Merging Case Relations into VSM to Improve Information Retrieval Precision

Wang Hongtao[1], Sun Maosong[1], and Liu Shaoming[2]

[1] The State Key Laboratory of Intelligent Technology and Systems,
Department of Computer Science and Technology,
Tsinghua University, Beijing 100084, China
wanght02@mails.tsinghua.edu.cn
[2] Future Technology Institute, Fuji Xerox Co. Ltd, Japan
Liu.shaoming@fujixerox.co.jp

Abstract. This paper presents an approach that merges case relations into the well-known Vector Space Model (VSM), leading to a new model named C-VSM (Case relation-based VSM). A Chinese case system with 23 case relations is established, and a Chinese Olympic news corpus of 7,662 sentences, denoted COCS, is constructed by manual annotation with these 23 case relations. We use 50 queries on COCS as a test set. Experimental results on the test set show that C-VSM outperforms W-VSM (Word-based VSM) by 3.4% on the average 11-point precision. It is worth pointing out that almost all the previous studies on semantic IR obtained no better, even worse, results than W-VSM, our work thus validates the usefulness of case relations in IR through the validation is still preliminary. The proposed model is believed to be language-independent.

1 Introduction

A majority of traditional models of information retrieval (IR) mainly make use of surface linguistic information such as words/terms. It is reasonable to expect better retrieval results if we can exploit deep linguistic information further. Previous studies of this sort have been carried out at both syntactic and semantic levels. Most of them focused on the former, because the recognition of syntactic structures is easier than that of semantic structures. Syntactic information possibly exploited in IR can be a simple syntactic relation between a pair of words, and can also be a complex structure tree. The use of simple syntactic relations in IR has found a small improvement in retrieval effectiveness (Croft, Turtle and Lewis, 1991; Hyoudo, Niimi and Ikeda, 1998). But the results of using complex structure trees are worse than keyword matching (Smeaton, O'Donnell and Kelledy, 1995).

It is natural to assume that semantic information is more useful in IR since it can capture the meaning of a sentence more precisely than syntax. Semantic information, both intra-sentential and inter-sentential, is usually represented by the so-called semantic relations between various entities involved.

Case relation is an intra-sentential semantic relation that exists between the core verb and other constituents of a sentence (Fillmore, 1968; Somers, 1987). Lewis

A. Gelbukh (Ed.): CICLing 2005, LNCS 3406, pp. 584–592, 2005.

(1984) addressed the possibility of IR based on case relation matching. Lewis' major hypothesis is that if index terms of a query and indexed terms of a document are more likely to co-occur with similar case relations, there will be a more significant similarity between the query and the document. In other words, if a document is judged to be associated with a query, the document would not only share many identical index terms with the query, but would share similar case relations between those index terms as well. Lewis just put forward this idea, without giving any experiment result. Lu (1990) proposed a simple structure tree-matching method in IR, according to the case-frame system in (Young, 1973), to formulate semantic meaning of sentences. The Experiment on a small test set demonstrated that the performance of this proposed method is worse than the vector-based keyword matching. Lu's study also suggested that the strict matching between case relations may hurt the performance of IR due to the resulting data sparseness problem. Liu (1997) incorporated the word concept, abstracted as the semantic category of a word, and the semantic role of the concept in a sentence into the Vector Space Model (VSM), taking a 2-tuple (word concept, semantic role of the concept), instead of the literal word, as the basic element of vectors. This method is named 'partial relation matching' because it is not based on full semantic structure tree (we shall continue to use this term in Section 3.1). The vector dimension can be controlled using upper-lower relations between semantic categories. The method yields an increasing in recall and a drop in precision, and almost same F-measure compared to conventional word-based VSM (W-VSM).

Inter-sentential semantic relations often exist between words beyond separate sentences in text. Khoo (2001) made an intensive study on exploring just one relation – the cause-effect relation. An algorithm is developed for recognizing cause-effect relations in text automatically. But the experiment on the Wall Street Journal corpus did not give better results than proximity-based word matching.

As can be observed, previous efforts of taking both syntactic and semantic information into consideration in IR have not reached satisfactory performance so far, and, obviously, the research concerned is very preliminary. This implies that there may be a large room of improvement for relation-based IR (in particular, for semantic relation-based IR).

This paper tries to introduce case relation into VSM, leading to a C-VSM (Case relation-based Vector Space Model). In C-VSM, the classic TF*IDF formula is adjusted by multiplying a weighting factor to each word according to its case relation in the sentence. Experiments on a test set show that the average 11-point precision of this model reaches 87.2% and outperforms the baseline, W-VSM, by 3.4%.

The rest of the paper is organized as follows: Section 2 describes a semantically annotated Chinese corpus used and the case relations defined in it, Section 3 discusses experiment-based design of the algorithm, in the context of comparing with W-VSM and the strategy of partial relation matching, and Section 4 is conclusion.

2 Semantically Annotated Corpus and Case Relations Defined

Case relations are semantic relations that hold between the core verb and other constituents in a sentence (Fillmore, 1968; Somers, 1987). For example, in the sentence *Harry loves Sally*, the case relation *experiencer* holds between *Harry* and

love, and the case relation *patient* holds between *love* and *Sally*. The verb *love* is said to assign the case relation of *experiencer* to *Harry* and the case relation of *patient* to *Sally*. Case relations can be sub-categorized into two groups, i.e., essential case relations and peripheral case relations. Essential case relations are those necessary for the verb while peripheral case relations are those optional to the verb.

Inspired by Fillmore's theory, Lin (1999) designed a Chinese case system with 22 cases. We simply adopt Lin's system with a minor expansion by adding one case particularly for the Olympic domain. As a consequence, a Chinese case system with 23 cases is established. A Chinese Olympic news corpus of 7,662 sentences, denoted COCS, is then constructed by manual annotation with these 23 case relations. Case relations defined and their distribution in COCS are listed in Table 1.

Table 1. A Chinese case system and the distribution of case relations in a semantically annotated Chinese corpus

Case Relation	Symbol	Coverage for case relations in COCS (%)
Agent (施事)	S	21.4
Experiencer (当事)	D	17
Genitive (领事)	L	0.12
Patient （受事）	O	12.6
Accusative (客事)	K	4.7
Comitative (共事)	Y	3.7
Link (系事)	X	4.3
Type (类别)	B	0.1
Object (对象)	T	2.4
Result (结果)	R	6.8
Manner (方式)	Q	3.3
Quantity (数量)	N	1.3
Scope (范围)	E	8.7
Time (时间)	H	8.8
Part (分事)	F	0.15
Benchmark (基准)	J	0.6
Instrument (工具)	I	0.03
Material (材料)	M	0
Location (位置)	P	2.8
Direction (方向)	A	0.07
Warranty (依据)	W	0.45
Cause (原因)	C	0.6
Purpose (目的)	G	0.4

To ensure the quality of the annotated corpus, a two-round annotation is performed. An sample sentence from COCS is as follows:

[S 中国/ns 选手/n 龚智超/nr]S1S2 [D 周五/t]H [D 在/p 奥运会/j 羽毛球/n
Chinese player Gong Zhichao Friday in Olympic Games badminton
女单/j 决赛/vn 中/f]E ，/w [D 以/p 2/m ：/w 0/m]Q
Women's singles final within with 2 : 0
[P 战胜/v]V1 [O 前/f 世界/n 排名/v 第一/m 的/u 丹麦/ns 名将/n
defeat former world ranking NO.1 of Denmark well-known player
马尔廷/nr]T1 ，/w [D 为/p 中国/ns 代表团/n]Y2 [P 夺得/v]V2 [O 本届/r
Camilla Martin for Chinese delegacy win this
奥运会()/j 上/f 的/u 第 1 4 ()/m 块/q 金牌/n]R2 。/w
Olympic Games in of the fifth piece Gold medal

(Gong Zhichao from China, defeated Camilla Martin, a well-known and former world ranking No.1 player from Denmark, on Friday, yielding the Women's Singles Badminton title. It is the 14th gold medal China has won in this Olympic Games.)

where: 'w/x' stands for part-of-speech x for word w, '[……]' gives a chunk in a given sentence, '[X' indicates the grammatical function of the associated chunk in the sentence (for example, '[S' means Subject), ']X#' indicates the case relation of the associated chunk to the core verb of the sentence ']V#'(for example, ']S' means Agent of V), and '#' is a sequence number for multiple sentences. The word underlined is the head of the associated chunk.

3 Experiment-Based Algorithm Design

COCS concerns news for Olympic sports. Each article in COCS is usually quite short, – on average, there are only 2-3 sentences in each article. So both query and retrieval in experiments here are based on a single sentence rather than a full article. We selected 100 sentences from COCS as queries, and hand-crafted the retrieval outputs, 635 sentences in total, for these 100 queries accordingly. Then we randomly split this data set into two equal parts: 50 queries for parameter estimation of the proposed model, and the remaining 50 queries for testing. W-VSM is regarded as the baseline throughout the experiments.

3.1 Solution 1: Partial Relation Matching

In attempting to incorporate case relations into IR, we try a solution similar to 'partial relation matching' at first. Each word and its associated case relation in a sentence constitute a 2-tuple (Note: the case relation of a word is conveyed from the case relation of the chunk containing that word), and this 2-tuple is used as an index item in vectors. For example, there exist three index items, (龚智超, S), (马尔廷, T) and (战胜, V), for the sentence "龚智超(*Gong Zhichao*) 战胜(*defeat*) 马尔廷(*Camilla Martin*)" (*Gong Zhichao defeated Camilla Martin*). Each 2-tuple is then weighted with TF*IDF where TF is the frequency of the 2-tuple in a sentence and IDF is the inverse sentence frequency of the word involved in the 2-tuple. We say that an index item (or a 2-tuple) is matched if and only if its word and case relation are matched simultaneously.

We explore two strategies in the experiment. Strategy 1 takes all of the 2-tuples in sentences as index items, whereas strategy 2 only takes 2-tuples relating to heads of chunks as index items (the others are still simply indexed as words). We compare the two strategies with the baseline, W-VSM, as illustrated in Fig. 1.

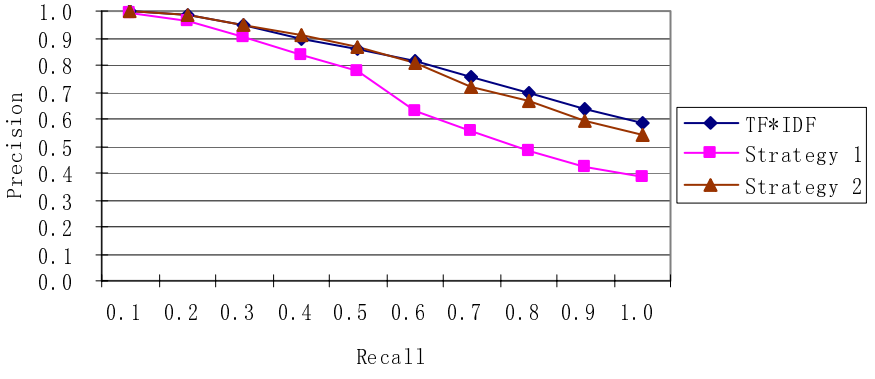

Fig. 1. Comparisons between partial relation matching and W-VSM

As can be seen, strategy 1 is poorer than strategy 2, and both strategies do not give better results than the baseline, suggesting that solution 1 may not be feasible for integration.

3.2 Solution 2: Merging Case Relations into Word Weighting

In experience gained in Section 3.1, we find that the condition for matching two 2-tuples is too strict. We thus present another possible solution: instead of using a case relation explicitly in a 2-tuple, we use it in a way of being viewed as just a weighting factor added to the traditional W-VSM model, that is, we still use a word as an index item, but re-estimate its TF*IDF weighting by multiplying a factor according to the type of the associated case relation. As stated earlier, the case relation of any word in a chunk is derived from the case relation of the chunk containing that word.

Obviously, the contributions of case relations to the meaning of a sentence are not identical. We categorize case relations into several groups in order to decrease the number of parameters to be estimated in the model. Case relations falling into the same group will be given an identical weighting factor, meanwhile those falling into different groups will be assigned distinct factors.

Agent, Experiencer and Genitive all belong to the source of an action, so we classify these 3 case relations into a group, denoted Group 1; In parallel, Patient, Result, Link, Part, Objective, Type and Accusative, the direct target of an action, are classified into another group, denoted Group 2; Verb is not a case relation, but deserves special attention in weighting, so it is treated as a separate group, denoted Group 3; Comitative and Benchmark belong to the indirect target of an action, so we classify them into a group, denoted Group 4; Scope describes an important situational aspect of an action (especially for sports news), we let it stand alone as a group,

denoted Group 5; and, all the rest case relations are classified into a group, denoted Group 6. The classification for case relations is summarized in Table 2.

Table 2. Classification for case relations

Group	Case Relation	Group	Case Relation
1	S D L	4	Y J
2	O R X F T B K	5	E
3	V	6	I, M, P, A, W, C, G, Q, N, H

Thus, we have six weighting factors that need to be estimated.

Suppose $wt(w)$ is the TF*IDF value of a word w in W-VSM, $q(w)$ is its weighting factor according to case relation of w in a sentence, then we have an adjusted weight for w:

$$wt(w)*q(w) \tag{1}$$

Furthermore, the head of a chunk is expected to be more significant than the other words in the chunk for IR. So we particularly design a set of weighting factors for heads, resulting in another six weighting factors, in accordance with the six groups in Table 2 respectively.

Consequently, the weighting for a head w is adjusted as:

$$wt(w)*r(w) \tag{2}$$

where $r(w)$ is head-related weighting factor according to case relation of w in a sentence.

We need to determine 12 weighting factors in total. We fixed the factor for Group 6 to be 1. Genetic algorithm (GA) is used to train the rest 10 factors based on the 50 queries in the training set. The setting of GA is: binary encoding, 40 populations in a generation, and the average 11-point precision as fitness function. The weighting factors obtained from GA are listed in Table 3.

Table 3. Weighting factors obtained from GA

Type	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6
q	2.147	1.575	1.315	1.100	1.545	1
r	2.761	2.144	2.383	2.900	0.133	1

Now, we yield a new IR model – Case Relation-based Vector Space Model, denoted C-VSM. Fig. 2 compares the performance of C-VSM with that of W-VSM on the test set.

As shown in Fig. 2, C-VSM outperforms W-VSM: there is a 3.4% improvement on the average 11-point precision.

We demonstrate the effectiveness of C-VSM with the following example.

Query：2000 年 9 月 25 日，北京时间周一下午刚刚结束的女子 400 米决赛上，澳大利亚名将弗里曼夺得金牌。

(Fuliman, an Australian well-known player, won the gold medal of women 400m in the afternoon, Monday 25 Sep 2000, Beijing time.)

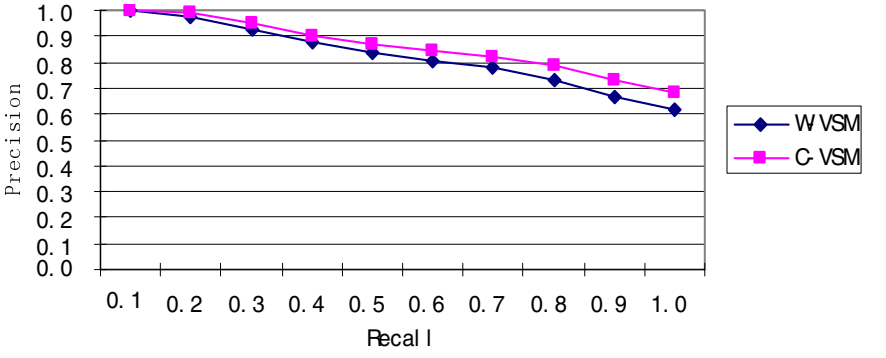

Fig. 2. Comparison between C-VSM and W-VSM

The top 3 retrieved sentences from C-VSM and W-VSM for this query are given in Table 4. Sentences with '*' are correctly retrieved results according to human judgments.

Table 4. An example for comparison between C-VSM and W-VSM

Response from C-VSM	Response from W-VSM
* 2000 年 9 月 25 日，北京时间周一下午刚刚结束的女子 400 米决赛上，澳大利亚名将弗里曼夺得金牌。*(Fuliman, an Australian well-known player, won the gold medal of women 400m in the afternoon, Monday 25 Sep 2000, Beijing time.)*	* 2000 年 9 月 25 日，北京时间周一下午刚刚结束的女子 400 米决赛上，澳大利亚名将弗里曼夺得金牌。*(Fuliman, an Australian well-known player, won the gold medal of women 400m in the afternoon, Monday 25 Sep 2000, Beijing time.)*
* 澳大利亚名将弗里曼夺得女子 400 米金牌。*(Fuliman, an Australian well-known player, won the gold medal of women 400m.)*	2000 年 9 月 25 日，北京时间周一下午刚刚结束的男子 400 米决赛中，美国名将约翰逊夺得金牌。*(Michael Johnson, an American well-known player, won the gold medal of men 400m in the afternoon, Monday 25 Sep 2000, Beijing time.)*
* 9 月 25 日，澳大利亚选手弗里曼在奥运会女子 400 米决赛中获得金牌。*(Fuliman, an Australian player, won the gold medal of women 400m on 25 Sep.)*	2000 年 9 月 25 日，北京时间周一下午刚刚结束的女子撑杆跳决赛上，美国德拉吉拉夺得金牌。*(Dragan, an American player, won the gold medal of women's pole vault in the afternoon, Monday 25 Sep 2000, Beijing time.)*

3.3 Smoothing with Similarity Between Words

An observation on C-VSM indicates that many unmatched words share the same or similar meanings, as '获得'(*gain*) and '夺得'(*seize*), and '金牌'*(gold medal*) and '冠军'(*champion*). A possible improvement for C-VSM is thus to resort to a sort of thesaurus as a means of smoothing as matching between two words is being done.

A thesaurus of words is constructed by automatic clustering on COCS. We consider the context of a word w to be a window with k words on the left and right of w respectively (k = 1 here). Two words are said to be semantically associated with each other if their contexts are similar in a document collection. In the process of automatic clustering, TF*IDF is used for word weighting, and cosine is used for similarity measuring.

In the process of retrieval, for any unmatched word w_1 in a query, we compute similarities between w_1 and any word in the target sentence. The word with the biggest similarity in the target sentence is fixed, denoted w_2. If the similarity between w_1 and w_2, $Sim_{w1,w2}$, is greater than a threshold, then we assert that w_1 is approximately matched with w_2, and the weighting of w_1 is estimated by:

$$Sim_{w1,w2}*TF_{w2}*IDF_{w1} \tag{3}$$

We try three strategies:

Strategy A: C-VSM + Approximate matching on all unmatched words in the query;
Strategy B: C-VSM + Approximate matching on all unmatched head words in the query;
Strategy C: C-VSM + Approximate matching only on the core verb in the query.

Experimental results are listed in Table 5:

Table 5. Experimental results for introducing word similarity into IR

Recall	Precision of C-VSM	Precision of Strategy A	Precision of Strategy B	Precision of Strategy C
0.1	1	1	1	1
0.2	0.993	0.987	0.990	0.993
0.3	0.953	0.927	0.944	0.952
0.4	0.911	0.896	0.911	0.914
0.5	0.880	0.869	0.871	0.880
0.6	0.853	0.820	0.843	0.852
0.7	0.822	0.806	0.825	0.824
0.8	0.791	0.778	0.788	0.792
0.9	0.725	0.706	0.715	0.729
1	0.669	0.654	0.661	0.678
The average 11-point precision	0.872	0.858	0.868	0.874

We can see from Table 5 that the noise brought by word similarity may hurt the performance of IR: only strategy C gains a bit improvement on the average 11-point precision (0.2%) compared to C-VSM.

4 Conclusion

Experimental results in the paper indicate that case relations can benefit the effectiveness of IR if they are properly combined with W-VSM, though the improvement is not as significant as expected. Approximate matching between words may also be beneficial to case relation-based IR.

We believe the largest contribution of this work is that we obtain a better performance with C-VSM (Note that the model is in fact language-independent) than W-VSM, whereas previous studies on semantic IR often obtained no better, even worse results. We preliminarily validate by experiments an assumption that semantic information is useful for IR, though the road ahead in this direction is still very long.

Acknowledgements. This research is sponsored by Fuji Xerox Co. Ltd.

References

1. Khoo, S.G.: Using Cause-effect Relations in Text to Improve Information Retrieval Precision. Information Processing and Management, 37, (2001) 119-145
2. Liu, G.Z: Semantic Vector Space Model: Implementation and Evaluation. Journal of the American Society for Information Science, 48(5), (1997) 395-417
3. Lin, X.G.: Lexical Semantics and Computational Linguistics. YuWen Press, Beijing, (1999)
4. Lu, X.: An Application of Case Relations to Document Retrieval, Doctoral dissertation, University of Western Ontario, (1990)
5. Fillmore, C.J.: The Case for Case. In: Universals in Linguistic Theory, New York: Holt, Rinehart and Winston, Inc, (1968)
6. Somers, H.L.: Valency and Case in Computational Linguistics. Edinburgh University Press, (1987)
7. Lewis, D.A.: Case Grammar and Functional Relations. Doctoral dissertation, University of Western Ontario, (1984)
8. Young, C.: Development of Language Analysis Procedures with Application to Automatic Indexing. Doctoral dissertation, The Ohio State University, (1973)
9. Croft, W.B., Turtle, H.R., Lewis D.D.: The Use of Phrases and Structured Queries in Information Retrieval. In: Proc. of the Fourteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval, (1991)
10. Hyoudo, Y., Niimi, K., Ikeda, T.: Comparison between Proximity Operation and Dependency Operation in Japanese Full-text Retrieval. In: Proc. of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, (1998)
11. Smeaton, A.F., O'Donnell, R., Kelledy, F.: Indexing Structures Derived from Syntax in TREC-3: System Description. In: Overview of the Third Text REtrieval Conference (TREC-3), National Institute of Standards and Technology Special Publication 500-225, (1995) 55-67