.

[6] Y. Niu and L. L. Li, "An Improved Chinese Segmentation Algorithm Based on New Dictionary Construction," *International Conference on Computational Science and Engineering*, Vol. 2, 2009, pp. 993-996.

[7] A. Choi, C. H. Cheng and Y. L. Ko, "Word Extraction from Chinese Documents by Occurrence Counts," *1988 International Conference on Computer Processing of Chinese and Oriental Languages*, Toronto, Canada, pp. 488-491.

[8] H. Y. Cui, "Research 0n an Improved Chinese Segmentation Algorithm based on Word Frequency Statistic," Information Technology, Vol. 04, 2008.

