International Journal of Fuzzy Logic and Intelligent Systems, vol. 12, no. 1, March 2012, pp. 15-19
http://dx.doi.org/10.5391/IJFIS.2012.12.1.15 pISSN 1598-2645 eISSN 2093-744X

A New Approach of Domain Dictionary Generation

Su mei Xi[1,2] • Young Im Cho[1]* • Qian Gao[1,2]

[1]College of Information Technology, The University of Suwon, San 2-2, Bongdam-eup, Hwaseong, 445-743, Korea
[2]Department of Information, Shandong Polytechnic University, Jinan, 250353, China

Abstract

A Domain Dictionary generation algorithm based on pseudo feedback model is presented in this paper. This algorithm can increase the precision of domain dictionary generation algorithm. The generation of Domain Dictionary is regarded as a domain term retrieval process: Assume that top N strings in the original retrieval result set are relevant to C, append these strings into the dictionary, retrieval again. Iterate the process until a predefined number of domain terms have been generated. Experiments upon corpus show that the precision of pseudo feedback model based algorithm is much higher than existing algorithms.

Keywords: Domain Dictionary, Pseudo Relevance Feedback, In Word Probability

1. Introduction

Domain Dictionary refers to the set of special terms or expressions of specific domains. Domain Dictionary automatic generation refers to the process of identifying the domain terms or expressions automatically through processing the related corpus of the particular fields.

Artificial compilation of Domain Dictionary not only costs a lot of statistical work but also the lack of real-time. Automatic generation of domain dictionary can remedy the weaknesses of artificial compilation of Domain Dictionary and can be applied to Information Retrieval, public opinion analysis, corpus construction and index word optimization directly, as the foundation of the Chinese information processing technology at the same time. Domain Dictionary automatic generation can be described formally as follows:

A given domain corpus C, general dictionary D_{com}, the system can automatically generate terms set D_{dom} of C, which meets that any element T of D_{dom} has specific semanteme, appearing one or several times on the C, and not the D_{com} element. Each element of domain dictionary is named domain terminology. Domain terminology has special semanteme, which belongs to a special domain. Generally candidate domain terminologies can be obtained by mining the Meaningful Strings. The Meaningful Strings are referred to those strings that have special semanteme and can be used independently.

The Meaningful String has several characteristics as follows:

1) It has some negotiability and appears frequently in the real corpus.
2) Its inner structure is stable and it has certain coagulability.
3) Its using environment is flexible and it can appear in some language environments.

Nowadays many researchers have studied in Meaningful String recognition aiming at some special application domains, such as using Meaningful Strings recognition result for retrieval and category domains to increase the efficiency of retrieval and category [1,2]. They also use the Meaningful Strings recognition for frequent key pattern extraction to extracting the text categories or cluster characteristics and so on.

The automatic generation of Domain Dictionary is regarded as a process of information retrieval in this article. The Pseudo Relevance Feedback technology of retrieval model is used to increase the precision of Domain Dictionary generation, added several top retrieval results into the Domain Dictionary, retrieval again, iteration again until the number of terminologies achieve the given threshold value.

2. Relevant Works

Given a massive documents set D and user input queries set Q, for each query q_i of Q, a sort function value R (q_i, d_j) can be associated to every document d_j of D, the bigger R(q_i , d_j) is, the more relevant about d_j and q_i. This process can be regarded as information retrieval process. The user's query usually only includes a few key words in the actual IR system such as Search Engine, which can cause the word mismatch between the relevant documents and user's query, so the negative effect of the retrieval performance may be serious. The Pseudo Relevance Feedback is a usual approach of query expansion. It assumes that the several top documents of the first retrieval results are relevant to the query, and then uses the standard Relevance Feedback technology to expand the query of user[3]. Many TREC evaluation results show that Pseudo Relevance Feedback is a simple but very efficient technology to query expansion.

Each domain terminology of Domain Dictionary is a semantic unit of special meaning and it can be used

Manuscript received Jan. 11, 2012; revised Mar. 7, 2012; accepted Mar. 14, 2012
*Corresponding author: Young Im Cho(ycho@suwon.ac.kr)

independently, therefore Domain Dictionary is a set of Meaningful Strings of special domain corpus. There are some works about mining Meaningful String have been applied for Domain Dictionary automatic generation. Feng [4] puts forward the notion of Adjacency Variety to describe the flexibility of strings.

The string's Left Adjacency Variety refers to the species number of words appearing on the string's left and the string's Right Adjacency Variety is also similar. The string's Adjacency Variety is defined to the minimum among Left Adjacency Variety and Right Adjacency Variety.

The bigger Adjacency Variety of the string is, the more flexible it uses, the more possible it is used independently, so the bigger probability it is a Meaningful String. Zougang [6] gets the candidate Meaningful Strings by counting repetitive strings on large scale Web corpus and then filters some junk strings through some principles, so the Meaningful Strings can be obtained. Hemin puts forward a solution of mining Internet Meaningful String on the basis of Feng's work. He realized a new word found algorithm of Internet and in his algorithm the context environment of string is used to denote using flexibility of string and inner structure to denote coagulability of string.

String's Adjacency Variety is used to denote flexibility of string in this approach and In Word Probability to denote coagulability of string. Among this approach In Word Probability and Position Word Probability are defined as follows:

In Word Probability, IWP: IWP(c) is defined to the ratio of independent frequency about a single Chinese character c in training corpus over the appearing frequency of the word [5], i.e.

$$IWP(c) = \frac{N(c,w)}{N(c)} \qquad (1)$$

N(c,w) denotes the frequency of a single character appearing in other words, N(c) is the total frequency of c appearing in the corpus. The string S's In Word Probability IWP(S) refers to the product of all of the IWP(c) which c composing the string S, i.e.

$$IWP(S) = \prod_{c \in S} IWP(c) \qquad (2)$$

The higher the In Word Probability of string S the more possible it is Meaningful String.

Position Word Probability, PWP: Some Chinese characters tend to appear in a special position of word, such as 老 usually appears at first of word but 性 maybe appears at rear of word. Position Word Probability PWP (c, pos) of Chinese character c is computed as follows:

$$PWP(c, pos) = \frac{N(c, pos)}{N(c)} \qquad (3)$$

Value range of pos is 0, 1, 2, they denote the first, ending and intermediate position in the word. If the string S includes some character c and PWP(c,pos) is less than some threshold value, S may not be the Meaning String. The string S's Position Word Probability PWP(S,C) can be defined to the product of all of the PWP(c,pos) which c composing the string S, i.e.

$$PWP(S,C) = \prod_{c \in S} PWP(c, pos) \qquad (4)$$

3. Domain Dictionary Generation Algorithm Based on Pseudo Feedback Model

In this paper a Domain Dictionary automatic generation algorithm based on Pseudo Relevance Feedback model is put forward, it regards the generation of Domain Dictionary as IR process, i.e. according to the corpus of user's given special domain documents to retrieve the domain terminology collection related with the corpus. The difference to traditional IR is that traditional IR list some relevant website collection related with user's query according to user's query words but the process of Domain Dictionary generation is to give the set of all domain terminologies according to the special domain corpus of user.

3. 1. Domain Correlation Degree of String

In traditional IR system the notion of Correlation Degree is used to evaluate the connection degree of query word and document. We can also define the notion of Correlation Degree like IR system in Domain Dictionary automatic generation system.

In the Domain Dictionary automatic generating system, Correlation Degree is used to evaluate the Correlation Degree of string and the domain.

Given corpus C which is constructed by some special domains documents and string S, the domain Correlation Degree R(S,C) about string S relative to corpus C is often described by characteristic frequency and Adjacency Variety of S among C. F(S,C) denotes the appearing number of S among C, which describing the generality of string. Adjacency Variety AV(S,C) describes the flexibility of string, it is defined to the minimum of Left Adjacency Variety and Right Adjacency Variety of S among C. The Correlation Degree R(S,C) is defined as follows:

$$R(S,C) = \lambda_1 \times \log(f(S,C)) + (1-\lambda_1)\log(AV(S,C)) \qquad (5)$$

λ_1 is a preference parameters, and $0 \le \lambda_1 \le 1$. We use it to adjust weight of characteristic frequency and Adjacent Variety of the domain Correlation Degree. If $\lambda_1 = 1$, it denotes the domain Correlation Degree only considering characteristic frequency; if $\lambda_1 = 0$, denotes only considering Adjacent Variety.

3. 2. Domain In Word Probability of String

Given domain corpus C and general dictionary D_{com}, for each Chinese character c, we can count the appearing frequency N(c) about c of corpus C; we also can count the dependence frequency N(c,w) about c of corpus C by segmenting words of corpus C according to some word segmentation strategies and using D_{com}, and then we can

calculate In Word Probability IWP(S,C) about string S in corpus C.

We can construct the Domain Dictionary D_{dom} of C by adding some or all of the domain terms into the general dictionary D_{com}. Segmenting the words of C by using D_{dom} according to the same word segmentation strategy, the frequency N'(c,w) of c appearing in C dependently can be counted. We also can calculate the In Word Probability IWP'(S,C) about string S in corpus C. The domain In Word Probability $IWP_{dom}(S,C)$ of string S in corpus C is defined as follows:

$$IWP_{dom}(S,C)=\lambda_2\times IWP(S,C)+(1-\lambda_2)IWP'(S,C) \quad (6)$$

Among them λ_2 is a preference parameters, and $0\leq\lambda_2\leq1$. We use it to adjust weight of In Word Probability using D_{com} and D_{dom} of domain In Word Probability. If $IWP_{dom}(S,C)$ is less than a given threshold value, it is regarded as a junk string.

Similarly we can define the domain Position Word Probability $PWP_{dom}(S,C)$ of string S in corpus C. If the $PWP_{dom}(S,C)$ is less than a given threshold value, it is regarded as a junk string. The specific calculation formula is as follows:

$$PWP_{dom}(S,C)=\lambda_3\times PWP(S,C)+(1-\lambda_3)PWP'(S,C) \quad (7)$$

3. 3. The Pseudo Relevance Feedback Model of Domain Dictionary Generation

Some studies show that the precision of Meaningful String found algorithm based on Word Segmentation is higher than Meaningful String found algorithm based on character, therefore, the Domain Dictionary automatic generation is generally done on the basis of Word Segmentation.

Because using common dictionary when segmenting words it is possible to make many segmentation mistakes when segmenting words of domain corpus and to affect the statistical data used in searching Meaningful Strings such as Adjacency Variety. Furthermore, In Word Probability computed using common corpus is not proper to domain corpus certainly. In this paper we use Pseudo Relevance model to solve these two problems. Assuming that top N Meaningful Strings are relevant to the domain, putting these results into Domain Dictionary, segmenting words using the new generated dictionary, finding Meaningful String again; meanwhile, updating In Word Probability by corpus that be obtained using Domain Dictionary. Repeating the above iteration process, the generation process of Domain Dictionary based on Pseudo Relevance Feedback is described as follows.

The first step is to select a general dictionary D_{com} and segment the words of corpus C. On the basis of Word Segmentation we find the frequent model set $FP(C)^1$, each of which characteristic frequency is greater than a certain threshold value, using frequent pattern found algorithm. Secondly, for each element S (S is not in the D_{com}) of $FP(C)^1$ we can calculate its domain Correlation Degree R(S,C) according to the formula (5) and can rank all elements of $FP(C)^1$ according to its domain Correlation Degree R(S,C). We can create the candidate domain terms set D^1_{can} by using the In Word Probability and Position Word Probability of string S in general dictionary D_{com} in C and filtering part of the elements of $FP(C)^1$.

Some strings sorting before their domain Correlation Degrees are greater in the candidate domain term set D^1_{can}, so we can assume that the top N strings sorting before of D^1_{can} are relevant to corpus C. The top N strings are added into D_{com} to create first Domain Dictionary D^1_{dom}. And then using the first Domain Dictionary D^1_{dom} to segment the words of corpus C, we can obtain the frequent model set $FP(C)^2$ of C. Similarly the candidate domain term set D^2_{can} can be obtained according to the domain In Word Probability and domain Position In Word Probability and then domain In Word Probability and domain Position In Word Probability will be updated.

Similarly we can assume that the top N strings of candidate domain term set are related to corpus C, adding these top N strings into the first Domain Dictionary D^1_{dom} to create the second Domain Dictionary D^2_{dom}. Similarly D^3_{dom}, D^4_{dom} ,… , D^n_{dom} can be calculated until the Domain Dictionary achieving the given scale or there is no string that satisfy the condition of domain Correlation Degree and domain In Word Probability can be added into the dictionary. The algorithm flow chart is as figure 1.

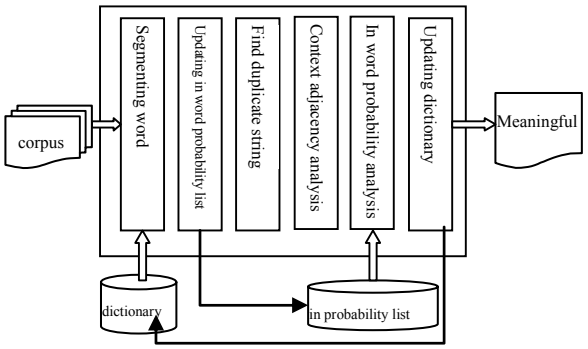

Figure 1 process of Domain Dictionary generation algorithm based on Pseudo Relevance Feedback

3. 4. Algorithm Process

Domain Dictionary generation algorithm based on Pseudo Relevance Feedback is showed as algorithm 1.

Algorithm 1: Domain Dictionary generation algorithm based on Pseudo Relevance Feedback model

Input: domain corpus C, common dictionary D_{com}, In Word Probability table T_1 and Position Word Probability table T_2 of each Chinese character by common corpus, repeat string frequency threshold θ_1, In Word Probability threshold θ_2, Position Word Probability threshold θ_3, terminology number N of adding into the Domain Dictionary each time during the Pseudo Relevance Feedback, string minimum frequency θ_4 of adding into the Domain Dictionary;

Output: Domain Dictionary D_{dom} of domain corpus C.

1) initialization $D_{dom}= D_{com}$;
2) segment words of C by using D_{dom}, update T_1 and T_2;
3) find the frequent mode that characteristic frequency is greater than θ_1 after segmenting words, construct FP(C) set;

4) for each string S of FP(C) execute step 5) to step 7);
5) calculate IWP(S) and PWP(S) of S;
6) if IWP(S) $<\theta_2$ or PWP(S) $<\theta_3$, delete S;
7) calculate domain Correlation Degree R(S,C) of S;
8) rank all strings of FP (C) according to domain Correlation Degree;
9) for each string of top N in FP (C), execute step 10;
10) if word frequency of S is greater thanθ_4, add S into D_{dom};
11) if there is no any S to be added into D_{dom}, stop the algorithm; otherwise turn to step 2).

4. Experiment Result and Analysis

We have implemented this Domain Dictionary generation algorithm by using Visual C++ on Windows XP operation system. There are two domain corpuses in this experiment. The one is BBS title corpus of several mainstream BBS sites in 2006 and the other is some patent data of computer field. The size of common dictionary is 29083 before Domain Dictionary generated. The system Word Segmentation method is the biggest positive matching method. There are 1000 domain terminologies found for each of BBS title corpus and patent corpus. We compute the precision of existing Domain Dictionary generation algorithm and the precision of Domain Dictionary generation algorithm based on Pseudo Relevance Feedback model, and the results is showed in figure 2.

From figure 2, we can see that the precision of Domain Dictionary generation algorithm based on Pseudo Relevance Feedback model is higher than traditional Domain Dictionary generation algorithm.

Table 1 and table 2 show the top 10 meaningful strings of BBS title corpus and patent corpus. These tables show that the top terminologies their precisions are very high so they can be used to execute Pseudo Relevance Feedback. Furthermore figure 2 shows that if we do not use Pseudo Relevance Feedback model the top 100 domain terminologies their precisions are high. Therefore in this Domain Dictionary generation algorithm based on Pseudo Relevance Feedback model we add the top 100 Meaningful Strings into the Domain Dictionary until 1000 domain terminologies are found.

Table 1 top 10Meaningful Strings result of BBS corpus

Word	Word frequency	Word	Word frequency
转帖	15778	组图	2302
原创	12340	火影	2164
帖子	11320	仙剑	1672
贴图	8143	菜鸟	1636
版主	4308	中文字幕	1511

Table 2 top 10 Meaningful Strings result of patent corpus

Word	Word frequency	Word	Word frequency
计算机系统	2938	高速缓存	1193
基板	2340	电脑主机	1191
客户端	1849	插槽	1136
机箱	1761	液晶显示器	1099
密钥	1744	散热片	1078

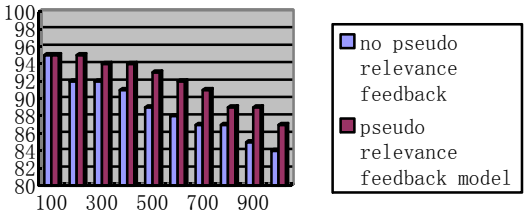

Figure 2 precision comparison of Domain Dictionary generation algorithm

5. Conclusions

The Domain Dictionary generation algorithm based on Pseudo Relevance Feedback model regards the generation of Domain Dictionary as IR process. It assumes that the several Meaningful Strings before are related to the corpus, using them for Pseudo Relevance Feedback, updating relevance probability, iteration as before, until finding the given number of domain terms or the Domain Dictionary increasing no longer. The Domain Dictionary generation algorithm based on Pseudo Relevance Feedback model can increase the precision of Domain Dictionary generation algorithm. We can apply the automatic generated Domain Dictionary for text classification and information retrieval systems in order to increase the efficiency of classification and retrieval.

References

[1] Jian Zhang, Jianfeng Gao, Ming Zhou, “Extraction of Chinese Compound Words: An Experimental Study on a Very Large Corpus”, *ACL2000 Second Chinese Language Processing workshop,* 2000.

[2] YuSheng Lai, Chung, “Meaningful Term Extraction and Discriminative Term Selection in Text Categorization via Unknown Word Methodology”, *ACM Transactions on Asian Language Information Processing* , vol.1,pp. 34-64, 2002.

[3] Rocchio, J. Relevance, “feedback in information retrieval”, *In: The Smart Retrieval System Experiments in Automatic Document Processing, G. Salton, Ed. Prentice Hall, Englewood Cliffs,* NJ, pp.313-323. 1971.

[4] Haodi Feng , Kang Chen, Xiao tie Deng et al, " Access or Variety Criteria for Chinese Word Extraction Computer Linguistics" , vol.30, no. 1, 2004.
[5] Haowei Qin, "A survey of Chinese new word recognizing characteristic", *computer engineering*, 2004, 12.
[6] Gang Zou, Yang Liu et al, "Chinese new word detection face to Internet", *Journal of Chinese Information Processing*, vol.18, no.6, pp.1-9, 2004.

Su mei Xi
2001 Bachelor, Shandong University of Science and Technology
2009 Master, Shandong University
Current Ph.D. Course student, University of Suwon
Research Area: Artificial Intelligence

E-mail : xiyanzi_79@sina.com.cn

Young Im Cho
1988. Bachelor, Korea University
1990. M.Sc. Korea University
1994. Ph.D. Korea University
Current. Professor at Univ. of Suwon Dept. of computer science
Research Area: Artificial Intelligence, Agent System, Ubiquitous System etc.

E-mail : ycho@suwon.ac.kr

Qian Gao
2001 Bachelor, Shandong University of Science and Technology
2008 Master, Shandong University
Current Ph.D. Course student, University of Suwon
Research Area: Artificial Intelligence

E-mail : gq@spu.edu.cn