Semantic Similarity Measurement of Chinese Financial News Titles Based on Event Frame Extracting

Ching-Hao Mao[1], Ta-Wei Hung[2], Jan-Ming Ho[2] and Hahn-Ming Lee[1,2]
Department of Computer Science and Information, National Taiwan University of Science and Technology, Taipei 106, Taiwan[1]
Institute of Information Science, Academia Sinica. Taipei 115, Taiwan[2]
D9415004@mail.ntust.edu.tw, {geel,hoho,hmlee}@iis.sinica.edu.tw

Abstract

The Chinese financial news titles has only few words so that it is hard for measuring the similarity between titles if compare all their keywords only. In this study, we proposed a method of semantic similarity measurement for Chinese financial news titles based on constructing the Event Frame structure as the template of a Chinese financial news title. It concerns the relation between the basic meanings of two news titles for similarity measurement. In addition, a semantic similarity function is used to integrate both the relation of Event Frames of the financial news titles and the relation between the keywords of these titles. In this matter, the proposed method can differentiate the Chinese financial news that mention the same event from all other Chinese financial news by the Event Frame, since it concerns the relation between the basic meanings of two news titles and reduces the comparing time. The result of this approach shows that the Event Frame extracting has high precision and the provided semantic similarity measurement can emphasize the relation between the connotations of two news titles.

1. Introduction

Information Retrieval (IR) is a technology in retrieving documents from collections [1]. Most real-world retrieval problems are involved in retrieving some of the most similar ones to a given unknown. Especially, grouping, indexing, or searching systems employ a method for retrieval based on their own similarity measurement [2]. In these similarity measurements, keyword-based similarity measurement is one of the most classic methods to get the relation of two documents. Clearly, several problems occur in computing the similarity measurement with only keyword-based method [3-5]. The first problem is that the number of keywords increases extremely fast. The second problem is the loss of the information about the relation between keywords. For the second problem, the semantic similarity measurement is one of the most popular ways to modeling the relation between keywords [6].

The study of semantic similarity between words has been a part of natural language processing and is a generic issue in a variety of applications in the areas of computational linguistics and artificial intelligence, both in the academic community and industry [6]. Recent investigations in information retrieval and data integration have emphasized the use of ontology and semantic similarity functions as a mechanism for comparing objects that can be retrieved or integrated across heterogeneous repositories [7]. For example, in Chinese word similarity domain, HowNet [8] is the soundest semantic network for explaining the relation of word and is useful to measure the Chinese semantic similarity.

In the financial news domain, the news titles show most information of total news articles. Therefore the semantic similarity measurement of financial news titles can be used to represent the semantic similarity measurement of whole financial news articles in order to improve the time consuming and reduce the number of keywords. Computing the semantic similarity between two news titles, we should determine that what kinds of events would occur in these news titles. Because a news title has only few words, we do not have enough information if compare all their keywords only. In other words, we need to retrieve additional information from the news titles in order to model the news titles more accurate. We can classify the news titles to several classes, and using the classes and keywords of the news titles to compute the semantic similarity of them. Clearly, the classes "events" of news titles will be the suitable way to classify the news titles. "Event" means the kinds of the financial event

IEEE International Conference on e-Business Engineering (ICEBE'06)
0-7695-2645-4/06 $20.00 © 2006 IEEE

which happens in financial news. We can use the rules in an event to be the features of keywords, and let the keywords with features be the basis of computing the semantic similarity between two news titles.

FrameNet [9-12] is used to produce frame-structure description of lexical items. In each action there are several different Frame Element Groups (FEGs) for different situations. In general, frames encode a certain amount of "real-world knowledge" in schematized form. Thus, we use the Event Frames with the structure like the frames in FrameNet to encode the events in the financial domain.

In this study, we present a concept structure "Event Frame" between the whole news title and the keywords. The structure "Event Frame" provides Event Name and Event Slots to compute the semantic. An Event Frame contains the Event Name used to describe the rough meaning in the Event Frame and the Event Slots used to describe the main roles in this event of the news titles. Event Name of the Event Frames can be used to determine that whether the two Event Frames have relation between their Event Names or not by Event Name ontology. Event Slots provide the information of news title's content and can be used to compare with other news title's Event Slot. Also, we proposed a Semantic similarity measure based on Event Frame for measuring the similarity between two news titles. The proposed method extracts the concept meaning of News Titles into Event Frame, and then applies the Frame based similarity measure to determine the similarity between news titles. This method lays stress on both literal meaning and concept implication that fit human common sense when comparing two news titles. The experiment result of this study shows that the Event Frame extracting has high precision as man-made and the provided semantic similarity emphasize the relation between the connotations of two news titles rather than the relation of keywords.

2. THE FRAMEWORK OF MEASURING THE SEMANTIC SIMILARITY OF CHINESE FINANCIAL NEWS TITLES

In this section, we propose the framework of measuring the semantic similarity of two Chinese financial news titles. There are two steps to compute the semantic similarity. First, we extract Event Frames from the Chinese financial news titles. Second, we use the similarity measurement based on Event Frame to compute the semantic similarity of two Chinese financial news titles. Figure 1 shows the view of these two steps. In subsection 2.1, we will describe the concept of Event Frame. In subsections 2.2 and 2.3, we will describe the details of these two steps, respectively.

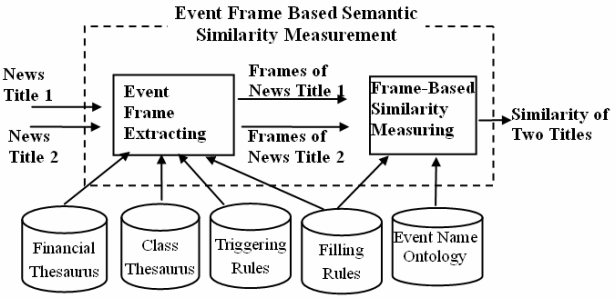

Figure 1. The framework of the Event Frame Based Semantic Similarity Measurement

2.1 Event Frame Concept

Event Frames can be seen as a subset of frames which encode the "real-world knowledge". An Event Frame records a financial event which happened in the real-world. As the frames in FrameNet, each Event Slot in an Event Frame is a role in this event. For example as in Figure 2, the second row describes the important roles in the frame "Commercial Transaction", and the frame "Commercial Transaction" explains that the event "BUYER gets GOODS, SELLER gets PAYMENT" in the real world. This event will be described in the third row in this frame.

When we read a financial news title, it is important to determine the kind of events in this news title. In this study, we use the pre-defined several Event Frames to represent the kind of events in a financial news title, because we not only need to determine the events of news titles but also use the events to decide which terms would be filled in the Event Frame.

We create some appropriate Event Frames as event templates to describe all events in the financial domain. Each Event Frame has its Event Name, Description, and Event Slots. The Event Slots of the Event Frame show the key role of the event. Besides, terms with similar meaning is common in Chinese Language, especially the verb class, so we keep the verb as an Event Slot value in each Event Frame to preserve this information. Figure 3 shows several Event Frames we defined here.

Frame(CommercialTransaction)
frame-element{BUYER, SELLER, PAYMENT, GOODS}
scenes (BUYER gets GOODS, SELLER gets PAYMENT)

Figure 2. A frame which describes the event of Commercial Transaction in FrameNet

Frame(StockPriceGoesUp)
Stock Stock type (ADR, Taipei stock price) Mood Stock price at event time Event time Ranges the stock price changes Contrast time of the stock going up Words describing that the stock goes up
Frame(RiseEstimate)
Estimating organization or analyst Stock being estimated Event time Suggestion Words describing that someone rises the estimate Target stock price Mood Contrast time of rising estimate

Figure 3. Event Frames with Event Names and their corresponding Event Slots as specific roles

2.2 Event Frame Extracting

We use the encoding concept like the frames in FrameNet to transfer the Chinese financial news titles to Event Frames. The semantic means of each term generated from Chinese News Titles are used to trigger the related News Event. Via Event Frame Extracting, we can understand the roles which every term extracted from the Chinese financial news title plays, and the extracted Event Frame can be prepared for frame-based similarity measure between two news titles. Figure 4 show the steps and will be described below.

- CKIP-Based term generating: identifying terms with the syntactic meaning in a sentence from the Chinese financial news title and determine the syntactic characteristic of each term.
- Semantic tagging: determining the semantic meaning of each term extracted from the Chinese financial news title and using an appropriate tag to keep this information.
- Event triggering: using the tags of terms to identify the events where the Chinese financial news title says.
- Frame Slot filling: for every Event Slot in each Event Frame, finding the appropriate term as the slot value using the tags of terms extracted from the Chinese financial news title.

For extracting terms from the Chinese financial news titles, we use CKIP lexicon and grammar software [13] to parse a Chinese financial news title and obtain the terms with syntactic tags. It is because CKIP can provide many terms with specific semantic meaning which are better than the terms generated by statistical method. CKIP can also determine a syntactic meaning of each term extracted from a Chinese financial news title and the syntactic tag can help us to decide the semantic meaning of each term. In addition, we use the financial thesaurus maintained by CKIP members to increase the accuracy of extracting special terms.

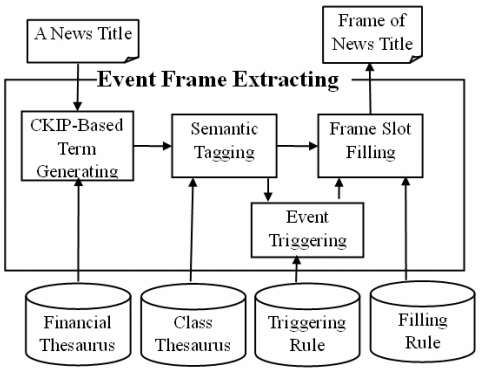

Figure 4. Event Frame extracting procedure which transforms a news title to a group of Event Frames

After terms are generated by CKIP-Based Term Generating, Semantic Tagging identifies the semantic meaning of each term. We create a Class Thesaurus to determine the semantic meaning of these terms generated from the step of Term Generating. Class Thesaurus includes the terms semantic meaning, divided into two kinds of groups "Formal From" and "General Form". Formal Form can be explained as the terms with numeric data, such as "January 31", "The 2nd season", "13%", "14 dollars". General Form is exclusive of Formal Form. Because the keywords in a financial news title would be restricted, so the most terms in Class Thesaurus are available for identifying the semantic meaning of each term. Semantic Tagging can identify the semantic meaning of each term by Class Thesaurus. After Semantic Tagging, Event triggering uses the combination of tags in a Chinese financial news title to determine which event the news title describes according to Triggering Rule. The Triggering Rule contains several rules for describing which tags of term related the specific events.

The final step is to fill the terms decided by relational events from Event Triggering into the slots of Event Frames. Because each slot has its specific meaning, we can use the semantic meaning to decide which term can be filled in corresponding slot.

2.3 Frame-Based Similarity Measuring

The Frame Based Similarity Measuring is used to compute the semantic similarity of two Chinese financial news titles. In order to compute the semantic similarity of two Chinese financial news titles, comparing the relation between the keywords of these two news titles is not enough. Therefore, the structure "Event Frame" provides Event Name and Event Slots with filling rules and semantic tags generated by Event Frame Extracting to help us compute the semantic similarity between two financial news titles.

There are two steps in this measuring are Event Name Related Measure and Event Slots values similarity Measure. Figure 5 is the overviews of the similarity measurement between two frames belonged to two news titles will be described below.

- Event Name Related Measure: determining whether the two Event Frames has relation between their Event Name or not by the Event Name of the Event Frames and the Event Name ontology.
- Event Slots Values Similarity Measure: measuring the both Event Frame slots values by either semantic similarity comparison or terms overlapping.

In Event Name Related Measure, we can begin with comparing the Event Names and measure that both of them have relation or not in the Event Name ontology. According to the result of Event Name Related Measure, the Event Slots Values Similarity Measure can use semantic similarity method to compare the slots values between Event frames. In Event Slots Values Similarity Measure, if the Event Names are related, the slot pairs will then be compared for the semantic similarity comparison by Filling Rules mentioned in Event Frame Extracting and HowNet [8]. However, if two Event Frames have names that are not related, we will compute the term overlapping of two Event Frames. Term overlapping can be used to find that two events are still considered similar when they both have the specific term. This method can take account of the situation that two news titles are similar when they have specific term, even if the event name seem not to be related.

Event Name Related Measure can be easy to determining that either two Event Frames has relation between their Event Name or not by Event Name Ontology. Event Name Ontology provides the information about the related degree between two News Events for Event Name Related Measure. Some Event Frames have inheritance relation, and therefore we can say that one Event Frame is the sub-event of another Event Frame if there are inheritance relations between these two Event Frames. For example, the Event Frame "StockPriceChanges" is the sub-event of the Event Frame "StockPriceStatus", and the Figure 5 shows a part of this event ontology. In this ontology, we can define the relation "related" if and only if two Event Names have inheritance relation or the distance between these two Event Frames in the ontology is very small. There are two advantages for this procedure. First, we can use the ontology to measure the similarity of two Event Frames, and keep the information of Event Name. Second, when the two Event Frames are not related, we can skip the similarity measure to reduce the measure time.

In Event Name Related Measure, we use the Frame-based similarity measure $Sim(A,B)$ for determining the events of two Chinese financial news titles A, B are related or not. We define the similarity measure $Sim(A,B)$ in (1):

$$Sim(A,B) = \sum_{i=1}^{n_a} f_{\max_{B_j \in EF_s(B)}} SF(f_{A_i}, f_{B_j}) + \sum_{j=1}^{n_a} f_{\max_{A_i \in EF_s(A)}} SF(f_{A_i}, f_{B_j}) \quad (1)$$

where $EF_s(A) = f_{A_1}, f_{A_2}, ..., f_{A_{na}} \mid A_i \in \{1,2,...,n\}$ is the set of frames of A, n_a is the number of frames in A. $EF_s(B) = f_{B_1}, f_{B_2}, ..., f_{B_{n_b}} \mid B_i \in \{1,2,...,n\}$ is the set of frames of B, where n_b is the number of frames in B. $SF(f_{A_i}, f_{B_j})$ denotes the frame similarity between f_{A_i} and f_{B_j} can be generated by Event Name Ontology. The reason of keeping the maximum value is to ensure that each frame in A can find the most suitable (similar) frame in B. Therefore we can know that for each Event Frame in a news title, we only need to find the Event Frame most similar in the other news title, and this comparison method can express the difference between two news titles for determining the two titles have relation or not.

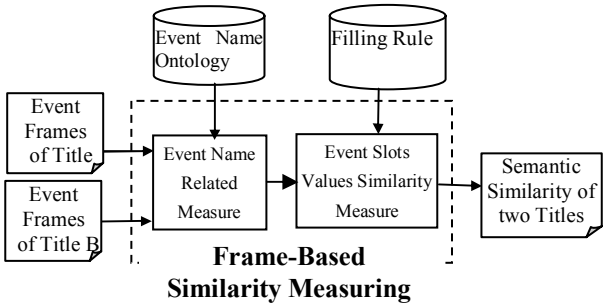

Figure 5. Semantic similarity measurement procedure based on Event Frame structure

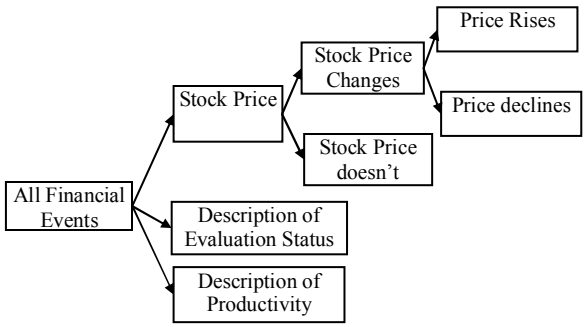

Figure 6. Part of Event Name ontology used to determine whether two Event Names are related or not.

After comparing with the Event Names and find their relation in the Event Name ontology, Event Slots Values Similarity Measure will compare with the Event Slot values in Event Frames. In terms of no relation between two Event Frames, the slot pairs will then be compared for the semantic similarity comparison, but in terms of having relation between, Term Overlapping will compute the terms overlapping of two Event Frames.

Because the values between in slot pairs may not be exactly the same, we must measure the semantic similarity between two different values. The value can be divided into three groups: numeric values, nominal values, and verb values. Numeric values denote the values which are numeric data. Nominal values denote the values which are not numeric data and are not a verb. The verb values denote the values which is a verb. For numeric values, we united the unit of two values and compare with their numeric data. For nominal values, we create a tree structure to maintain the semantic similarity between these values. For verb values, we use the similarity measure based on HowNet [8] to compute the similarity between them.

If two events name are related, Slot-Value Matching determines the slot pairs which need comparing to each other in these two frames. Let $S_{A_i} = \{S_1^{A_i}, S_2^{A_i}, ..., S_{ns_{A_i}}^{A_i}\}$ are the slots of A_i, where ns_{A_i} denotes the number of slots in frame A_i, and $S_{B_j} = \{S_1^{B_j}, S_2^{B_j}, ..., S_{ns_{B_j}}^{B_j}\}$ are the slots of B_j, where ns_{B_j} denotes the number of slots in frame B_j. We define that $(S_a^{A_i}, S_b^{B_j})$ is a slot pair. After that we need to determine whether the slot pair needs to be compared with or not. Here we use the filling rules of the slots $S_a^{A_i}$ and $S_b^{B_j}$ to estimate the degree that the slot pair $(S_a^{A_i}, S_b^{B_j})$ needs to be compared with. If the tag of term filled in $S_a^{A_i}$ is the same as the tag of term filled in $S_b^{B_j}$, the slot pair $(S_a^{A_i}, S_b^{B_j})$ should be compared with; else this pair should not be compared with.

In Slot-Value Semantic Comparison, let $P(f_{A_i}, f_{B_j}) = \{p_1^{A_iB_j}, p_2^{A_iB_j}, ..., p_{np_{A_iB_j}}^{A_iB_j}\}$ be the slot-pairs which need comparing in A_i and B_j, where $np_{A_iB_j}$ denotes the number of slot-pairs needed comparing between A_i and B_j. After that we can define the frame similarity $SF(A_i, B_j)$ between news titles of A_i and B_j in (2):

$$SF(A_i, B_j) = (1 - CW) * \sum_{m=1}^{np_{A_iB_j}} w_m SS(p_m^{A_iB_j}) + CW * SC(f_{A_i}, f_{B_j}) \quad (2)$$

where $w_m \geq 0, \sum_{m=1}^{np_{A_iB_j}} w_m = 1$ is the weighting parameters of each slot values. $SC(f_{A_i}, f_{B_j})$ denotes the Event Name similarity which can be computed from the Event Name Related Measure by Event Name ontology, and $SS(P_m^{A_iB_j})$ are the similarity of Event Slot values in this Event Slot pair $P_m^{A_iB_j}$.

If two Event Frames have not related names, we do not need to compare with the Event Slot values of these two Event Frames. But in some situation, two events are still considered that they are similar when they both have the specific term. When the Event Names are not related, we need to keep the information of term overlapping between two Event Frames. Let $nd(f_{A_i})$ and $nd(f_{B_j})$ be the number of the non-duplicate terms in the Event Frame A_i and B_j, and $nd(f_{A_i}, f_{B_j})$ be the number of terms belonged to both A_i and B_j. After that the term overlapping $TO(A_i, B_j)$ between A_i and B_j is in (3):

$$TO(A_i, B_j) = CW * \frac{2 * nd(fA_i, fB_j)}{nd(f_{A_i}) + nd(f_{B_j})} \quad (3)$$

where CW is the parameter used in the function (3). When $nd(f_{A_i}, f_{B_j})$ is large, even the Event Names of A_i and B_j are not related, they may be considered that they are similar.

Finally we can conclude the frame-based similarity measurement $Sim(A, B)$ as (4) (5):

$$Sim(A,B)=\sum_{i=1}^{n_a} f_{B_j} \in \overset{\max}{EFs}(B)SS(f_{A_i},f_{B_j})+ \sum_{j=1}^{n_b} f_{A_i} \in \overset{\max}{EFs}(A)SF(f_{A_i},f_{B_j}) \quad (4)$$

$$Sim_{norm}(A,B)=Sim(A,B)/(n_a+n_b) \quad (5)$$

where the $Sim_{norm}(A,B)$ in function (5) is the normalization of (4) that ensure the range of similarity is [0,1], then $SF(f_{A_i},f_{B_j})$ is defined as:

$$SF(f_{A_i},f_{B_j})=\begin{cases}(1-WC)*\sum_{m=1}^{np_{A_iB_j}} w_m SS(p_m)+WC*SC(f_{A_i},f_{B_j}), \\ \text{if name of } A_i \text{ and } B_j \text{ are related} \\ TP(A_i,B_j)=(1-WC)*\dfrac{2*nd(f_{A_i},f_{B_j})}{nd(f_{A_i})+nd(f_{B_j})} \\ \text{if name of } A_i \text{ and } B_j \text{ are not related}\end{cases} \quad (6)$$

3. Experiments and Analysis

In this section, we will introduce the experiment result and discuss the feature of the semantic similarity measurement. We describe the properties of the Chinese news title dataset collected from Yahoo [14] in subsection 3.1. We also focus on the accuracy of Event Frame extracting in opposition to the human hand tagging result in 3.2. After that, we use a strategy in statistic methods to determine the weighting parameters *CW* mentioned in (6) and discuss the variation between strategies in subsection in 3.3. Finally, we compare the correctness between proposed method and keyword based method using TF-IDF in 3.4.

3.1 Dataset Description

For dealing with the Chinese financial news titles and analyze it in the Event Frame-base method, we collect the financial news titles from Yahoo from 2002.10.30 to 2003.07.21. Because the number of Even Frames would increase rapidly when news coverage increase, so we choose the news about the TAIWAN SEMICONDUCTOR MANUFACTURING (TSM) and some other electric categorized stocks to observe the feature the semantic similarity measurement makes. There are 531 Chinese financial news titles in this dataset.

3.2 The Accuracy of Event Frame Extracting

We will analysis and discuss the accuracy from several views in the result of Event Frame Extracting experiment.

Event Frame Extracting can deal with about 88% (468 news titles) Chinese news titles in the dataset. In the remaining 63 (12%) news titles, which can't be deal with, we can observe that most news titles are about human speech and the evaluation of other appraisal corporations. In the Table I, we can see the statistic result of the news titles not be dealt with in a rough classification.

The accuracy of the number of news titles allocate to appropriate Event Frames. In the remaining 463 news titles of the dataset, which can be dealt by Event Frames Extracting, the accuracy is 90.16% (427 news titles). By observing the news titles allocated incorrectly, the result in Table II shows that the news titles with more than one Event Frame have more chance to be allocated incorrectly. For solving this problem, we have tried to use some symbol to divide a news title to several term sets if the news title would be allocated to more then one Event Frame, but the result is worse than the current method of term combination. There are still some left problems of allocating Event Frames to news titles included sometimes one of the major terms does not appear in the news title, but it still has the meaning of the Event Frame. Finally, there are some news titles with unknown Event Frames.

In briefly, the Event Frame extracting can transfer Chinese financial news titles to Event Frames. In 531 Chinese financial news titles, the Event Frame Extracting can allocate the appropriate Event Frames and can deal with 419 news titles with correct Event Frame. The accuracy of Event Frame extracting is about 78%.

Table 1. The statistic result of the news titles without appropriate Event Frames

Reasons of titles without appropriate Event Frames	Numbers (Rate %)
Human speech	15 (23.81%)
Estimation by corporation, etc.	22 (34.92%)
Loss of Event Frame construction	11 (17.46%)
Others	15 (23.81%)
All the news without appropriate Event Frames	63 (100%)

Table 2. The list of news titles which are not allocated correctly.

Reasons of titles are not allocated correctly	Numbers (Rate %)
The news titles with more than one Event Frames	16 (39.02%)
Loss of major terms	10 (24.39%)
Unknown Event Frame	10 (24.39%)
Others	5 (12.20%)
All the news titles without correct Event Frames	41 (100%)

3.3 Strategy for Parameter Determination

We provide several strategies in statistic-based method for parameter determination of each slot for describing an event more precisely. The strategy is based on two concepts, one is the probability of null value in the Event Frames, the other is that the major term is more important than other terms which may not appear in a news titles.

For describing an event precisely by an Event Frame, we use several Event Slots to record the detailed status in this Event Frame, but sometimes the news titles would not describe an event so precisely. It causes the problem of null value. The appearance probability of null value can help us to determine the importance of a specific Event Slot. When an Event Slot has null value frequently, it reminds us that the importance of this Event Slot would be down. Let $P_m^{A_i B_j} = (S_a^{A_i}, S_b^{B_j})$ denotes a pair want to be compared with. Thus we define that the weighting parameter w_m in equation (6) as follows:

$$w_m = \frac{AP(p_m^{A_i B_j})}{\sum_{m=1}^{np_{A_i B_j}} AP(p_m^{A_i B_j})} \quad (7)$$

where $AP(p_m^{A_i B_j})$ is defined as:

$$AP(p_m^{A_i B_j}) = (1 - NP(S_a^{A_i})) * (1 - NP(S_b^{B_j})) \quad (8)$$

where $NP(S_a^{A_i})$ denotes the appearance probability of null value in the Event Slot $NP(S_a^{A_i})$.

In this strategy, because the appearance probability of null value in major terms in an Event Frame is closed to 0, $NP(S_a^{A_i})$ would be closed to 1 where $S_a^{A_i}$ the Event Slot with major terms is. Therefore the parameter W_m would be large if the Event Slot pair $P_m^{A_i B_j} = (S_a^{A_i}, S_b^{B_j})$ contains the major terms in Event Classes. Figure 8 shows the difference between the strategy and the equal weighting parameter. We can know that appearance probability of null value will make the verb difference goes up.

3.4 Comparison with keyword based method using TF-IDF

Comparison with keyword based method using TF-IDF. In the dataset of Yahoo, we try to use the semantic similarity to find out the news titles which mention the same event in financial domain. Clearly, a semantic similarity value as the threshold can be used to determine if the two Chinese financial news titles mention the same event in financial domain. Considering the time dependency, we divided the dataset to subsets; the news titles in the same subset happened in the same day. Removing the subsets in which there is only one news title or there are no titles describing the same event. According the result in Table 3, we can know that the proposed method has superior correctness than keyword based method for differentiation between the financial Chinese news titles mention the same event or not.

Title A: 21 日台積電 ADR 收盤價 7.55 美元，較前一交易日下跌 0.04 美元 Taiwan Semiconductor Manufacturing Co. ADR closing price is 7.55 dollars on the 21st, drops by 0.04 dollars compared with the last bargain day

Title B: 費城半導體指數大跌，權值股台積電開盤跌破季線 Philadelphia semiconductor index slump, right value gangs of Taiwan Semiconductor Manufacturing Co. opening quotation break season line by a fall

The strategy used	Semantic similarity value
The probabilistic strategy	0.790
Equal weighting strategy	0.822

Figure 8. The statistic result of the news titles without appropriate Event Frames

4. Conclusion

In this study, we proposed a method of semantic similarity measurement in Chinese financial news titles. It concerns the relation between the basic meanings of two news titles and reduces the comparing time. In the experiment results, the event frame extracting would keep most information of original news titles, and the similarity measurement would give higher priority to the financial news titles with related Event Names in the Event Name ontology. The proposed method can help users to find the related financial news titles in an event-driven concept instead of the keyword-driven concept. We can differentiate the Chinese financial news titles which mention the same event from all the Chinese financial news titles by the semantic similarity measurement based on Event Frame extracting. In addition, we anticipate that the semantic measurement based on Event Frame extracting can help users to find the related financial news titles in an event-driven concept instead of the keyword-driven concept.

5. Acknowledgements

This work is supported in part by the National Digital Archive Program-Research & Development of Technology Division (NDAP-R&DTD), the National Science Council of Taiwan under Contract No. NSC 94-2422-H-001-006, and by the Taiwan Information Security Center (TWISC), the National Science Council under Contract No. NSC 94-3114-P-001-001-Y.

6. References

[1] J. Kleinberg, "Authoritative Sources in A Hyperlinked Environment," *Journal of the ACM (JACM)*, vol. 18, no. 5, pp. 604-632, 1999.

[2] L. Eikvil, "Information Extraction from World Wide Web- A Survey," *Technical Report 945*, Norwegian Computing Center, 1999.

[3] D.L. Lee, H. Chuang, K. Seamons, "Document ranking and the vector-space model," *IEEE Software,* vol. 14, no. 2, pp. 67-75, Mar/Apr 1997.

[4] J.J. Kim, B.W. Hwang, S.W. Lee, "Retrieval of the top N matches with support vector machines," *Proc. of 15th International Conference on Pattern Recognition, 2000,* vol. 2, pp. 716-719, Sept 2000.

[5] T.H. Haveliwala; "Topic-sensitive PageRank: a context-sensitive ranking algorithm for Web search," *IEEE Transactions on Knowledge and Data Engineering,* vol. 15, no. 4, pp. 784-796, July/Aug 2003.

[6] Y. Li, Z.A. Bandar, D. Mclean, "An approach for measuring semantic similarity between words using multiple information sources," *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, no. 4, pp. 871-882, July/Aug 2003.

[7] V. Oleshchuk, A. Pedersen, "Ontology based semantic similarity comparison of documents," *Proc. of the 14th International Workshop on Database and Expert Systems Applications, 2003,* pp. 735 – 738, Sept 2003.

[8] HowNet, http://www.keenage.com/html/e_index.html

[9] S. Atkins, M. Rundell, H. Sato, "The Contribution of Framenet to Practical Lexicography," *International Journal of Lexicography*, vol. 16, no. 3, pp. 333-357, 2003.

[10] F. Baker; Collin and J. Ruppenhofer, "FrameNet's Frames vs. Levin's Verb Classes," In J. Larson and M. Paster (Eds.),*Proc. of the 28th Annual Meeting of the Berkeley Linguistics Society*, pp. 27-38, 2002.

[11] F., Baker, Collin, J., Fillmore Charles, B., Lowe, John; "The Berkeley FrameNet project," *In Proc. of the COLING-ACL,* 1998.

[12] C. Boas, Hans, "Frame Semantics as a framework for describing polysemy and syntactic structures of English and German motion verbs in contrastive computational lexicography," In: Rayson, Paul, Andrew Wilson, Tony McEnery, Andrew Hardie, and Shereen Khoja (eds.), *in Proc. of the Corpus Linguistics 2001 conference on Technical Papers*, vol. 13. Lancaster, UK: University Centre for computer corpus research on language, 2001.

[13] CKIP, http://godel.iis.sinica.edu.tw/CKIP/.

[14] Yahoo, http://www.yahoo.com.tw

Table 3. THE CORRECTNESS OF USING SIMILARITY TO FIND OUT THE SIMILAR NEWS TITLES.(A) Semantic Similarity Measurement Based on Event Frame Extracting (B) TF-IDF method with cosine measure

(A)

THRESHOLD VALUE / EVENT FRAME EXTRACTING	0.9	0.7	0.5	0.3
CORRECTNESS OF SIMILAR PAIRS OF NEWS TITLES	31.5%	47.3%	97.2%	97.2%
CORRECTNESS OF NON-SIMILAR PAIRS OF NEWS TITLES	100%	91.1%	91.1%	88.8%

(B)

THRESHOLD VALUE / WITH COSINE MEASURE	0.55	0.1	10^{-2}	10^{-4}
CORRECTNESS OF SIMILAR PAIRS OF NEWS TITLES	36.8%	89.4%	97.2%	97.2%
CORRECTNESS OF NON-SIMILAR PAIRS OF NEWS TITLES	95.5%	91.1%	91.1%	88.8%