Term Mining for Relation Visualization and Exploration: The Case of Agricultural News Reports in Taiwan[3]

Yuen-Hsien Tseng[1], Yi-Yang Lee[2], and Te-Yi Chan[3]

[1]*samtseng@ntnu.edu.tw*
National Taiwan Normal University, No.162, Sec. 1, Heping East Road, Taipei City (Taiwan)

[2]*d24257@tier.org.tw*
Taiwan Institute of Economic Research, Biotechnology Industry Study Centre, Taipei (Taiwan)

[3]*tychan@mail.stpi.org.tw*
Science & Technology Policy Research and Information Center. No. 106, Heping E. Rd., Sec. 2, Taipei (Taiwan)

Abstract
An efficient term mining method to build a general term network is presented for term relation visualization and exploration. Terms from each document in the corpus are first identified. They are subject to an analysis for their association weights, which are accumulated over all the documents. The resulting term association matrix is used to build a general term network. A set of terms having similar attributes can then be given to extract the desired sub-network from the general term network for visualization. This analysis scenario based on the collective terms of the same type enables evidence-based relation exploration. Our application examples show that term relations, be it causality, coupling, or others, can be effectively revealed and verified by the underlying corpus. This work contributes by presenting an efficient and effective term-relationship mining method and extending the applicability of term networks to a broader range of informetric tasks.

Introduction
Network analysis is a powerful methodology in informetrics for various knowledge explorations. Different networks, including citation networks, social networks, and term networks, have been proposed for specific applications. For example, citation networks are used to derive various relationships among documents, such as topical relatedness, evidence of knowledge diffusion, or paradigm shift. Social networks are used to investigate the individual relationship sociological motivated. And term network based on document co-occurrence is often employed to describe the topical landscapes of various knowledge domains. (Jacobs, 2002; Noyons & van Raan, 1998)

Among the above networks, term networks seem to be under-explored so far, due to the un-clear relationship inherited from their document-level co-occurrence. To allow more in-depth exploration like those in citation and social networks, we believe that a sentence-level co-term network would be more suitable because terms co-occur in the same sentences tends to exhibit clearer relationship than those in the same documents only. As an evidence, (Inui, Inui, & Matsumoto, 2005) showed that for causal relation, cause and effect are in the same sentence in 80% of the case. That means among various relations that we may find from the same sentences (or similar-size snippets), some of them are causality and this can be easily checked by the relative short snippets containing those terms. Likewise, other relations can also be verified instantly by retrieving their corresponding snippets, which nowadays is nearly a trivial task thanks to the rapid development of information retrieval tools and technology.

However, such term networks based on snippet co-occurrence seems to be more costly to build than those based on document co-occurrence using traditional methods. Fortunately, we have a method that reduces the building complexity by an order of magnitude.

This article was motivated by the above observation and technique. By combining network analysis and information retrieval technology, such term networks can be analyzed with

[3] This work was supported in part by Science & Technology Policy Research and Information Center and National Science Council under the grant NSC 97-2631-S-003-003- .

Proceedings of ISSI 2009, pages 189–193.
Edited by B. Larsen and J. Leta.

evidence at hand without sacrificing efficiency. As such, our aimed applications are to explore the network created for any given list of terms (with the same attributes such as those under the same taxonomy or name entity) in relation to any text corpus. This application is different from those in information retrieval, where a seed term is used to trigger other associative terms for better query term selection (Zhang, 2008) or for serendipity search. It is also different from the bottom-up topical landscape mapping (although the resulting network is able to do so), because it is more about the verification of a top-down analysis, i.e., to find support for relation evidence from the corpus of interest.

Method

Traditional co-term network is built by representing terms in vector form which denotes the occurrence of the terms in all documents. The terms for the network can be obtained from a given list, output of an NLP parser, or any key term extraction algorithms (Tseng, 1998). Term similarities are then calculated based on these vectors to form the similarity matrix of terms (Salton, 1989), which is the input to the co-term network. If there are n terms from m documents, the time complexity can be on the order of $O(n^2m)$, where m steps are required to calculate similarity between any of n^2 term pairs.

The major difference of ours from the above is that we limit the terms to be associated to those that co-occur in the same logical segments of a smaller text size, such as a sentence. Association weights are computed in this way for each document and then accumulated over all documents. This changes it into a roughly $O(mk^2s)$ algorithm, where k is the average number of selected key terms per document and s is the average number of sentences in a document. As can be seen, the larger the n and m, the bigger the difference between $O(mk^2s)$ and $O(n^2m)$, because k can be kept less than a constant and so can s by breaking large documents into smaller ones.

Specifically, key terms identified from each document are first sorted in decreasing order of their term frequencies (*TF*), or *TF* x *Term_Length*, or other criterion such as *TF* x *IDF* (Inverse Document Frequency) if the entire collection statistics are known in advance. Then the first k terms are selected for association analysis. A modified Dice coefficient was chosen to measure the association weights as:

$$wgt(T_{ij}, T_{ik}) = \frac{2 \times S(T_{ij} \cap T_{ik})}{S(T_{ij}) + S(T_{ik})} \times \ln(1.72 + S_i)$$

where S_i denotes the number of sentences in document i and $S(T_{ij})$ denotes in document i the number of sentences in which term T_j occurs. Thus the first term is simply the Dice coefficient similarity. The second term $ln(1.72+S_i)$, where ln is the natural logarithm, is used to compensate for the weights of those terms in longer documents so that weights in documents of different length have similar range of values. This is because longer documents tend to yield weaker Dice coefficients than those generated from the shorter ones. Association weights larger than a threshold (1.0 in our implementation) are then accumulated over all the documents in the following manner:

$$sim(T_j, T_k) = \frac{\log(w_k \times n / df_k)}{\log(n)} \times \sum_{i=1}^{n} wgt(T_{ij}, T_{ik})$$

where df_k is the document frequency of term k and w_k is the width of k (i.e., number of constituent words). It is noted that the above procedure is an approach to derive a general term network. In fact, the association weight can be changed to a numeric indicator representing the (degree of) existence of a predefined set of relations between the term pairs. In this way, a specialized term network subject to the predefined term relationship can be captured.

Computation of the similarities among all term pairs can be carried out as the inverted index file in a retrieval system is constructed for the entire collection. Weights of term pairs from

each document are calculated and accumulated just like the index terms accumulating their document frequencies and postings. In this way, a global term relation structure can be obtained efficiently. As an example, for a collection of 381,375 documents (469 MB of texts), it takes only 50 minutes on a notebook computer for indexing, key term identification, and term association computation. Compared to traditional approaches, this method improves the efficiency drastically while maintaining sufficient effectiveness.

Data

In a project supported by Science & Technology Policy Research and Information Center (STPI) in Taiwan to analyze the social and economic demands for agriculture-related topics and events, a total of 14,230 Chinese news stories were collected from the Great China Knowledge Bank database. These stories range from May 2005 to May 2008 and all are from the agriculture category tagged by the database provider. Our algorithm takes 70 seconds to identify key terms, extract term relationship, and build the document index on a notebook computer. In the resulting general term relationship, there are 14,147 terms consisting of a total of 69,735 relations. The general term relation structure was then matched with four sets of terms concerning the products, techniques, functions, and statistic nouns of a broad range of agricultural topics/events provided by STPI for exploration. In the following, we show the results from the last two sets to demonstrate the feasibility of our approach.

Results

For the first set of terms, a list of 483 predefined terms concerning the effect/function of agricultural topics/events was provided for exploration. Using this list to filter the general term network, a sub-network consisting of 222 nodes and 200 links was obtained and visualized, as shown in Figure 1. A dozen of salient cliques representing coherent topics can be readily identified, such as the topics of leisure agriculture, agricultural disaster, and agricultural quality issue, as circled in Figure 1. These 3 clusters were selected and translated into English, as shown in Figure 2, by the Google translation service with a few terms manually corrected. It can be seen that the quality of crops or fruits relates to the taste, colour, and sweetness of the products (the left part of Figure 2), which may eventually affect the consumer confidence or annual turnover, as evidenced by the relations extracted from the news reports. The agricultural disaster (in the middle) mainly comes from heavy rain and pests in Taiwan. The ensuing events affected include flowering and production or maturation periods of the crops. Some policy issues (responsive programs or counter measures) such as agricultural aid and natural ecological conservation are often mentioned when it comes to agricultural disaster. As to leisure agriculture, it relates to the scenery, tourism, eco-landscape, and health and life span of citizens. A seemingly unrelated term is Soil and Water Conservation. After examining the snippets from the dataset, the evidence becomes clear as shown in Table 1. Without prior knowledge of the mission of government bureaus, one of our authors who majors in agricultural science never thought of the relatedness between these two concepts represented by the two terms.

Another set of terms provided for exploration consists of 83 statistic nouns. They were visualized in Figure 3. The dataset does not contain plenty of them, but some basic economic causality was revealed in the corpus, such as the economic growth rate and consumption. Based on the coupling effect between these two events, the government in Taiwan has just issued consumption vouchers worth of 120 USD to each citizen in the hope to stimulate the sluggish economy. Other ripple effects are also observed from and supported by the corpus: starting from the oil price, rice price, commodity price, inflation, unemployment, and economy growth are all affected, which precisely reflects the phenomenon happened in mid 2008.

Discussion and Conclusions

Just as the essential functions of documents or individuals, terms are also important entities in network analysis. Detailed relations between key terms are enormous and some of them are valuable for investigation. However, the methodologies to effectively explore them are not as salient as those for citation and social networks. This article tries to propose and demonstrate an approach viable for this purpose. As shown in the above examples, term relations, be it causality, coupling, or others, were revealed and verified by the corpus. Furthermore, such analysis can be done without starting from term similarity computation, but by extracting the sub-network directly from the whole network already built, thus allow dynamical network exploration for various sets of terms without re-scanning over all the texts.

In conclusion, this paper contributes by presenting a practical term relation mining method, proposing evidence-based analysis for term network, and demonstrating the feasibility and usability of the above claims. Future work will implement the above methodology in an integrated environment for ease of use. Another future direction would be to apply the ideas to more practical issues, such as those to sieve relevant terms for measuring civic scientific literacy in order to understand where to improve science education in a democratic society that is more and more technologically demanding (Brossard & Shanahan, 2006).

References

Brossard, D., & Shanahan, J. (2006). Do They Know What They Read? Building a Scientific Literacy Measurement Instrument Based on Science Media Coverage. Science Communication, 28(1), 47-63.

Inui, T., Inui, K., & Matsumoto, Y. (2005). Acquiring causal knowledge from text using the connective marker tame. Transactions on Asian Language Information Processing, 4(4), 435-474.

Jacobs, N. (2002). Co-term network analysis as a means of describing the information landscapes of knowledge communities across sectors. Journal of Documentation, 58(5), 15.

Noyons, E. C. M., & van Raan, A. F. J. (1998). Mapping Scientometrics, Informetrics, and Bibliometrics. Retrieved November 23, 2006, from http://www.cwts.nl/ed/sib/home.html

Salton, G. (1989). Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. MA: Addison-Wesley.

Tseng, Y.-H. (1998, Aug. 24-28). Multilingual Keyword Extraction for Term Suggestion. Paper presented at the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '98, Australia.

Zhang, J. (2008). Visualization for Information Retrieval: Springer.

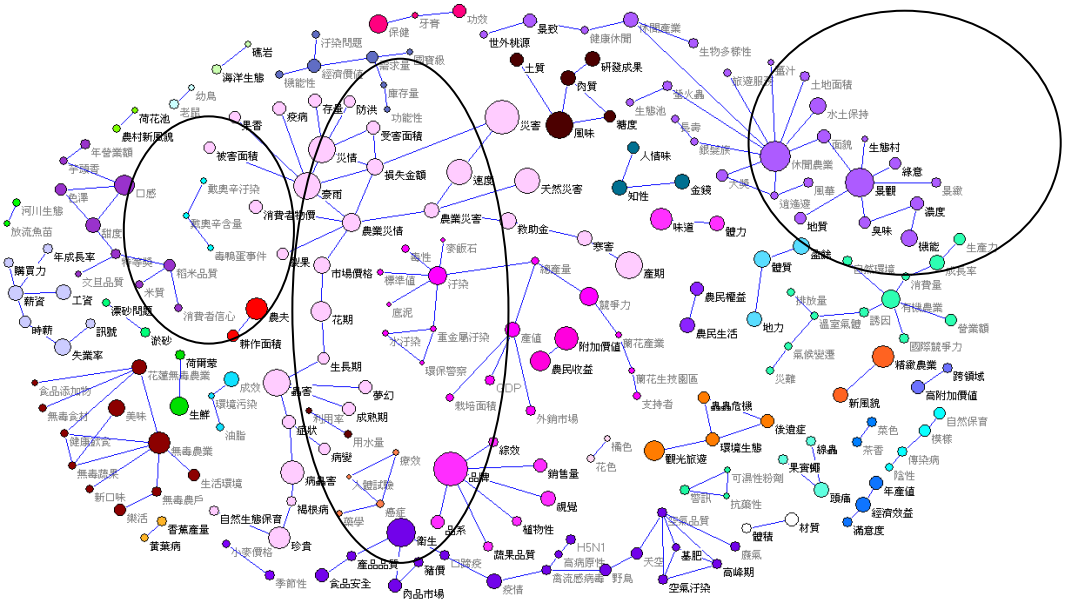

Figure 1. Term network for the set of function/effect terms.

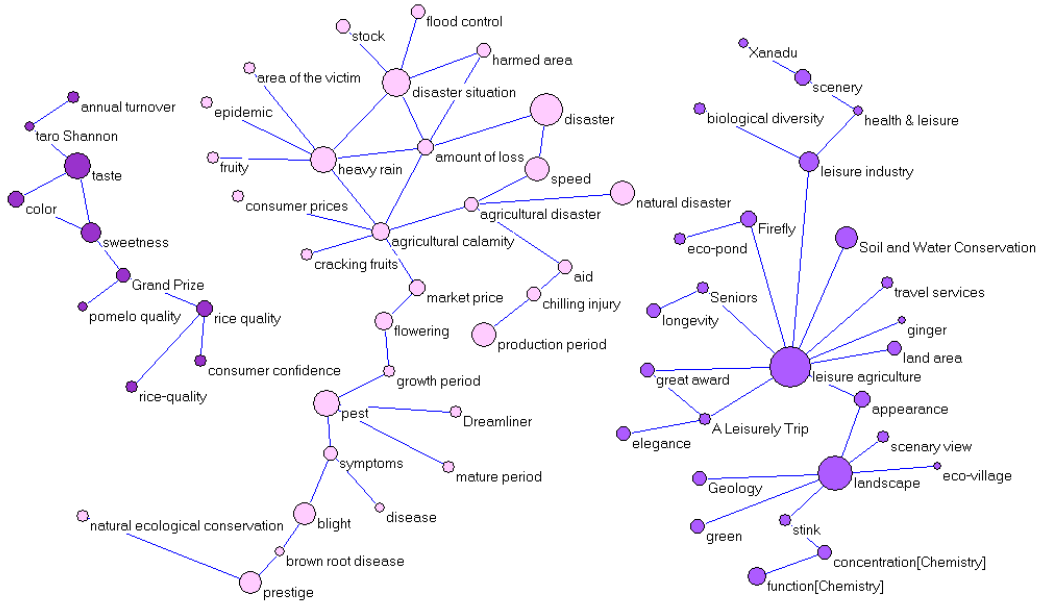

Figure 2. Selected sub-networks from Figure 1 for discussion.

Table 1. Two example snippets retrieved from the dataset showing the evidence for the highlighted terms (Texts were translated into English for illustration.)

1. □□□□□ □□□□□ / The beauty of North County farm was yet to be known After performing **soil and water conservation** and flood prevention stage by stage, the **Soil and Water Conservation** Bureau drew up a **leisure agriculture** park **...**
2. □□□□□□□□□□ □□□□ / Soil and Water Conservation officials visit the classical village for future strategic planning The successful development of a local tea and coffee-based **leisure agriculture**, especially the combination of ecological conservation practices, **water and soil conservation** efforts, attracts large numbers of tourists **...**

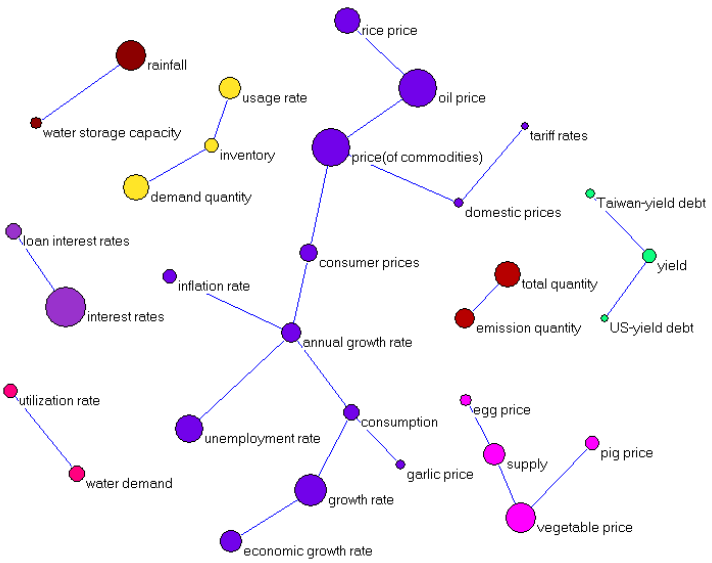

Figure 3. Term network for the set of statistic nouns.