Applied Mathematics & Information Sciences 5 (2) (2011), 9S-16S
– An International Journal

GRAPH: A Domain Ontology-driven Semantic Graph Auto Extraction System

Chunying Zhou, Huajun Chen and Jinhuo Tao

College of Computer Science, Zhejiang University, Hangzhou, China,
Email Address: cyzhou@zju.edu.cn

Received January 20, 2010; Revised January 20, 2011

This paper presents sGRAPH – a domain ontology-driven semantic graph auto extraction system used to discover knowledge from text publications in traditional Chinese medicine. The traditional Chinese medicine language system (TCMLs), composed of an ontology schema and a knowledge base containing 153,692 words and 304,114 relations, is used as the domain ontology. The sGRAPH comprises two components: a user interface that interacts with users and the domain ontology-based semantic graph extraction algorithm. This algorithm is divided into five steps: text processing, semantic graph extraction, graph identification, keyword-based semantic graph search and the selectable enrichment to the knowledge base. When the knowledge base of TCMLs is used, the domain-specific words are extracted from sentences more accurately; and the hierarchical structure of the ontology can also be used to help identify the extracted graphs. The algorithm not only can extract relations between words that have already been annotated by relations in the knowledge base but also can predict the relations between words that have never been annotated by relations. The sGRAPH was developed and evaluated by extracting semantic graphs from 2000 publications which predicted 6778 relations that have never been found.

Keywords: Semantic graph extraction, domain ontology, text mining, knowledge acquisition, traditional Chinese medicine.

1 Introduction

Traditional Chinese medicine (TCM) has been preferred to provide treatments of diseases and to take care of the health of Chinese people for over thousands of years [1]. A large amount of valuable medical knowledge has been accumulated in text publications. Two methodologies of the knowledge acquisition (KA) from texts are distinguished [2]. The first methodology develops general-purpose natural language processing algorithms to understand the text and to extract concepts and their relations from texts [3-5]. However, TCM is a special domain in which most publications were written in Chinese, even ancient Chinese, that is extremely difficult for computer systems to understand so far. The second methodology relies upon interactions with knowledge

engineers or uses the existing semantically tagged knowledge base [2, 6-8] to extract the knowledge from texts.

Because of the large amount of manual work undertaken by TCM researchers and engineers over last ten years, a knowledge base containing 153,692 words and 304,114 relations has been built. TCMLs that has an ontology schema and this knowledge base is used as the domain ontology. Thus the second methodology is more suitable to be used in TCM.

This paper presents sGRAPH that is a domain ontology-driven semantic graph auto extraction system of the second methodology. It consists of a friendly user interface and an ontology-based semantic graph auto extraction algorithm. The algorithm is divided into five steps: (1) text processing where TCMLs is used as the dictionary to extract valid domain-specific words, (2) semantic graph extraction, (3) semantic graph identification, (4) keyword-based semantic graph search and (5) selectable knowledge enrichment to the knowledge base. In addition sGRAPH has three technical features:

- It supports highly efficient semantic graph auto extraction from text documents;
- It supports keyword-based semantic graph search to help users easily browse the extracted graphs;
- It supports knowledge base enrichment using the extracted graphs users select.

The sGRAPH system was developed well. To evaluate its efficiency experiments were run on 2000 publications which produced 7038 relations that contain 6778 relations that have never been found and are good complementary to the knowledge base.

2 Systematic Methodology

As is shown in Fig.1, the sGRAPH comprises two components: the user interface that interacts with users to get documents and to demonstrate the processing results, and the kernel semantic graph extraction algorithm that is to extract latent knowledge from texts automatically, to support keyword-based knowledge search and to enrich the knowledge base using the selected knowledge of users. The algorithm is divided into five steps:

- Text processing: The valid domain-specific words are extracted from a sentence, where TCMLs is used as the domain dictionary to improve accuracy;
- Semantic graph extraction: The frequent-word group is first discovered from a sentence and is stored in a word-vector model. Then potential relations are extracted from this word group;
- Semantic graph identification: Determine whether the two words have been annotated by relations in the knowledge base. If yes, find all relations and

compute their probabilities of occurrence in the knowledge base. If no, predict the relation by using mapping information from the words to concepts and the hierarchical structure of the ontology and compute probabilities of occurrence of the predicted relations in the knowledge base;

- Keyword-based semantic graph search: after the above steps, users can search their graphs of interest from the extracted graphs by using keywords;
- Selectable knowledge base enrichment: the existing knowledge base can be enriched by the extracted knowledge which is approved by the user;

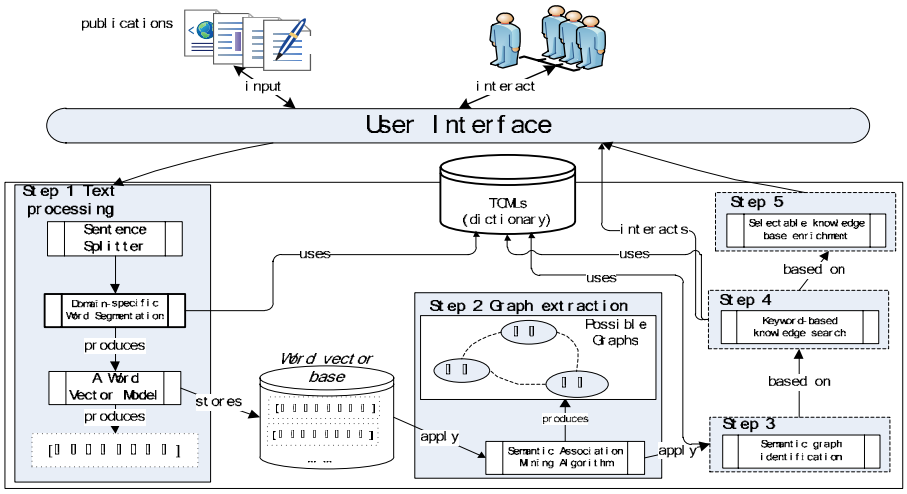

Fig.1 the architecture consists of the user interface and the kernel graph extraction algorithm.

2.1 Traditional Chinese Medicine Language System (TCMLs)

The TCMLs is defined using RDF [9], which has an ontology schema of 127 concepts and 56 relations, and a knowledge base of 153,692 words and 304,114 relations. It provides three types of knowledge for the kernel algorithm:

- Mappings from words to ontological concepts that are used to predict relations;
- Aliases of words, such as “general” is an alias of “rheum officinal”, which can be used to help extract domain-specific words in the step of text processing and predict relations in the step of graph identification;
- Semantic relations that have been semantically annotated between words in the knowledge base, which helps the graph identification.

2.2 Text Processing

A sentence is the basic processing unit in this step. As is shown in Fig.1, a document is firstly split into sentences by scanning and splitting it into sentences if encountering a period separator. TCMLs is used to segment words from a sentence. A vector model in Definition 2.1 is used to store extracted words and their position-sequence information.

Definition 2.1 A vector-space model $V_s = (W_1, W_2, ...W_n)$ stores sequenced words that are extracted from a sentence S, where W_i is the i_{th} valid words extracted from S.

2.3 Semantic Graph Extraction

After the text is processed, a vector storing sequenced words extracted from a sentence is obtained. This step is to discover the frequent-word group and to extract semantic graphs. The association rule mining method [10] is used to discover the frequent-word group. It assumes that words that always co-occur are interrelated. As is shown in Fig.2, a frequent-word group $[f_{1i},..., f_{1j}]$ is discovered from the vector $[w_{11},..., w_{1n}]$ in which words $f_{1i}, f_{1(i+1)}$ are at most K words spacing. The relative positions of two words in the frequent-word group probably determine whether a relation between them exists or not. This method assumes that the probability of the existence of a relation is higher if the positions of the two words are closer. Two interrelated words are assumed to have at most M words spacing in the frequent-word group. A new vector in Definition 2.2 is proposed to store the relative positions of words.

Definition 2.2 $VM_i = \{(r_{jk}, w_{jk}) \mid 0 < j < k \le n\}$ is a vector of $V_i = (W_{i1}, ...W_{in})$, where $r_{jk} = \frac{1}{k-j}$ $(k \ge j)$ is the relative position distance between W_{ij} and W_{ik}.

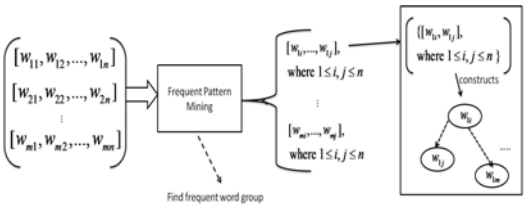

Fig.2 Extracting graphs from the word vectors.

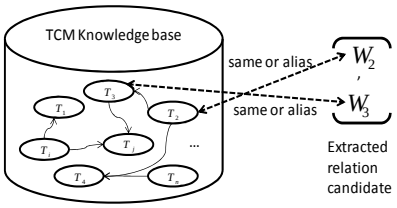

Fig.3 an annotated relation between W_2 and W_3.

2.4 Semantic Graph Identification

A frequent-word group is discovered from a sentence. The semantic graph representing this sentence is composed of words as nodes and relations between words as edges. The crucial step of identifying a graph is to identify its relations. The relations differentiate into two types: the annotated relation and the never-annotated relation. The relation is an annotated relation in two situations: the two words or their aliases have been annotated by at least one relation in the knowledge base. Fig.3 shows an example.

The candidates of an annotated relation are relations between the two words or their

alias words in the knowledge base. Fig. 4 (a) shows an example of an annotated relation, where relations between words T1 and T2 in the knowledge base are denoted as $S=\{R_1,....,R_n\}$. The identification step is to find relations ($R_i \in S$) having higher probabilities of occurrence. The probability of R_i over S is defined in Definition 2.3.

Definition 2.3 The occurrence probability of R_i is $P(R_i)=\dfrac{num(i_1,R_i,i_2)}{\sum_{R_j \in S} num(i_1,R_j,i_2)}$, where i_1, i_2 are or aliases of T1 and T2, S is the set of relations between T1 and T2.

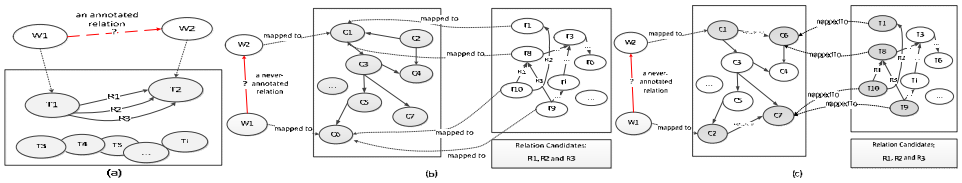

Fig. 4 (a) identifying an annotated relation, (b) identifying a never-annotated relation; (c) identifying an never-annotated relation in which no words are mapped to concepts directly.

To identify a never-annotated relation one uses the ontological schema as the mediation layer between the extracted words and the knowledge base. The procedure of identification comprises five steps: (1) find concepts (C_1,C_2) to which extracted words are mapped; (2) find words also mapped to C_1,C_2; (3) find relations annotating the pair of words, one of which is mapped to C_1 and another is mapped to C_2; (4) if there are no other words mapped to C_1,C_2, find concepts to which C_1,C_2 are connected to by 'rdf:subClass' in the ontology and go back to step 2 again taking the concepts as C_1,C_2; (5) compute the probability of occurrence in Definition 2.4 of every relation candidate. In the example of Fig.4 (b) there are words mapped to C_1,C_2, while in Fig.4(c) no words are mapped to C_1,C_2 directly.

Definition 2.4 For two words T1, T2 the probability of occurrence of a relation R_i is $P(R_i)=\dfrac{\sum_{i_1\in S_{C_1},i_2\in S_{C_2}} num(i_1,R_i,i_2)}{\sum_{i_1\in S_{C_1},i_2\in S_{C_2}}\sum_{R_j\in R_{i_1,i_2}} num(i_1,R_j,i_2)}$, where T1, T2 are mapped to concepts C1, C2, S_{C1}, S_{C2} contain all words mapped to C1 and C2, R_{i_1,i_2} contains all relations between i_1 and i_2, $num(i_1,R_i,i_2)$ is the number of R_i in R_{i_1,i_2}.

A semantic graph is an item of the Cartesian product of all sets of relations between the pairs of extracted words. Fig. 5 shows an example. In this example the relation between W2 and W1 is found to be the type of annotated relation and three relations (R1,

R2, R3) are found in the knowledge base that are annotated between W2 and W1. The relation between W3 and W2 is of the type of never-annotated relation. Three relations (R7, R5, R6) are predicted. There are 3*3=9 possible graphs constructed by the extracted relations. The probability of a graph is defined as the product of its relations. The graph constructed by R2 and R5 has the highest probability, computed as 0.32*0.432.

2.5 Keyword-based Semantic Graph Search and Selectable Knowledge Enrichment

The semantic graph search allows users easily browse the extracted semantic graphs. The keyword-based search is a simple and efficient way to help users find their interested knowledge. It supports searching graphs that contain nodes the labels of which are matched to the keywords. The selectable knowledge enrichment allows users to select their knowledge of interest from extracted graphs to enrich the knowledge base. The users firstly search their graphs of interest from the extracted graphs by using the keyword-based search and select their approved graphs from the results to enrich the knowledge base.

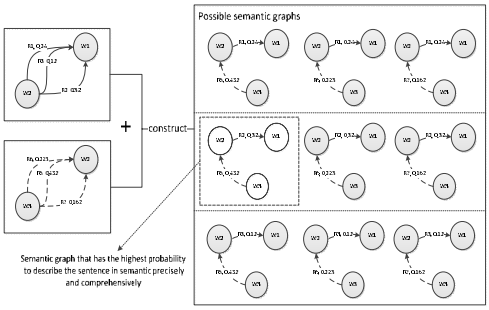

Fig.5 The graph of words W1, W2 and W3.

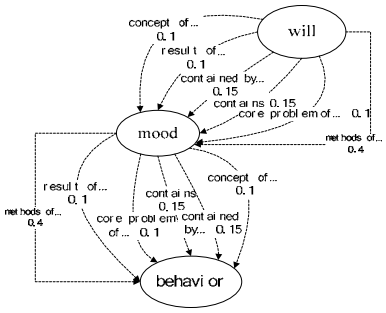

Fig.6 an example of the extracted graphs.

3. Implementation & Evaluation

sGRAPH was developed using JAVA with a friendly user interface shown in Fig.7. In the frequent-word group discovery algorithm, the parameter K is assigned to 1 and M is assigned to 0. It provides a user interface that interacts with users, and especially it provides users with four functionalities: (1) allow users to select publications to be processed and list them; (2) semantic graph auto extraction from the listed textual files the statistical results of which is demonstrated through the user interface to users; (3) keyword-based knowledge search that allows users to browse all extracted graphs; (4) enriching the existing knowledge base using selected and approved knowledge by users.

To evaluate the performance of the sGRAPH system, 2000 TCM publications were batch processed to extract semantic graphs (knowledge). The experiments produced a total of 7038 relations that consist of 260 annotated relations and 6778 never-annotated relations. In particular these 6778 predicted never-annotated relations were a good complement to the existing knowledge base in TCMLs. Fig.6 shows an example of an

extracted graph from the frequent word group ([‘意志’ (will), ‘情绪’ (mood), ‘行为’ (behavior)]). Two pairs of interrelated words ([‘will’, ‘mood’], [‘mood’, ‘behavior’]) were extracted. The relation between ‘will’ and ‘mood’, and the relation between ‘mood’ and ‘behavior’ are both of the type of never-annotated relation. In the figure they are drawn by dashed lines that are associated with their predicted probabilities as the weights.

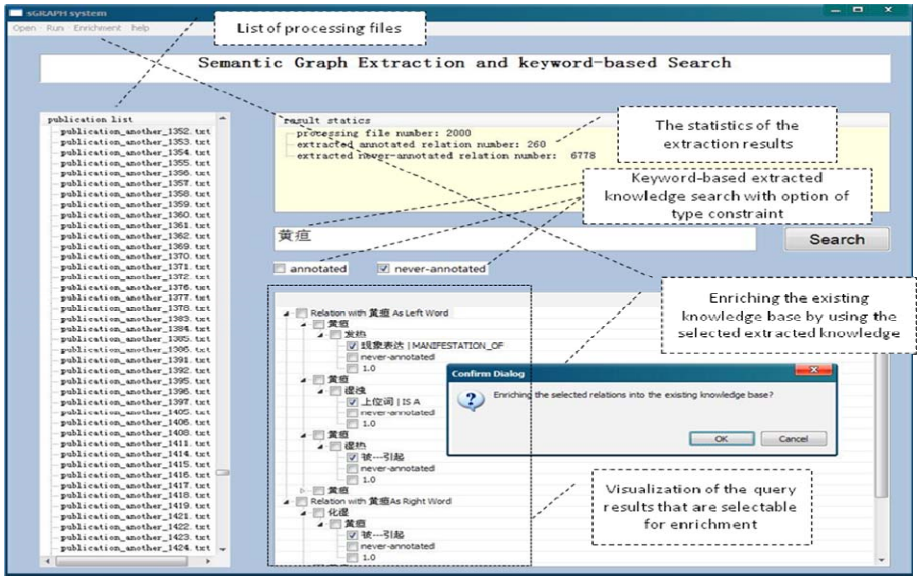

Fig.7 The implementation of the sGRAPH system.

4. Conclusion and Future Work

In this paper a domain ontology-driven semantic graph auto extraction system named sGRAPH is presented and implemented for discovering knowledge from publications in TCM. The system consists of a user-friendly interface and the semantic graph extraction kernel algorithm. Particularly, by using the TCMLs as the domain ontology, this system does not support extracting relations that have already be annotated in the knowledge base. It does support predicting relations between words that have never been annotated. The predictions are a good complement to the knowledge base. Furthermore sGRAPH provides a user-friendly interface for users to search over the extracted knowledge and to add selected knowledge to enrich the knowledge base. More work is needed to optimize the performance of processing large scales of publications. Map-Reduce [11] and Hadoop [12] can be used to leverage the parallel computing on the knowledge discovery from large scales of textual publications in TCM to improve the time performance.

Acknowledgements: This work is funded by NSFC Programs NO. NSFC61070156, NSFC60873224, important programs of Zhejiang Sci-Tech Plan (2008C03007).

References

[1] G. Nestler, Traditional Chinese medicine, in: Medical Clinics of North America, 86(1) (2002), 63-73.

[2] C. Cao, Medical knowledge acquisition from the electronic encyclopedic texts, in: Artificial Intelligence in Medicine, Berlin: Springer-Verlag, 2101 (2001) 268--271.

[3] R. Hull, F. Gomez. Automatic acquisition of biographic knowledge from encyclopedic texts, in: Expert Systems with Applications, 16(3) (1999) 261--270.

[4] S. D. Richardson. Determining similarity and inferring relations in a lexical knowledge base, in: Doctoral Dissertation, City University of New York New York, NY, USA, (1997)

[5] Z. Wu, X. Zhou, B. Liu, H. Chen. Text mining for finding functional community of related genes using TCM knowledge, in: Proceedings of 8th European Conference on Principles and Practice of Knowledge Discovery in Databases, (2004) 459--470.

[6] C. Cao, H. Wang, Y Sui, Knowledge modeling and acquisition of traditional Chinese herbal drugs and formulae from text, in: Artificial Intelligence in Medicine, 32(1) (2004) 3-13.

[7] R. Lu, C. Cao, Towards knowledge acquisition from domain books, in: Current Trends in Knowledge Acquisition, edited by B. Wielinga, Amsterdam: IOS Press, (1990) 289-301.

[8] D.R. Swanson, N.R. Smalheiser, An interactive system for finding complementary literatures: a stimulus to scientific discovery, in: Artificial Intelligence, 91(2), (1997) 183--203.

[9] Information on http://www.w3.org/RDF/

[10] R. Agrawal, R. Srikant. Fast algorithms for mining association rules, in: Proceedings of 20th Int. Conf. Very Large Data Bases, Chile, (1994) 487—499.

[11] J. Dean, S. Ghemawat. MapReduce: simplified data processing on large clusters. In: OSDI'04: Sixth Symposium on Operating System Design and Implementation, San Francisco, CA, 2004.

[12] M. F. Husain, P. Doshi, L. Khan et al. Efficient Query Processing for Large RDF Graphs Using Hadoop and MapReduce: Technical Report UTDCS-41-09, Department of Computer Science of the University of Texas at Dallas, 2009.

Chunying Zhou now is a PhD candidate in the College of Computer Science, Zhejiang University, China. She received the Bachelor degree in Computer Science from Zhejiang University in 2006. Her research interests are in the areas of Data Mining, Knowledge Discovery, Semantic Web and Social Computing.

Hua-jun Chen is now an Associate Professor in the College of Computer Science, Zhejiang University, China. He received his Ph.D. degree from College of Computer Science, Zhejiang University, China, in 2004. His research interests include Cloud Computing, Bioinformatics and the Semantic Web.

Jinhuo Tao is now a masters candidate in the College of Computer Science, Zhejiang Univesity, China. He received the Bachelor's degree in Computer Science from Zhejiang University in 2008. His research interests are in the areas of Knowledge Discovery and Semantic Web.