Experiments Study for Scientific Texts Domain Keyword Acquisition *

Xiangfeng Luo, Ning Fang, Weimin Xu, Sheng Yu, Kai Yan and Huizhe Xiao

Semantic Grid and Web Content Analysis Group, Key Lab. of Grid Technology,
Shanghai University, Shanghai, 200072, China.
{luoxiangfeng@163.com, fangning@graduate.shu.edu.cn and wmxu@staff.shu.edu.cn }

Abstract

Scientific texts domain keyword is one of the basic elements of the text high-level semantics acquisition, domain ontology building and the knowledge representation in semantic grid, knowledge grid and e-science environment. It is also the indispensable foundation and prerequisite work of Web scientific texts automatic classification, clustering and personalized services. TFIDF based TDDF formula is proposed to extract scientific texts domain keyword. The experiments proved that TDDF formula extracting texts domain keyword is superior to the classic TFIDF formula does. Above discussions and achievements can provide certain support not only for the establishment of semantic grid, knowledge grid and e-science environment, but also for the Web knowledge acquisition, representation and text information retrieval and so on.

Keywords: Domain keyword, Knowledge Acquisition, Semantic Grid, Knowledge Grid

1. Introduction

Texts domain keyword is one of the basic elements of the scientific texts knowledge representation, high-level semantics acquisition and text domain ontology building. It affects the precision and the quality of scientific texts knowledge acquisition, representation and the establishment of scientific texts domain ontology. The keywords of single scientific text are difficult to accurately reflect text domain knowledge and user research interests, which will cause the Web more difficult to provide high-quality personalized services for researchers.

Firstly, the significant achievements have been acquired for the research of single text keyword extraction [1-7]. However, as e-science, semantic grid, knowledge grid and the enormous scientific e-texts emergence, people need to extract multi-texts keywords to reflect the domain knowledge of texts, in order to provide high-quality personalized services for researchers. Secondly, the domain keyword are not only the indispensable foundation and prerequisite work of user's behaviors based scientific texts automatically clustering, classification and personalization services, but also the basic elements of text representation and domain ontology building in semantic grid, knowledge grid and e-science environment[8][9]. Lastly, the domain keyword extracting precision directly affects the services of the Web scientific resources.

2 Related Works

The main methods of the keyword extraction are TFIDF (Term Frequency Inverse Document Frequency) [1], mutual information [2] [3], information gain[4], relevancy score [5], chi-square [6], NGL coefficient [7] and odds ratio [2] and so on. TFIDF and mutual information are the mainstream methods.

The significant effect has been proved in the practical applications using TFIDF formula to acquire single text keyword [1]. Mutual information is commonly used in statistical language models to evaluate the correlated degree between strings [2] [3]. Bigger mutual information between strings indicates the stronger correlation in the viewpoint of statistics. But the small mutual information does not always means that there is weaker correlation between strings and in computing the string requires minimum number. So in this paper we use TFIDF formula as the main method to discuss and compare.

In our analysis usually about 15 texts given by researcher are the greatest patience for building his own interests. So for the establishment of researcher's interests we select the texts around 15 files.

3 TDDF Formula Extracting Domain Keyword

3.1 TDDF Formula

For the sake of discussion, we give the following definitions:

* Research work is support by the National Science Foundation of China (grants 60402016), the Great Research Project of National Science Foundation of China (grants 90612010), the National Basic Research Program of China (973 project no. 2003CB317000) and the Special Research Foundation of Shanghai Science and Education Committee.

Definition 1: (Document Frequency)
The number of texts that word *t* occurrences in the texts set *D* is called document frequency.
Definition 2: (Document Domain Frequency m_t)
The number of texts that word *t* occurrences in the same domain texts set *D* is called document domain frequency.
Definition 3: (Word Common Possession Rate $c = m_t / M$)
The rate of the document domain frequency m_t and the text numbers *M* of the domain texts set *D* is called word common possession rate.

c reflecting the possibility of word *t* becomes the domain keyword. Higher *c* means that the word *t* becomes keyword with greater possibility, vice versa.

Although TFIDF formula extracts single text keyword very well, it is not suitable for extracting domain keyword. Figure 1 shows the relation between single text keyword and domain keyword. For example, a researcher whose interest is grid computing providing 7 texts, the frequency of words *t1* and *t2* see Table 1. If we extract the researcher interests according to TFIDF formula, from Table 1 we can see that "SARS" as a domain keyword's possibility more than "Architecture". Why is this happening? TFIDF considers word *t* occurrence only in unrelated texts, but in domain keyword extraction word *t* occurrence considers not only in unrelated texts, but also in related texts as well as word common possession rate *c*. So we propose TDDF (TFIDF based Document Domain Frequency) formula for the domain keyword extraction. TDDF as follows:

Table 1: Word *t1* and *t2* occurrence in text set D

	$\vec{d}_1$	$\vec{d}_2$	$\vec{d}_3$	$\vec{d}_4$	$\vec{d}_5$	$\vec{d}_6$	$\vec{d}_7$	Total Frequency	The possibility of word *t* because key- word judged by TFIDF
t1 （SARS）	22	0	0	0	15	0	0	37	higher
t2 （Architecture）	3	5	3	5	6	1	2	25	low

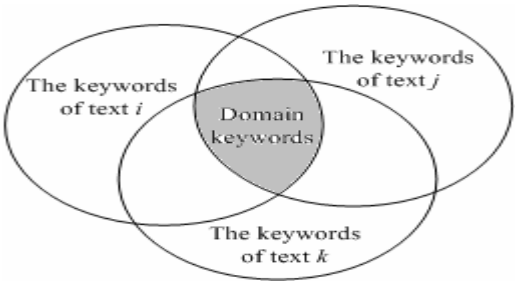

Figure 1: The relation between single text keyword and domain keyword

$$W(t,D,M) = \frac{(W^t_{\vec{d}_i}(t,\vec{d}_i)+0.01)\times\beta\frac{m_t}{M}}{\sqrt{\sum_i^M\left[(W^t_{\vec{d}_i}(t,\vec{d}_i)+0.01)\times\beta\frac{m_t}{M}\right]^2}} \quad (1)$$

where $W^t_{\vec{d}_i}(t,\vec{d}_i) = \frac{tf(t,\vec{d}_i)\times\log(N/n_t^{\vec{d}_i}+0.01)}{\sqrt{\sum_{t\in\vec{d}_i}\left[tf(t,\vec{d}_i)\times\log(N/n_t^{\vec{d}_i}+0.01)\right]^2}}$ is the weight of word *t* in text $\vec{d}_i$ (that is TFIDF formula); $tf(t,\vec{d}_i)$ is the frequency of word *t* in text $\vec{d}_i$; *D* is the texts set that belongs to the same domain; *M* is the texts number of texts set *D*; *N* is the number of unrelated texts; $n_t^{\vec{d}_i}$ is the document frequency in unrelated texts set. β is the impact degree of the document frequency on domain keyword; m_t is the document domain frequency; the 0.01 in $(W^t_{\vec{d}_i}(t,\vec{d}_i)+0.01)$ is to prevent the word *t* frequency from becoming zero in $\vec{d}_i$ which causing the denominator is also zero; denominator for returning a standardization factor.

If β is bigger enough in (1), that means the document frequency impact on the keywords extracting is very significant, formula (1) can be represented as follows:

$$W(t,D,M) = \begin{cases} \frac{(W^t_{\vec{d}_i}(t,\vec{d}_i)+0.01)}{\sqrt{\sum_i^M\left[(W^t_{\vec{d}_i}(t,\vec{d}_i)+0.01)\right]^2}} & if \quad m_t/M >= c \\ 0 & if \quad m_t/M < c \end{cases} \quad (2)$$

(2) is called TDDF formula with fixed word common possession rate *c*.
By analysis (1) and (2) - we found that:

1) If β =1 and *c*=1 in formula (1) and (2), TDDF formula retrogress to TFIDF;
2) Use formula (1) to extract domain keyword, the number of keywords is determined by the user; the quality of the extracted keywords is determined by the quantity of extracted keywords;
3) Use formula (2) to extract domain keyword, the quantity and quality of extracted keyword are determined by *c*; once *c* is determined the number of keywords will be determined automatically;
4) If *c* is higher, the number of the extracted keywords is lower and the quality is higher. Few domain keyword extracted by formula (2) will cause the texts domain knowledge do not covered.

3.2 The Optimal Parameters Determined of TFIDF formula

We found through experiments on the keyword extraction that *N* has greater impact on the quality of the keywords extraction. In order to better compare the results with TDDF formula's extracted ones, we need to investigate the parameter *N* existed in TFIDF. We hope to compare the keywords extracted by TDDF with the keywords extracted by the optimal TFIDF, which will make TDDF formula credible.

3.2.1 The Number of the Unrelated Texts Impact on the Single Texts Keywords Extraction

We randomly select 26 papers from the proceeding of the international conference on Grid and Cooperative Computing (GCC2003). 40 keywords are extracted from the above single text with 5, 10, 20 and 30 unrelated files respectively. The experiment results see Figure 2. We know from this figure that we should select around 10 unrelated files for the single text keyword extraction. If many unrelated texts are selected, the extraction accuracy is difficult to improve, and even decline. The reason for this result is the information of the unrelated files may submerge the single file's information. We can also see from this figure that the extraction accuracies of 1s288 and 12s110 are very low. Through inspecting original papers, we found that many sections in the extracted papers are the formula descriptions and inferences which cause the extraction accuracies are low.

3.2.2 The Number of the Unrelated Texts Impact on the Domain Keyword Extraction

We use 5, 10, 20 and 30 unrelated files to extract 500 domain keywords respectively. In order to analyses the correct rate of the extracted keywords, we divide the 500 extracted keywords into 1~50, 1~100, 1~150, 1~200, 1~300, 1~400, 1~500 respectively. Four experiments are made to extract domain keyword. The specific texts are the proceeding of GCC2003's session 1, session 2, session 3 and session 4 respectively. Figure 3 is the experimental results. We can see from this Figure that with the increase of the number of unrelated texts, the keyword extraction precision also increases. If we choose 20 and 30 unrelated texts to extract domain keyword respectively, the difference of the accuracies of extracted keyword is within 5%. Therefore, if conditions are permissive, we

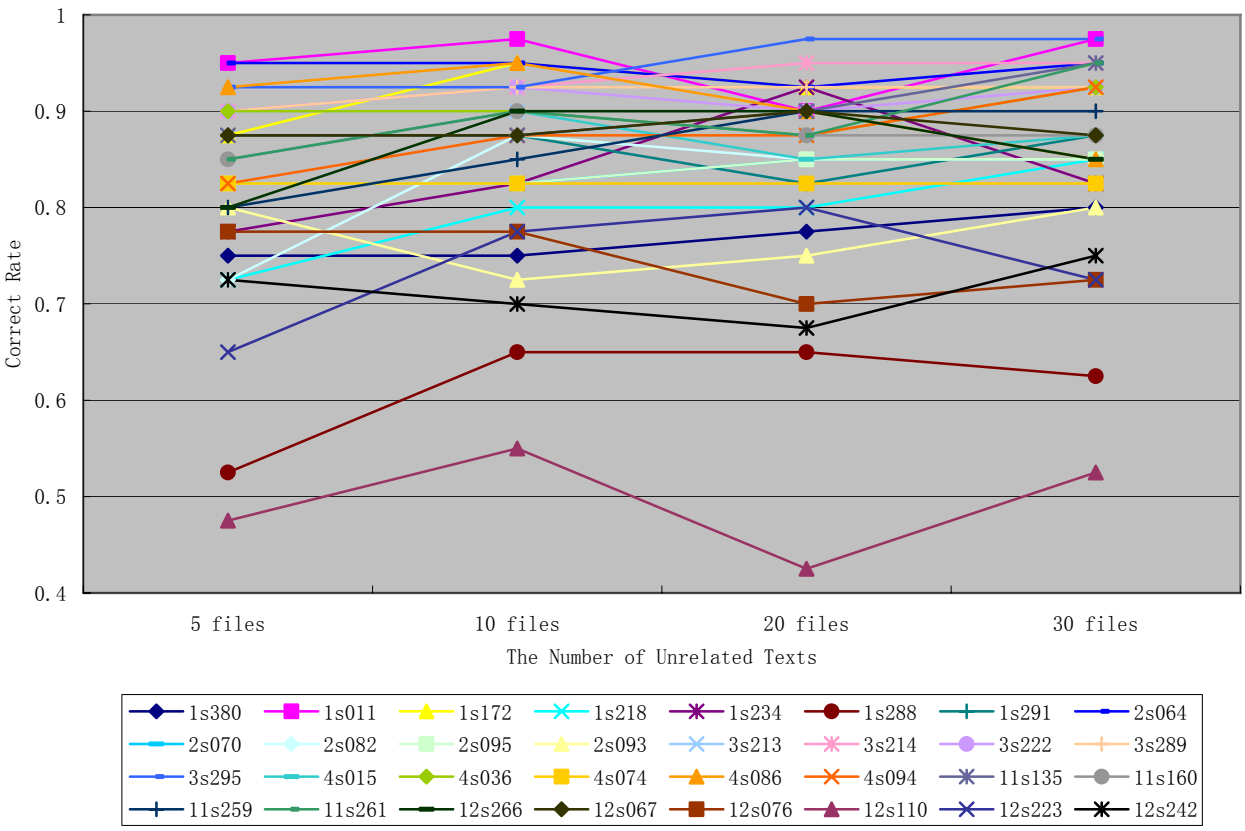

Figure 2: The number of the unrelated texts impact on the single text keyword extraction

(1s380 in this figure donates the NO. 380 paper in the session 1 of GCC2003's CDROM, the same below)

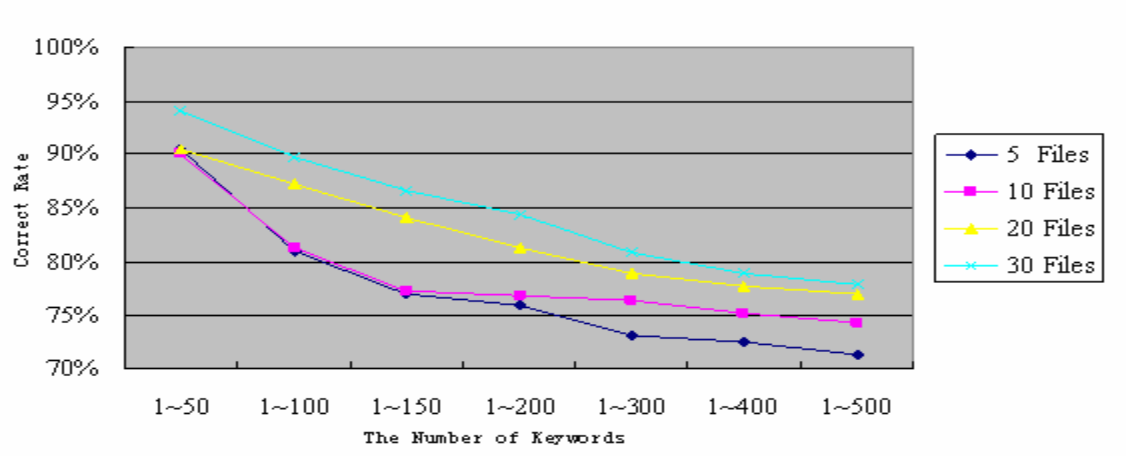

Figure 3: The number of the unrelated texts impact on the domain keyword extraction

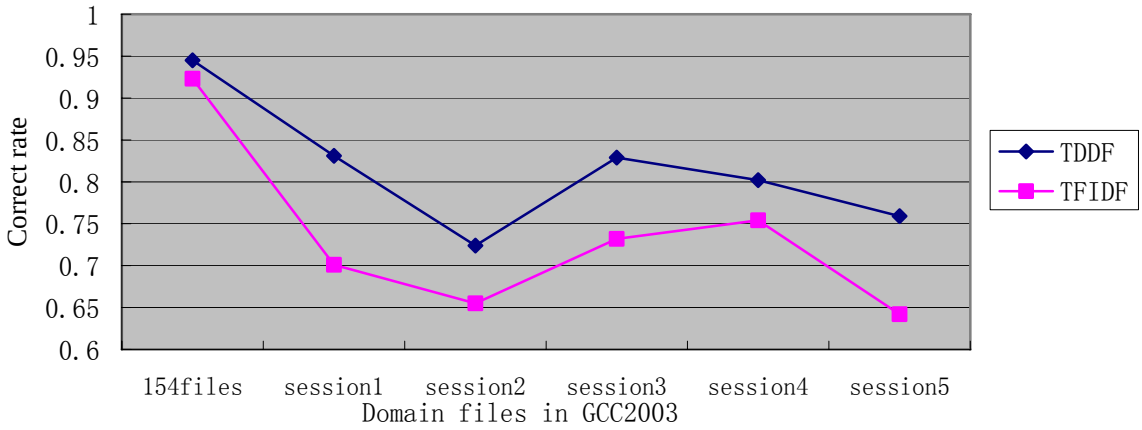

Figure 4: The results of TDDF based domain keyword extraction comparison with optimal TFIDF
（session 1 contains 15 scientific texts; session 2, session 3 and session 5 contain 16 scientific texts; session 4 contains 17 scientific texts and the all can handle papers in the proceeding of GCC2003 are 154）

should choose more unrelated texts. If conditions are restrictive (such as insufficient number of unrelated text, or real-time requirements), 20 unrelated texts are the moderate texts to extract domain keyword. Figure 3 also shows that with the increase of the number of the extracted keywords, the correct rate of keyword extraction has continued to decline.

3.3 Using TDDF to Extract Domain Keyword and Comparison with TFIDF

We use session1~session5 and the can handle papers in the proceeding of GCC2003 as the experimental database. Domain keywords are extracted by TDDF and optimal TFIDF from the above texts sets respectively. The same number of extracted keywords is compared; the results are show in Figure4. We can see from Figure 4 that the TDDF based domain keyword correct rate is higher than TFIDF does. The pseudo-keywords, such as "SARS" and "GIS", disappear in the extracted keywords set of session1.

3.4 Word Common Possession Rate Impact on the Extraction Performance

In formula (1) and (2), the significant impact on $W(t,D,M)$ is the word common possession rate c. To study c how to impact on the extraction performance of domain keyword, different c values are used. We use the experimental database as subsection 3.3. Six domain keyword extracted experiments are made. Different c values impact on the number of extracted domain keyword see Figure 5. The correct rates of extracted domain keyword see Figure 6. In Figure 5 and Figure 6, the word common possession rate c values are 0.5, 0.4, 0.3, 0.25, 0.2, and 0.18 respectively. As we can see from Figure 5~6, with the decrease of word common possession rate c, the number of the

extracted keywords increases, but the correct rates become lower. If there are few domain keywords, the domain knowledge is hard to provide the necessary basic elements for building text domain ontology and acquisition the high-level semantics of scientific text. Therefore, there exists a balance between the correct rate and the number of extracted keywords. If we want to automatically pre-process the scientific texts (such as for the scientific texts rough automatic classification), we can choose higher *c* value. Higher *c* value (> 0.4) implies that extraction of domain keyword is less, but the correct rate is higher. So it can be applied to automatic texts pre-processing. If we need to build text domain ontology or analysis text semantics, we need to choose the smaller *c* value (0.2~0.4) in order to gain more domain keyword.

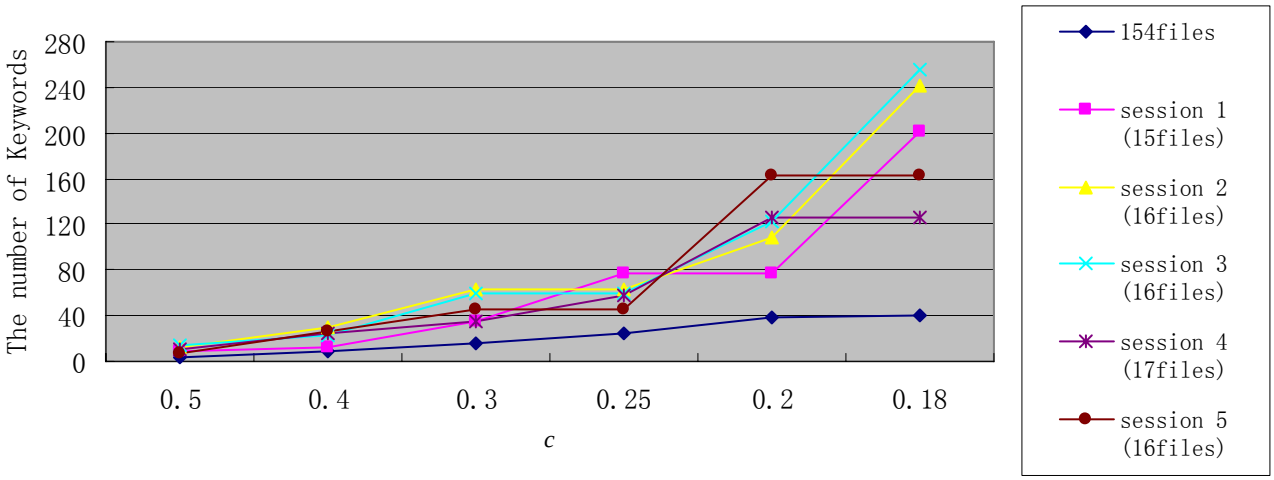

Figure 5: Different values of the word common possession rate *c* impact on the number of extracted domain keyword

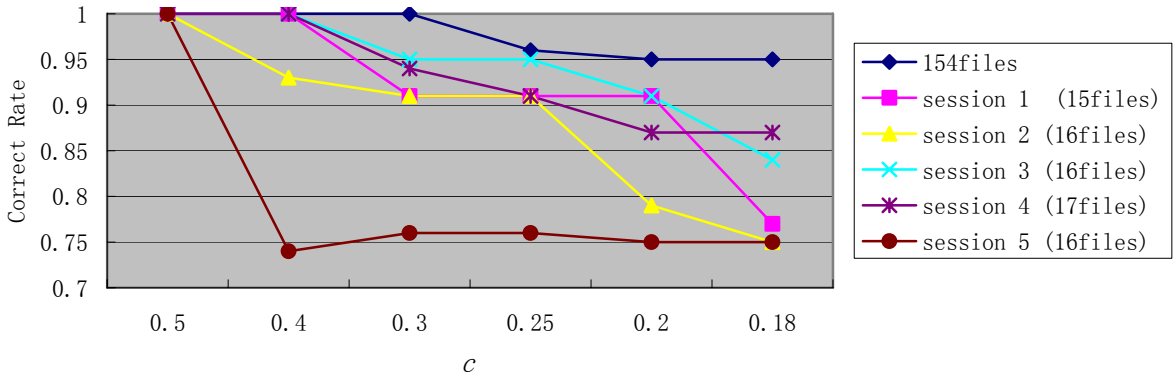

Figure 6: Different values of the word common possession rate *c* impact on the correct rate of the extracted domain keyword

Table 2: Use different word common possession rate *c* to extract domain keyword comparison with optimal TFIDF method

TFIDF			TDDF					
Extraction rate	Extraction volume	Correct rate	*c*=0.5	*c*=0.4	*c*=0.3	*c*=0.25	*c*=0.2	*c*=0.18
0.001	26	86%	100%	100%	96%	96%	96%	85%
0.002	52	82%				94%	94%	90%
0.003	78	79%				94%	94%	91%
0.004	104	78%						88%
0.005	129	75%						90%

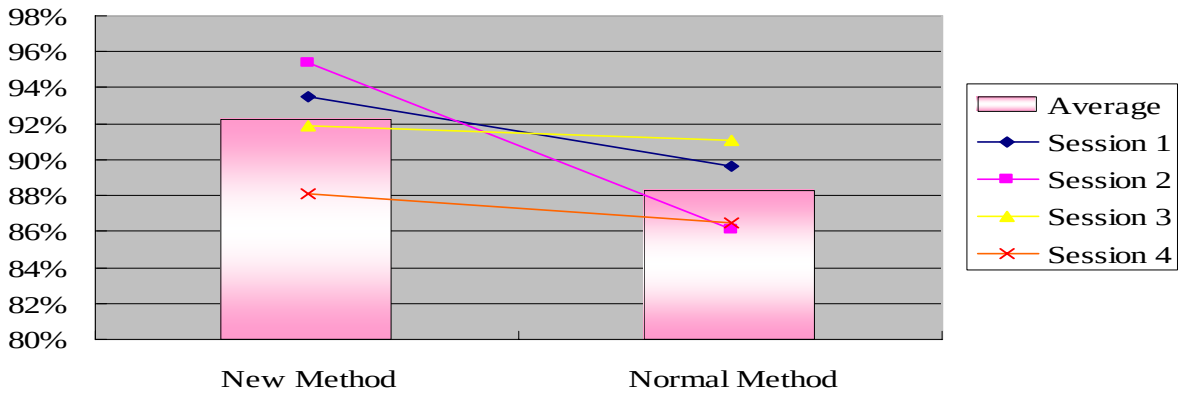

Figure 7: The domain keyword correct rate using TDDF formula with fixed word common possession rate comparison with optimal TFIDF does

3.4 Domain Keyword Extraction with Fixed Word Common Possession Rate and its Comparison with TFIDF

We choose session1~session 4 in the proceeding of GCC2003 as the texts sets of domain keyword extraction, use formula (2) with fixed word common possession rate *c* and optimal TFIDF to extract domain keyword respectively. The correct rates see Figure 7. Table 2 is the specific experimental results of session1. We can see from Figure 7 and Table 2, the correct rate that using TDDF with fixed *c* is higher than the optimal TFIDF to extract the domain keyword.

4 Conclusions

The domain keyword of scientific texts is one of the basic elements of the text domain ontology building, high-level semantics acquisition, and text knowledge representation in semantic grid, knowledge grid and e-science environment. It is also the indispensable foundation and prerequisite work of the Web based scientific texts automatic classification, clustering and personalized services. The number of text domain keyword and the correct rate directly affect the quality of Web resources services.

The experiments show that the proposed TDDF formula can extract multi-texts' domain keyword more effectively than optimal TFIDF. The quantity and quality of domain keyword can be flexibly controlled by word common possession rate *c*.

TDDF formula is only for the extraction of domain keyword and can not be used to extract single text keyword.

References

[1]. G. Salton, C.Buckley. Term-weighting approaches in automatic text retrieval. Information Process. Man. 24, 5, 513–523, 1988.

[2]. S. Fabrizio. Machine Learning in Automated Text Categorization, ACM Computing Surveys, Vol. 34, No. 1, March 2002.

[3]. W. Lam, K. F. Low et.al. Using a Bayesian network induction approach for text categorization. In Proceeding of the 15th International Joint Conference on Artificial Intelligence (Nagoya, Japan, 1997), 745–750.

[4]. L. S. Larkey. Automatic Essay Grading using Text Categorization Techniques. In Proceeding of the 21st ACM International Conference on Research and Development in Information Retrieval (Melbourne, Australia, 1998), 90–95.

[5]. E. D.Wiener, J. O. Pedersen, et.al.. A Neural Network Approach to Topic Spotting. In Proceeding of the 4th Annual Symposium on Document Analysis and Information Retrieval (Las Vegas, NV, 1995), 317–332.

[6]. M. F. Caropreso, S. Matwin, et.al. A Learner-Independent Evaluation of the Usefulness of Statistical Phrases for Automated Text Categorization. In Text Databases and Document Management: Theory and Practice, A. G. Chin, ed. Idea Group Publishing, Hershey, PA, 78–102.

[7]. H. T.Ng, W. B.Goh, et.al. Feature Selection, Perceptron Learning, and a Usability Case Study for Text Categorization. In Proceeding of the 20th ACM International Conference on Research and Development in Information Retrieval (Philadelphia, PA, 1997), 67–73.

[8]. H.Zhuge. The Knowledge Grid, World Scientific Publishing Co., Singapore, 2004.

[9]. H.Zhuge. China's E-Science Knowledge Grid Environment, IEEE Intelligent Systems, 19 (1) (2004) 13-17.