Constructing Hierarchical Topic Knowledge Window for Chinese News Browsing and Retrieval

Chuen-Min Huang[1] Sheng-Fu Chang[2]

[1,2]Department of Information Management
National Yunlin University of Science & Technology, Taiwan, ROC
{huangcm, g9323722}@yuntech.edu.tw

Abstract

This paper proposes an event-based hierarchical knowledge structure for Chinese news display. We apply Hierarchical Topic Detection and Tracking technique to group news reports into clusters. Considering the extensibility of events, the knowledge structure is designed with robustness to merge similar reports and reorganize the structure automatically. This brings the merit of tracking associated follow-up event. To verify the applicable features, we design an assessing mechanism and call for participation for two weeks. There are 75 percent of participants expressing their high remarks on visualized knowledge expression. Eighty-two percent of participants consider the event heading is meaningful and acceptable. Over 90 percent of participants consider this study to be applicable.

Keywords: Hierarchical Knowledge Structure, Event Detection and Tracking, Association Rules.

1. Introduction

Gartner Group defines Knowledge Management as "a discipline that provides an integrated mechanism to identifying, capturing, retrieving, sharing and evaluating information assets of an enterprise." [8] Documents, no matter structured or semi-structured, could be turned into information according to retrieval technology by which to capture knowledge easily from large amount of information.

As we observe that on-line news portals only provide simple reading mechanism based on categories such as policy, society etc. without any arrangement of hierarchical conception. Previous researchers also primarily focused on topic categories and knowledge map to retrieve and present knowledge hidden from diverse information source as a new approach of knowledge management [2][3][5][12]. Those researches who pointed out the above two methods could optimize browsing behaviors, however, only a tiny portion of researchers employ hierarchical visualization in their experiment.. Although Ong claimed to propose a novel approach for news knowledge representation [8], that merely displayed keyword categories rather than the hierarchy of a news event. In consideration of this, we take a further step by merging topic category and knowledge map as Hierarchical Topic Knowledge Window (HTKW), which enables users to understand news topic easily and subsequently to explore the context of whole news events.

HTKW is the main system to perform all procedures including preprocessing, segmentation, weighting, clustering, summarization and visualization. For visualizing, we also deploy the knowledge windows over Internet that users could browse those windows from remote interface by his/her browser.

In this paper, we briefly describe the notions of topic, event and relevant backgrounds (Section 2). Then we continue introducing HTKW in detail (Section 3). Next is to evaluate this study (Section 4). Lastly, we remark on future work and conclusion.

2. Literature Review

According to the report of TDT2004, each topic contains a number of attributes including topic's title, description, seminal events, what, when, where, etc. An event is defined as "a particular thing that happens at a specific time and place ". It might be a plane crash, or court adjudication [1][10].

A topic is composed of several events and could track associate follow-up events with relationships. Hence, Maedche and Steffen applied association rules to construct combinations of different conceptual relationships with support and confidence [6]. Knowledge based rules must train lots of rules and include basic patterns in the database. In order to express knowledge entities, those rules are going to be utilized to combine for relationships [11].

A map is used to illustrate certain relationships of objects which are more acceptable for human. Knowledge map is using map to take on knowledge by looking at icons, lines or fewer texts. The knowledge map could be separated several ways: subject hierarchy, manual and automatic construction, respectively. The subject hierarchy such as Yahoo! subject category or library book category that constructs regular categories by letters of the alphabet. Manual construction could use visual tools to construct knowledge map, however that not only costs more resources but need professionals to establish and maintain. Other method is automatic construction which forms knowledge map through Information Technology, such as data mining or IR technique [8]. Chen proposed hierarchical topic category to classify web documents [4]. Afterward, Chen improved the efficiency of previous project [3]. The latest issue was that Ong employed Self Organize Map to construct Newsmap [8].

3. Methodology

The system architecture of this study, as figure 1 illustrating, could be separated into several tasks: Text Processing, Event Detection and Tracking, Hierarchical Clustering, and Multi-Documents Summarization.

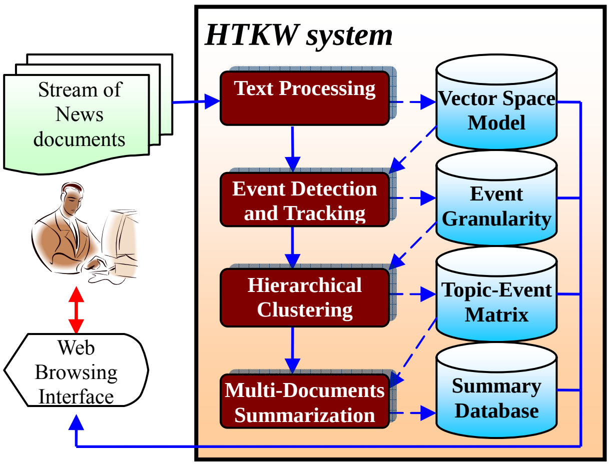

Fig. 1: Main system architecture

3.1 Text Processing

This research applied two top-performance methods to deal with text retrieval produces: Cutting off terms by Chinese corpus and CKIP component based on Service-Oriented Architecture developed by Academia Sinica. To turn the textual data into numeral data, this research will assign the weight to every term and store them in the Vector Space Modeling (VSM). We divided the terms into nouns, which may affect the accuracy of clustering or the quality of summarization; and proper nouns, which may not affect so mush as nouns, but there may be some unexpected result due to its variation.

3.2 Hierarchical Clustering

When one of the topics has the narrower sub-events, the system has to be able to cluster all streams of inputted documents hierarchically. The sub-events must relate to the upper topic semantically. One of the incremental clustering algorithms, Single-pass clustering, is used to detect the inputted news documents by adjusting the threshold to obtain the granularity in different levels. To track the news event, we use the 2-way kNN method to achieve it. Below, we applied a step-by-step method to describe the hierarchical-layered TDT algorithm:

- Utilize the procedure of weighting mentioned above to initialize every document into VSM. After filtering the noises and removing the stopwords, the system will assign a score for every term and pick out the top K term as the input vector, because other lower-weight terms will not affect clustering.
- Calculate the similarity score between the inputted document and all documents in the objective events. The Cosine similarity is used to measure the distance between two VSMs. If the average similarity score of the event was less than the predefined threshold, the document would be classified into the new event in that it does not has much similarity.

$$sim(x, c_y) = \frac{\sum_{i=1}^{M} Wx_i \times Wcy_i}{\sqrt{\left(\sum_{i=1}^{M} Wx_i^{\,2}\right) \times \left(\sum_{i=1}^{M} Wcy_i^{\,2}\right)}} \qquad (1)$$

According to the Equation (1), assume that we have two weights: Wx_i and Wcy_i, and two documents: Dx and Dy. The former is the weight of term i in the Dx and the latter is the weight of i in the Dy in event c.

- This research found out in the observation that an event will decay by time shifting, hence, we need to take time window into account when dealing with the event decay. The decaying-weighting function is applied based on Nallapati' research showing as Equation (2) where x is the new inputted document, c_i is used to express event i, t_x and t_{ci} are the time stamp of document x, and event c. T is the time interval between t_x and t_{ci}; α is time decaying constant to determine the degree of event decay and transform [7].

$$score(x) = \max(sim(\vec{x}, \vec{c_i}) \times e - \frac{\alpha\left|t_x - t_{c_i}\right|}{T}) \qquad (2)$$

- Classify the inputted document into proper events if it could be classified to more than one event. In order to track a specific event, we adapted the supervised text classification method of k-Nearest Neighbor (kNN) algorithm to achieve high-performance.

$$score(\vec{x}, kp, kn, D) = \frac{1}{|U_{kp}|} \sum_{\vec{y} \in U_{kp}} sim(\vec{x}, \vec{y}) - \frac{1}{|V_{kn}|} \sum_{\vec{z} \in V_{kn}} sim(\vec{x}, \vec{z}) \qquad (3)$$

See Equation (3), x is the inputted document and y (z) is the positive (negative) training document in the objective (non-objective) event. U_{kp} consists of kp and the nearest neighbors x among the positive documents in the training event; and V_{kn} consists of kn the nearest neighbors x among the negative documents in the training event. D is the training event of documents. k is the number of the nearest neighbors of x in D, which is used to compute the relevance score.

- Apply the above steps recursively. The system will perform a recursive thread procedure of executing TDT algorithm, until each map contains no more than t news documents.

3.3 Knowledge Representation

After carrying out HTDT, the following tasks are the key factors to users. The tasks are summarizing, labeling for every topic and event, and knowledge window. Summarization means that a brief statement describes the context of an event by compressing a great quantity of information. The heuristics to select sentences from the event to construct the summarization are as follows: (1) Select the candidate sentences from the event by comparing the similarity of news title and the threshold; (2) Remove the sentence in which TFIDF is less than the average TFIDF of candidate sentences; (3) Select the highest TFIDF as the most representative sentence in the cluster; (4) Repeat step 1 to 3, until all sentences from the clusters of an event are picked out; (5) Repeat step 1 to 4, until all events are processed.

Another task is to affix the label to topics and event. As what we mentioned in literature review, through the Apriori algorithm of association rule, the system mines lots of combinational relevant label terms. We can observe in a phenomenon that it is less applicable to extract only the label through the Apriori algorithm. Therefore, in order to achieve higher efficiency and capture the fittest label, we reused these label sets to intersect with their TFIDF and Location Information.

4. System Evaluation

The data collections were mainly collected from Yahoo! news portal. We found that after performing the clustering algorithm, the better orders of news categories are politics, entertainment, and society. Other categories are too dispersed to organize into the same cluster. We total collected approximately 23300 news documents to be the experiment data, and selected randomly topic/event labels and knowledge windows as evaluated data for web-based survey. Because the news documents collected from Taiwan Yahoo! news portal, the experiment was conducted using Taiwanese as participants.

4.1 Evaluation Result

The purpose of system evaluation is to evaluate the semantic accuracy of topic/event labels and the quality of the knowledge window. To evaluate the accuracy of topic/event labels on the measurement of degree, we calculated the rate of degree made from comparing the research results with the selected results from participants. So the proposition is: Are the topic/event labels in this research similar to the topic/event labels selected by participants?

The experiment procedures and calculated equations in this research are as follows.

- To evaluate the topic label, we designed a web-based survey system for the participants to read a brief summarization with a number of keywords highlighted. Then we asked the participants to select a keyword that they thought best correspond to the summarization.
- To evaluate the event label, we picked out two events randomly and listed a string of event label sets. After the participants had read the summarization, we ask the participants to select an event label.

Table 1 Research Results

Label	System	Participants
Topic label	70 %	74.75 %
Event label	45 %	57.19 %

As table 1 show, in the topic label, the evaluated data between the system and participants is nearer than the event label. It may because the topic label's candidate keywords are possible more specific for participants to select. Another possible reason is that people has the ability to do semantic analysis to distinguish the topic label which is the most representative to the topic summarization; however, the system, on the contrary, can not distinguish the topic label in the semantic computation, because it only use the algorithms or rules such as TFIDF to extract the topic label.

In the event level, the evaluated data between system and participant is far than the topic label. It is because the system might not match people's thinking in the event label so that the system got lower degree. Also, there might be some reasons similar to the topic label we mentioned above.

In order to normalize the evaluated standard, we used the five-point scale to indicate the preference of participants from high to low. In the statistical graph (Figure 5), the curves reveals the normal distribution which indicates that most of the participants considered the exhibition of knowledge window and topic summarization have high remarks in the visualized knowledge representation.

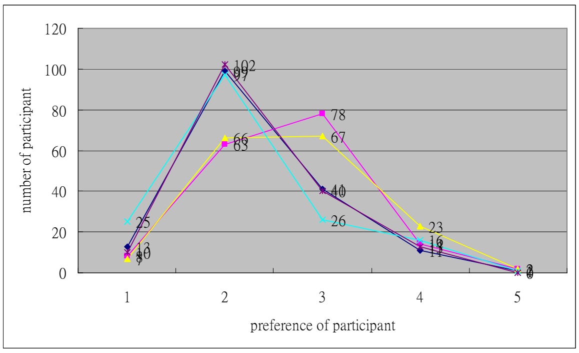

Figure 1 Results of statistics

5. Conclusion and Future Work

The information overloading problems had become a critical concern of how to extract valuable information in real-time application. To provide more convenient browsing mechanism, we merge the topic category and knowledge map as HTKW to explore the context of news events. In order to evaluate result, we conducted an experimental study to verify that the browsing mechanism surely help readers capture knowledge without extra efforts. The result shows that most participants prefer visualization representation while reading news and HTKW with summarization. It also indicates that readers understand the whole picture of news events effectively with such arrangement.

It is encouraging to understand that over 90 percent of participants consider this study to be applicable. The outcome of this study implies that the concerns of visual knowledge representation would be continuously emphasized. As for the task of transforming content into graphic expression is still a demanding challenge. It is hoped that the naming label for news topics and events could be improved, and the relationship between topics and events could be enhanced. It is also expected to combine Topic Maps or ontology [9] with similiar studies to express knowledge in the future. If more hierarchical relationship between topics and events is proposed, the field of Ontology could be a more potential and interesting research issue to explore.

6. Reference

[1]. Allan, J., Carbonell, J., Doddington, g., Yamron, J., and Yang, Y. "Topic Detection and Tracking Pilot Study Final Report."

[2]. Carmel, E., Crawford, S., and Chen, H. "Browsing in Hypertext: A Cognitive Study," *IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS,* 22(5) 1992, pp 865–884.

[3]. Chen, H., Houston, A.L., Sewell, R.R., and Schatz, B.R. "Internet browsing and searching: User evaluation of category map and concept space techniques," *Journal of the American Society for Information Science* 47(9) 1998, pp 582– 603.

[4]. Chen, H., Schuffels, C., and Owig, R. "Internet categorization and search: a machine learning approach," *Journal of Visual Communications and Image Representation Science* 7(1) 1996, pp 88– 102.

[5]. Lai, H., and Yang, T.-C. "A system architecture for intelligent browsing on the web," *Decision Support Systems* 28(3) 2000, pp 219-239.

[6]. Maedche, A., and Steffen, S. "Discovering Conceptual Relations from Text," in: *Proceedings of the 14th European conference on artificial intelligence*, 2000, pp. 321-325.

[7]. Nallapati, R., Feng, A., Peng, F., and Allan, J. "Event Threading within News Topics," *Proceedings of the ACM Thirteenth International Conference on Information and Knowledge Management*, 2004, pp. 446-453.

[8]. Ong, T.-H., Chen, H., Sung, W.-k., and Zhu, B. "Newsmap: a knowledge map for online news," *Decision Support Systems* 39(1) 2005, pp 583-597.

[9]. Pepper, S. "The TAO of Topic Maps Finding the Way in the Age of Infoglut," in: *XML Europe 2000*, Paris, 2000.

[10]. TDT "TDT 2004: Annotation Manual Version1.2," 2004.

[11]. Weng, S.-S., Tsai, H.-J., Liu, S.-C., and Hsu, C.-H. "Ontology construction for information classification," *Expert systems with Application* 2005, pp 1-12.

[12]. Wu, C.-H., Lee, T.-Z., and Kao, S.-C. "Knowledge discovery applied to material acquisitions for libraries," *Information Processing and Management* 40(4) 2004, pp 709–725.

[13]. Yang, Y., Ault, T., Pierce, T., and Lattimer, W.C. "Improving text categorization methods for event tracking," in: *Proceedings of the Annual International ACM SIGIR Conference*, 2000.

[14]. Yeh, J.-Y., Ke, H.-R., Yang, W.-P., and Meng, I.-H. "Text summarization using a trainable summarizer and latent semantic analysis," *Information Processing and Management* 41(1) 2005, pp 75-95.