http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 8 (7): 1033-1038, 2009
ISSN 1812-5638

A Study on Unified Term Co-Occurrence Model

Qiao Ya-Nan, Qi Yong and Hou Di
Department of Computer Science and Engineering, Xi'an Jiaotong University, Xi'an, China

Abstract: In order to improve the comprehensive performance and expand the scope of application of traditional term co-occurrence models, this study proposes Unified Term Co-occurrence Model. It unites two types of traditional term co-occurrence models (which are called mother models of Unified Term Co-occurrence Model) and could make a series of compound models of them for various research conditions. Precision and stability are two key performance indicators of term co-occurrence models. The first type of traditional term co-occurrence models are good at stability and the second type of traditional term co-occurrence models are good at precision. The experimental results in this study confirm that precision and stability of Unified Term Co-occurrence Model (UTCM) are not lower than both of its mother models. Then, a new measure for comprehensive performance is proposed and Unified Term Co-occurrence Model (UTCM) achieves better comprehensive performance compared with both of its mother models. Researchers can use unified term co-occurrence model instead of traditional models as an important tool to get more rational experimental results in relative research fields such as information retrieval, natural linguistic processing and computational linguistics, etc.

Key words: Information retrieval, term co-occurrence, co-occurrence window

INTRODUCTION

Term co-occurrence is one of important subjects of information retrieval research. Co-occurrence terms describe the context and language circumstance of the core terms. Term co-occurrence model is based on this hypothesis: in a sufficiently large corpus, if some terms often appear in the same co-occurrence window, then we can consider there are semantic relations between them; furthermore, the higher co-occurrence frequency, the closer semantic relations. So, terms' correlation can be quantitatively compared using term co-occurrence models. Now term co-occurrence has applied to many research fields such as information retrieval, natural linguistic processing and computational linguistics, etc.

Co-occurrence window, is one of the central concepts in term co-occurrence research. Easy to conclude, the reasonable size of co-occurrence window should be determined based on the average length of sentences, but it is instable in various language circumstances. In spoken language, it will be very short and in research study, it perhaps is very long. In English term co-occurrence research, the size of co-occurrence window is usually set to 20~40. Researchers have to do experiments for many times to verify the best window size for different test collection and language circumstance.

Therefore, the performance of term co-occurrence models should be estimated by at least two indicators: precision and stability. Firstly, the correlation calculated in co-occurrence models should be close to the estimation by human beings. Secondly, the correlation should not be very sensitive for the change of co-occurrence window size.

In this research field, researchers usually analyzed term co-occurrence models from two points of views: (1) How to evaluate the correlation of two terms in a single co-occurrence window? (2) How to evaluate the correlation of two terms based on all co-occurrence windows in a document even in a corpus? For these two problems, researchers proposed many term co-occurrence models, but these models often only concerned one of two problems above and simplified even ignored the other problem. The performance and application scope of these existing term co-occurrence models were impacted on the absence of considering both two problems. We call the model for the first problem the first model and call the model for the second problem the second model.

The simplest first model is constant model, in which terms' correlation is a constant in the same co-occurrence window without considering the position of them (Yarowsky, 1993). This constant model over-simplify the problems obviously, so researchers proposes some

Corresponding Author: Qiao Ya-Nan, Department of Computer Science, School of Electronic and Information Engineering,
Xi'an Jiaotong University, No. 28 Xianning West Road, Xi'an Shaanxi, People's Republic of China
Tel: 0086-29-82668645, 0086-13572103109

improving models, such as polynomial descending model (Lu and Bai, 2001), exponential descending model (Gao *et al*., 2002) and term field model (Qiao *et al*., 2007) etc. Except term field model, all of these existing models are descending model, namely the correlation of two terms in a co-occurrence window should be decreased when their distance increased. But in term field model, correlation is almost a constant when two terms in a small distance and when the distance is more than a threshold, the correlation will be decreased with their distance increased. Compared with descending model, term field model is more in accord with common senses in linguistics practice.

The simplest second model is frequency model, in which the correlation of two terms only depends on the number of co-occurrence windows. From frequency model, researchers proposed many improving models such as COSINE (Bayardo *et al*., 2007), DICE (Lewis *et al*., 2006), TANIMOTO (Cha *et al*., 2006), Z-Score, T-Score (Broda and Piasecki, 2008) and mutual information (Lin *et al*., 2008), etc. and these models are all based on some important statistics in information theory.

Two simplest models (constant model and frequency model) for the first problem and the second problem are one model indeed. In this model, the correlation of term pair is the number of co-occurrence windows and the correlation of term pair in each window is a constant 1. So, we can conclude, every first model includes a frequency model impliedly and every second model includes a constant model impliedly, too. This is the basis of the discussion in the next section in fact.

This study firstly analyzes existing term co-occurrence models and divides them into two types and then proposes Unified Term Co-occurrence Model (UTCM). This Model unites two types of traditional term co-occurrence models (we call them mother models) and could make a series of compound models of them for various conditions. The experimental results show that the compound models based on UTCM balances the advantages and shortcomings of their mother models and achieves better comprehensive performance.

MATERIALS AND METHODS

The study of UTCM was a part of the Doctor Authorizes Point Fund Project of China (No. 20060698018). This project was conducted in Xi'an Jiaotong University since March 2006 and expected to accomplish in December 2009. In this section, we will introduce the main framework of UTCM.

We call term a and term b in a document D Term Pair [a, b]. Let there be m term co-occurrence windows in D, denoted by [a, b]1, [a, b]2... [a, b] m-1, then the correlations of term pair [a, b] can be denoted by r ([a, b]1), r ([a, b]2)...r ([a, b]m-1), r ([a, b]m). How to evaluate r ([a, b]i) is the first term co-occurrence problem proposed in the introduction.

To term pair [a, b] in document D, the correlation of this term pair for whole the document can be denoted by R_D ([a,b]). How to evaluate R_D ([a,b]) is the second term co-occurrence problem proposed in the introduction. It does not utilize the concept of co-occurrence window explicitly and utilize the simplified concept the frequency of co-occurrence of a and binstead. That is:

$$R_D\,([a,b)]=G(a,b,f(a,b) \tag{1}$$

where, f (a, b) is the frequency of co-occurrence of a and b, namely R_D ([a,b]) depends on the statistics of a and b itself and the frequency of co-occurrence of a and b.

Then, introduce the concept of co-occurrence window into the second term co-occurrence problem using the mode of the first term co-occurrence problem,

$$R_D\,([a,b])=G(a,b,\int_D \frac{r([a,b]\theta)}{E(r([a,b]))}g(\theta)d\theta) \tag{2}$$

That is the correlation expression of term pair in Unified Term Co-occurrence Model. Where, θ is a certain co-occurrence window of [a, b] in D; r ([a, b]θ) is the correlation of term pair [a, b] in window θ; E (r ([a, b])) is the mathematical expectation of r ([a, b]θ), which will normalize the correlation; g (θ) is the weigh of window θ, $0\le g\,(\theta)\le 1$.

Specially, if we treat all of the windows equally, namely g (θ)=1, then

$$\begin{aligned} R_D\,([a,b])&=G(a,b,\int_D \frac{r([a,b]\theta)}{E(r([a,b]))}g(\theta)\,d\theta) \\ &= G(a,b,\sum_{\theta=1}^{n}\frac{r([a,b]\theta)}{E(r([a,b]))} \end{aligned} \tag{3}$$

where, n is the number of co-occurrence window in whole document.

Let r ([a, b]θ) in Eq. (3) and then E (r ([a, b])=1 obviously, so:

$$\begin{aligned} R_D\,([a,b]) &= G(a,b,\sum_{\theta=1}^{n}\frac{r([a,b]\theta)}{E(r([a,b]))}) \\ &= G(a,b,n)=G(a,b,f(a,b)) \end{aligned}$$

This is the correlation expression of the second term co-occurrence problems. In other words, the second term co-occurrence problems is the special case of UTCM when simplify the evaluation of r ([a, b]θ] to Constant Model.

In the framework of whole term co-occurrence models, the first term co-occurrence problems concerns the internal structure of co-occurrence window and it is microcosmic; the second term co-occurrence problems concerns the relations between co-occurrence windows and it is macroscopic. These two problems are united in UTCM, which regards the first term co-occurrence problems as the cell, the second term co-occurrence problems as the tissue consists of many cells. So we can utilize the existing research results of these two problems and unite two types of traditional term co-occurrence models to make a series of compound models of them for various conditions.

For example, choose COSINE model as the second term co-occurrence model, and its expression (Peat and Willett, 1993) is:

$$\text{COSINE}(X,Y) = \frac{F(X,Y)}{\sqrt{F(X)\times F(Y)}} \tag{4}$$

where, F (X, Y) is the frequency of term pair X and Y in a co-occurrence window; F (X) is the frequency of term X and F (Y) is the frequency of term Y.

Then, choose linear descending model as the first second term co-occurrence model and treat all of the windows equally, that is:

$$\text{SIM}_{\theta}(X,Y) = 1 - \frac{D\theta(X,Y)}{W} \tag{5}$$

where, SIM θ (X, Y) is the correlation of X and Y in co-occurrence window θ; Dθ (X, Y) is the distance of X and Y in θ; W is the size of co-occurrence window.

Obviously, the range of SIM $_{\theta}$(X, Y) in linear descending model is (0, 1), so the mathematical expectation is 0.5.

Finally, pack Eq. 4 and 5 based on Eq. 3, that is:

$$\text{SIM}(X,Y) = \frac{\sum_{\theta=1}^{N} \frac{\text{SIM}\,\theta(X,Y)}{0.5}}{\sqrt{F(X)\times F(Y)}} = \frac{2}{\sqrt{F(X)\times F(Y)}} \sum_{\theta=1}^{N} \text{SIM}\,\theta(X,Y) \tag{6}$$

Equation 6 is the UTCM compounded from COSINE model and linear descending model.

RESULTS

We choose Reuters RCV1 (Lewis *et al*., 2004) as the test collection of corpus. This is a large collection of reuters news stories for use in research and development of natural language processing, information retrieval and machine learning systems etc., which is significantly larger than the older, well-known Reuters-21578 collection heavily used in the text classification community. Reuters RCV1 contains about 810,000 reuters, english language News stories and requires about 2.5 GB for storage of the uncompressed files.

The experimental platform is The lemur toolkit for language modeling and information retrieval (Ogilvie and Callan, 2001). The newest version of Lemur is 4.9.

We will analyze the validity of UTCM by two statistics: precision and stability, so designed two groups of experiments for three intentions: (1) validating the precision of UTCM is not lower than both of mother models, (2) validating the stability of UTCM is not lower than both of mother models and (3) validating the comprehensive performance of UTCM is higher than both of mother models.

For a criterion of the correlation of term pairs, we choose WordSimilarity-353 (Finkelstein *et al*., 2002) as the test collection of terms. WordSimilarity-353 contains 353 pairs of terms and their correlations estimated by 13~16 humans.

The choices of term co-occurrence models and their expressions are shown in Table 1.

Two groups of experiments are: (1) To all term pairs of WordSimilarity-353, use 6 original models and 6 compound models in Table 1 to evaluate their correlations (the co-occurrence window sizes are 20, 30 and 40) and measure the difference between the results and the estimation of humans by Normalizing Discounted Cumulative Gain (NDCG), is a measure of effectiveness of a Web search engine algorithm or related applications, often used in information retrieval, (2) To 10 term pairs of WordSimilarity-353 selected randomly, use all 12 models to evaluate the correlation of each term pair for 21 times (the co-occurrence window sizes are from 20 to 40, step by 1) and measure the average dispersion of the results by Coefficient of Variance (CV), the quotient of

Table 1: The choices of term co-occurrence models

Model No.	Term co-occurrence models	Type
1	Constant/frequency	First/Second
2	Linear descending	First
3	Exponential descending	First
4	Term field	First
5	COSINE	Second
6	DICE	Second
7	COSINE+Linear descending	UTCM
8	DICE+Linear descending	UTCM
9	COSINE+Exponential descending	UTCM
10	DICE+Exponential descending	UTCM
11	COSINE+Term field	UTCM
12	DICE+Term field	UTCM

Table 2: The experimental results of group 1

Model No.	NDCG
1	0.9355
2	0.9362
3	0.9422
4	0.9350
5	0.9593
6	0.9586
7	0.9597
8	0.9588
9	0.9576
11	0.9568
12	0.9560

Table 3: The experimental results of group 2

Model No.	CV
1	1.845404×10^{-1}
2	1.399065×10^{-1}
3	2.362202×10^{-16}
4	3.202454×10^{-2}
5	1.845403×10^{-1}
6	1.845403×10^{-1}
7	1.399065×10^{-1}
8	1.399064×10^{-1}
9	2.047626×10^{-16}
10	2.693335×10^{-16}
11	3.202450×10^{-2}
12	3.202438×10^{-2}

standard deviation and mean value. The definitions of NDCG (Croft *et al*., 2009) and CV are:

$$\mathrm{DCG} = \mathrm{rel}_1 + \sum_{i=2}^{p} \frac{\mathrm{rel}_i}{\log_2 i}$$

$$\mathrm{NDCG} = \frac{\mathrm{DCG}}{\mathrm{IDCG}}$$

and

$$\mathrm{CV} = \frac{1}{\mu}\sigma = \frac{1}{\mu}\sqrt{\frac{I}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

where, DCG is Discounted Cumulative Gain; rel_i is the graded relevance of the result at position i, IDCG is the ideal DCG (the max value), μ is the mean value; σis the standard deviation.

The experimental results of group 1 and group 2 are shown in Table 2 and 3.

DISCUSSION

In this section, we will analyze the experimental results and examine whether the UTCM can achieve better performance compared with previous models.

As we know, this study is the first to propose a united framework to combine the existing term co-occurrence models, so we should compare the performance with non-compound models directly. We will compare UTCM with previous traditional models from two sides: (1) single UTCM with its mother models and (2) whole UTCM with whole previous traditional models.

Table 4: The Comprehensive Performance of All models

Model No.	CP
1	0.7993
2	0.8191
3	0.9358
4	0.8681
5	0.8230
6	0.8223
7	0.8426
8	0.8417
9	0.9512
10	0.9502
11	0.8899
12	0.8891

Choose 3 first models (Linear Descending (Lu and Bai, 2001), exponential descending (Gao *et al*., 2002) and Term Field (Qiao *et al*., 2007)) and 2 second models (COSINE (Bayardo *et al*., 2007) and DICE (Lewis *et al*., 2006)) as baseline models. These 5 models and 6 UTCM based on them are shown in Table 1. Note that Model No. in Table 1 is corresponding to Model No. in Table 2, 3 and 4.

As shown in Table 2, NDCG of UTCM approximate the higher one of its mother models; and in Table 3, CV of UTCM are very close to the lower one of its mother models. The experimental results show that UTCM have both advantages of their mother models. The precisions of the second models are higher than the first models and the precisions of UTCM are close to the second model of their mother models. Likewise, the stabilities of the first models are higher than the second models (the stabilities are inversely proportional to CV shown in Table 3) and the stabilities of UTCM are close to the first model of their mother models.

Why does UTCM have such a great advantage? The first model only concerns isolatable co-occurrence window without overall considerations (such as ignoring the frequency of single term in whole document), so its precision is affected. Moreover, whether two terms are co-occurrence in the second model only based on the number of co-occurrence window, so it is instable for the change of co-occurrence window size predicatively. We can see the shortcoming of first model is the advantage of the second model and vice versa.

UTCM inherits and expands the idea of previous findings about term co-occurrence models and abstracts term co-occurrence from another perspective. UTCM can combine the advantages of previous models which divided into the first model and the second model in this study, so UTCM can be used in every occasion using previous term co-occurrence models before.

In order to illustrate this conclusion further, define the measure of Comprehensive Performance (CP) as:

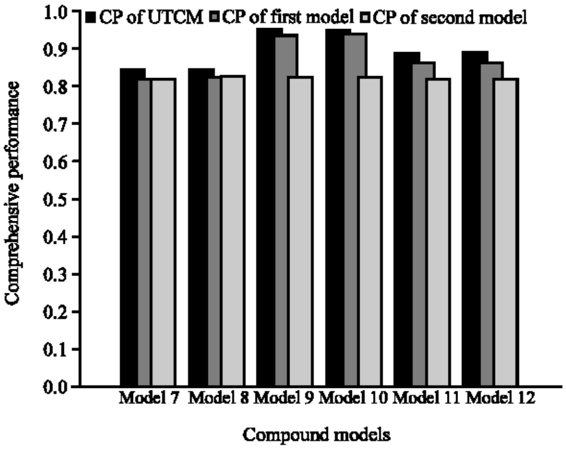

Fig. 1: The comparison of comprehensive performance of all models

$$CP = \lambda NDCG + (1-\lambda)\frac{1}{\log(CV)} \quad (7)$$

where, λ is the weigh of precision, $1>CV>0$. We set $\lambda = 0.90$ in this study. Equation 7 is significative when CV is lower than about 0.50. The CP of all models are shown in Table 4.

For easy comparison, the CP of UTCM and their mother models are showed in Fig. 1, where left bar in every model denotes CP of UTCM; middle bar denotes CP of the first model of mother models; right bar denotes CP of the second model of mother models. Compared with mother models, the compound models achieve better CP obviously.

For whole UTCM and whole previous traditional models, average CP of previous traditional models is 0.8537 and average CP of UTCM is 0.8941, so UTCM achieve about 4.73% improvement compared with previous findings as a whole.

In addition, model (1) in Table 1 is the first model and also the second model, so the performance of model (1) is always low not only precision but also stability. Model (1) is the simplest term co-occurrence model. The improvements in the side of first model enhance its stability and improvements in the side of second model enhance its precision. UTCM combines the first model and the second model, so it enhances both stability and precision.

Since, UTCM evaluate the correlation of term pairs more effectively, so it can almost entirely replace traditional term co-occurrence models in relative research fields. Besides these significances, the basic idea of UTCM can be extended in other situations. For example, to a certain problem, if there are some existing solutions with different focus, we can try to combine them in a suitable way and the new compound solution learns from other's strong points to offset one's weakness. Of course, there will be some trade-offs in this process inevitably.

CONCLUSION

Term co-occurrence model is an important tool to quantify terms' correlation and it has applied to many relative research fields widely such as Information retrieval, natural linguistic processing and computational linguistics, etc. In this study, UTCM is proposed to unite two types of traditional term co-occurrence models. The comprehensive performances of compound models are higher than their mother models. In the meantime, the precisions and stabilities of compound models are not lower than their mother models. So, we can use the compound models in relative research instead of their mother models to achieve more rational experimental results.

It is undeniable that, UTCM should be improved to lift its limitations. In the first place, UTCM combine only two traditional models and it cannot learn the advantages of three or more traditional models simultaneously. In the second place, UTCM replaces the frequency of term pairs in the expression of the second model by the expression of the first model, but remains the other parts unchanged simply. Finally, UTCM should be tested about computational complexity compared with traditional models. These works we will carry out in the next step of research.

ACKNOWLEDGMENTS

This work is supported by and the Doctor Authorizes Point Fund Project of China (No. 20060698018) and the National High Technology Research and Development Program of China (863 Program, No. 2006AA01Z101).

REFERENCES

Bayardo, J.R., Y. Ma and R. Srikant, 2007. Scaling up all pairs similarity search. Proceedings of the 16th International Conference on World Wide Web. Banff, Alberta, Canada, May 8-12, ACM Press, New York, pp: 131-140.

Broda, B. and M. Piasecki, 2008. Super Matrix: A general tool for lexical semantic knowledge acquisition. Proceedings of International Multiconference on Computer Science and Information Technology, Oct. 20-22, IEEE Computer Society, Wisia, Germany, pp: 345-352.

Cha, S.H., C. Tappert and S. Yoon, 2006. Enhancing binary feature vector similarity measures. J. Pattern. Recognit. Res., 1: 63-77.

Croft, B., D. Metzler and T. Strohman, 2009. Search Engines: Information Retrieval in Practice. 1st Edn., Addison Wesley, London, UK., ISBN: 978-0136072249.

Finkelstein, L., E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman and E. Ruppin, 2002. Placing search in context: The concept revisited. ACM Trans. Inf. Syst., 20: 116-131.

Gao, J., M. Zhou, J.Y. Nie, H. He and W. Chen, 2002. Resolving query translation ambiguity using a decaying co-occurrence model and syntactic dependence relations. Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Aug. 11-15, ACM Press, pp: 183-190.

Lewis, D.D., Y. Yang, T. Rose and F. Li, 2004. RCV1: A new benchmark collection for text categorization research. J. Mach. Learn. Res., 5: 361-397.

Lewis, J., S. Ossowski, J. Hicks, M. Errami and H.R. Garner, 2006. Text similarity: An alternative way to search medline. Bioinformat., 22: 2298-2304.

Lin, J.F., S. Li and Y. Cai, 2008. A new collocation extraction method combining multiple association measures. Proceedings of International Conference on Machine Learning and Cybernetics, Jul. 12-15, IEEE Computer Society, Kunming, China, pp: 12-17.

Lu, S. and S. Bai, 2001. Quantitative analysis of context field in natural language processing. Chinese J. Comput., 24: 742-747.

Ogilvie, P. and J. Callan, 2001. Experiments using the lemur toolkit. Proceedings of the 10th Text Retrieval Conference, Nov. 13-16, NIST Special Publication, Gaithersburg, USA., pp: 103-108.

Peat, H.J. and P. Willett, 1993. The limitations of term co-occurrence data for query expansion in document retrieval systems. J. Am. Soc. Inf. Sci., 42: 378-383.

Qiao, Y.N., Y. Qi and H. He, 2007. The research on term field based term co-occurrence model. Proceedings of the 3rd International Conference on Semantics, Knowledge and Grid, Oct. 29-31, IEEE Computer Society, Xi'an, China, pp: 471-474.

Yarowsky, D., 1993. One sense per collocation. Proceedings of the ARPA Human Language Technology Workshop, Mar. 21-24, Morgan Kaufmann, Princeton, New Jersey, pp: 266-271.