International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol. 8, No. 2 (2015), pp. 57-64
http://dx.doi.org/10.14257/ijsip.2015.8.2.07

A Graph-based Algorithm to Build Knowledge Map for Minority Languages

Lirong Qiu

School of Information Engineering, Minzu University of China
Beijing, China
E-mail:qiu_lirong@126.com

Abstract

Knowledge mapping will undoubtedly bring great convenience to application users for being behind the strong support of knowledge base. In this paper, we study how to discover the evolution of knowledge map in multi-languages. Our approach is uniquely designed to capture the rich topology of semantic items and to link the sub-graph to a global knowledge map. Instead of building a knowledge map start from scratch, we conceptually define semantic classes as a quantized unit of evolutionary link in sub-graph and discover new knowledge with multi-language dictionaries. Discovered new knowledge items are then connected to form an evolution knowledge map using a measure derived from the underlying semantic classes. We integrate these noisy items and entities into a unified probabilistic knowledge map using ideas from graph-based algorithm.

Keywords: *knowledge map, graph-based model, ontology integration, knowledge mapping*

1. Introduction

During the long historical development process, each minority language in China has a unique ethnic customs, whose civilians create the great culture with their diligence and wisdom. The minority language websites also reflect the national characteristics, such as the language, religion, culture and art.

With the fast development and widely use of internet, the number of websites in minority languages is increased tremendously in China. The growth speed of Mongolian, Tibetan and Uygur forums and blogs is amazing: the growth rate of minority ethnic netizens is more impressing compared with netizens of China. For example, the raise of Tibetan netizen is 86%, much higher than the average level of whole country.

Gathering, reorganizing comprehensive utilization of national websites can promote the ethnic culture, be beneficial to the accumulation for social integrated administration experience, and provide objective data and timely information service for people using minority languages. However, the intensive processing technology of minority languages is relatively backward, which is unable to provide effective technical support on entity extraction templates from those websites, and on hot issue tracking.

Distinguished from English, Chinese and other mainly useful languages, the problem of data sparseness supposed to be taken into account in the minority language processing technology. "Small data" does need intensive processing, of which knowledge graph is an effective means.

A Knowledge Map (KM) is a discipline-specific form of concept map, which is generally a representation of "knowledge about knowledge" rather than of knowledge itself [1]. Capturing and representing knowledge is critical in knowledge management [2]. Knowledge mapping provides a basic tool for managers and employees to retrieve necessary knowledge and to analyze the relationships between knowledge sources [3].

This work aims to orient to the needs of monitoring public opinion, to study the

ISSN: 2005-4254 IJSIP

creation and fusion techniques of knowledge map in minority languages, based on the techniques used on website feature templates, such as the extraction of entity knowledge, fusion and verification of new knowledge, which is able to extend the entity knowledge of the Uyghur knowledge integration map. The building technology research of that map can greatly alleviate the problem of minority language data sparseness. Also, it can provide effective technical support for monitoring public opinion.

Based on multiple minority language and test co-existing network information resource, this paper will try to utilize the existing semantic resources and use the method of analyzing and linking semantic entity to realize the semantic connections for multiple languages information resource (Chinese, Mongolian, Tibetan and Uygur network information resources will be included). A new algorithm will be proposed to combine the ethnic text knowledge profile and to verify the new generated knowledge.

The remainder of this paper is organized as follows. In the next section we will describe the motivation of our work with an overview of related work. In Section 3 we outline the methodology of the work and the main challenges. We describe the main research content with those challenges in Section 4. In Section 5 we draw the conclusions and mentions future work.

2. Related Work

One of the ultimate aims of natural language processing domain and artificial intelligent research domain is to develop the test reading and understanding ability like human being for computer. Lacking the knowledge base which can support intelligent deduction and decision for computer is one of the bottle necks [4]. Scientists and research agents have been working hard to build such knowledge base and corpus.

Knowledge maps are the representation of 'detailed, interconnected, nonlinear thought' [6]. Knowledge mapping serves as both an instructional and assessment tool to illustrate both declarative knowledge (facts, definitions, and statements) and to a lesser extent, procedural knowledge (how something is done, *e.g.*, processes for problem solving, plans, decision making).

In 1985 Princeton University had developed English vocabulary knowledge library: WordNet. Since 1997 FrameNet, a knowledge library based on frame semantic is started to build in University of California, Berkeley. Until now, 1164 frames, 8180 word elements and totally 194 thousands sentences' frame semantic information are tagged.

In 2007 DBpedia, an online data connection knowledge library project, had realized the structure data's subtraction from Wikipedia vocabulary entries. It can also build the connection between other data cluster and Wikipedia and public these data to internet in the form of connection data. These online connection data can be used by online network application, social interaction website and other online connection data knowledge library.

In May of 2009, Wolframalpha, a new search engine developed by Wolfram Research, Inc. was published [14]. This engine can understand the question and give the expected answer. For example, when Wolframalpha is questioned about the height of Mount Qomolangma, it not only can present the height above sea level, but also display information on the geographical location, near cities and diagrams which is relative to Mount Qomolangma.

In 2012 Google published its latest research achievement, knowledge map, in the past two years. The main function of knowledge map is that once a word or a phrase is typed in the Google search engine, a new column will be added to display the relative information about this word entry at the right side of traditional search result. This relative information comes from public network resources such as Freebase, Wikipedia, CIA World, Facebook, *etc*. And abundant of collected and marshaled materials from other website are also included.

In Chinese information processing field, researchers made lots of study on the building

techniques about Chinese knowledge map and Chinese knowledge application techniques in search engine based on some latest research results about the semantic knowledge library such as HowNet, synonym library and HNC. In 2012 Sogou published its knowledge map, "knowledge cube" which can provide the knowledge connected information query service.

Currently there is not much research result about ethnic text knowledge map building technique. The websites in different minority languages have developed rapidly, and the growth rate has shocked the world. Network public opinion is the main way for the nation to understand public opinion.

3. Approach Overview

3.1 The steps of the approach

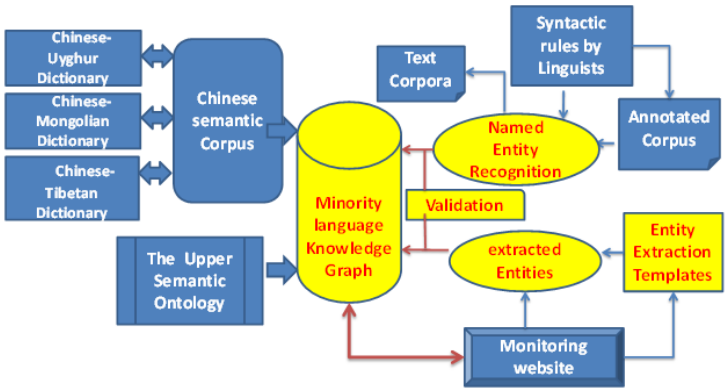


Figure 1. The Main Architecture of the Approach

Technical route is shown in Figure 1. The main steps are as follows:

(1) According to the existing dictionary, Chinese Knowledge Base and the Upper Semantic Ontology, the minority language knowledge graph should be preliminarily constructed with linking method.

(2) According to the selection of monitoring sites, it needs to analyze the site layout, to design the entity extraction template, to develop the visual template editing tools, and to provide the flexible custom design of template extraction.

(3) To analyze the named entity recognition rule with minority language linguists, to design the named entity recognition method, and to extend the knowledge graph with named entities extracted from the text corpus.

(4) To study the fusion and verification technology of new knowledge, and to guarantee the consistency and accuracy between the new knowledge and knowledge graph.

3.2 Main Challenges

In Figure 1, the blue part shows the foundation of existing research and yellow part shows what should be ongoing studies.

The semantic relevance technology research of Mongolian, Tibetan and Uyghur information resources need to consider the unique characters of the language. For example, The Tibetan language is the 'Subject-Object-Predicate' type language with rich syntax marks. In Tibetan language, the syntactic components usually have syntactic marks which can express the semantic meanings, such as agent, patient, time and tool.

As another example, Uyghur has obvious morphological rules, and this feature provides us with a lot of help when we need to do lexical analysis, part-of-speech tagging, and semantic understanding with Uyghur texts. But not all Uyghur language's

morphology obeys the rules. For instance, root forms of multi-category words (such as the noun گات (horse) and the verb گات (hit), گات (throw)) haven't change of form when they are used.

From the characteristics of the Mongolian, Tibetan and Uyghur language themselves, we need to study the semantic annotation technology.

Establishing a knowledge map entity extraction templates, and researching minority languages' entity linking methods should be studied.

We need to build entity extraction template for features of minority languages. For example, Uygur language is written from right to left, top to down, and its web page layout has uniqueness. Entity extraction templates require to be designed according to the different language writing habits.

Because of lack of aligned bilingual corpus, and there are great difference between different languages, especially among Mongolian, Tibetan and Uyghur. This problem will result in most words cannot find the corresponding vocabulary in other languages if we establish the mapping knowledge domain according to the traditional vocabulary with matching patterns. So the semantic similarity analysis method of fuzzy words, phrases and simple sentence which are based on semantic similarity are designed to extend the coverage of the multilingual knowledge mapping [9].

The entity's classification system of mapping knowledge domain is built to adapt to the need of the public opinion monitoring. Creating mentioned model and researching entity linking method which is based on the entity-mentioned model with comprehensive consideration of lexical semantic similarity, word co-occurrence, textual theme, and many other problems are also our research content.

How to fusion and validation technology of mapping knowledge domain need to be considered.

For the feature of multilingual mapping knowledge domain will continue to expand and improve, we establish a separate inverted index table for each language and extend search keywords between different languages. The advantage is that the establishment and maintenance of the inverted index maintenance independent of the extension of knowledge map and the inverted index table don't need to be changed frequently due to the increase of the document and the new semantic word.

We need to calculate the possibility of new knowledge which is compatible with existing knowledge and verify the new knowledge with comprehensive consideration of new knowledge verification technology with the feature of authority, redundancy and diversity, and so on. So we can guarantee the consistency and accuracy of the new knowledge and knowledge map, and guarantee the continuous updating of knowledge.

4. The algorithm to Build Knowledge Graph

Graph-based entity propagation has been widely used and shown to outperform other models [11]. Typically, graph-based algorithms are run in two main steps: graph construction and joint disambiguation when adding new knowledge.

4.1 Graph Construction

The graph construction provides a natural way to represent data in a variety of target domains, which is

Definition 1: knowledge graph relational model $G=(V,E,W)$:

$V=\{v_1, v_2, \ldots, v_m\}$ is defined as the set of vertices which covers all nodes which represent all vocabularies (or knowledge terms) , and the vocabularies come mainly from the dictionary.

$E=\{e_{12}, e_{ij}, \ldots, e_{mn}\}$ is a set of all edges connect all the vertices.

$W=\{w_{12}, w_{ij}, \ldots, w_{mn}\}$ is the weight of the edges. The value of w_{ij} is calculated by the semantic similarity and semantic relation of the word v_i and v_j.

Definition 2: Semantic notation

$S=\{s_1, s_2, ..., s_n\}$ is a set of all semantic source pairs labeling the semantic source of the node v_i, in which $s_i =< v_i, tag_i >$ is a semantic item and $tag_i =\{sa_1,sa_2,...,sa_t\}$is the key label set of the semantic item.

$O=\{o_1,o_2, ...,o_n\}$ is a set of ontologies, in which $o_i=\{item_1,item_2,...,item_s\}$.

$Sim(o_i, o_j)$ is denoted the semantic relation value of o_i and o_j, and the possibility of tag_i semantic related to a semantic set o_t is calculated by the similarity distance between the words. The similarity between two words is calculated based on a semantic dictionary HowNet and the method is introduced in our work [13].

One particular ontology set o_t we will repeatedly use when confront with v_i is the one that maximizes the similarity values.

Definition 3: conditional probability:

$\forall v_i \in V$ is the i[th] node in the knowledge graph G, and v_i is related to the semantic notation set o_i. The semantic conditional probability $p(v_i, o_j)$ denotes the semantic relation between v_i and o_j, which also means semantic conditional probability of the semantic units o_j relative to the given keywords v_i in the knowledge map G.

$$P(< v_i, o_j >) = (\sum_{\forall tag(< v_i, tag_i >\in s), \forall item_j \in o_j} (\ln \frac{\|tag_i\| - df(v_i, tag_i) + 0.5}{df(o_j, tag_j) + 0.5} * \frac{\|tag_i\| - df(v_i) + 0.5}{df(v_j) + 0.5} * \frac{1 + \ln(1 + \ln tf(item_j, o_i))}{(1-\gamma) + \gamma \frac{\|O\|}{avW1}} * \frac{1 + \ln(1 + \ln tf(item_j, o_j))}{(1-\delta) + \delta \frac{\|O\|}{av1}}) / \|v_i\|$$

Where $df(o_i, tag_i)$ is the number of tag_i whose items appear in o_i.

Definition 4: The weight of the edges:

The weight w_{ij} of two edges is calculated by the semantic coherence between nodes v_i and v_j:

$$Sim(v_i, v_j) = (\frac{\sum_{\forall o_m \in Va_i \cup Va_j} \left(\frac{2 * |P(< v_i, o_m >) - P(< v_j, o_m >)|}{|P(< v_i, o_m >) + P(< v_j, o_m >)|} \right)^2}{\|s_j \cup s_j\|})^{\frac{1}{2}}$$

The semantic coherence captures to what extent two semantic items occur in the same context. To be able to compare semantic items of different semantic ontologies, we need to extend this to knowledge mapping and joint disambiguation.

4.2 Knowledge Mapping and Joint Disambiguation

Given a knowledge sub-graph, a word distribution θ is a multinomial distribution over the words in the vocabulary V of the sub-graph. The probability of a new knowledge word is calculated and added into the sub-graph.

The result of knowledge mapping and disambiguation is to align sub-graphs, yielding the most coherent mappings. Some necessary definitions are listed as Figure 2:

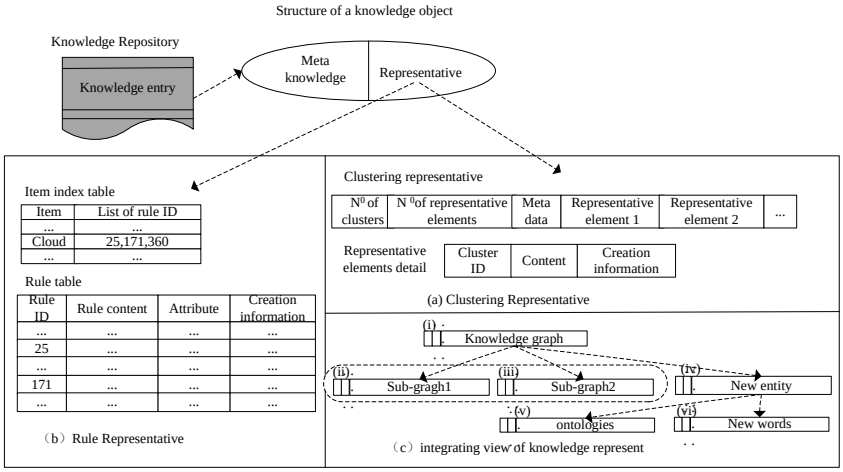


Figure 2. A Example of Knowledge Graph Linking

5. Conclusion and Future Work

The network information security in ethnic minority area will be enhanced and the national unity will be promoted by applying minority language processing technology, thus can improve ethnic researchers' capability for independent innovation and prompt the development of economics and culture.

By promoting the developing of minority language processing technology, the scheme that can strengthen the base for natural language processing technique on Tibetan and Uygur can be explored, thus can transplant the popular researched language processing technology to ethnic languages area.

Gathering, organizing and utilizing comprehensively these ethnic information resources can help with information searching and can provide technical support for the research, inheritance and protection of ethnic culture. Minority language processing has a problem, data sparsity issue. Chinese-Tibetan and Chinese-Uygur semantics relevance corpus can be built by utilizing current existing minority language resources and use Chinese as intermediary for researchers to meet the research requirement and extends the application of ethnic language processing technique.

In this paper, how to discover the evolution of knowledge map in multi-languages is studied. The approach is designed to capture the topology of semantic items and to link the sub-graph to a global knowledge map. Instead of building a knowledge map start from scratch, we conceptually define semantic classes as a quantized unit of evolutionary link in sub-graph and discover new knowledge with multi-language dictionaries. Discovered new knowledge items are then connected to form an evolution knowledge map using a measure derived from the underlying semantic classes. We integrate these noisy items and entities into a unified probabilistic knowledge map using ideas from graph-based algorithm.

Acknowledgements

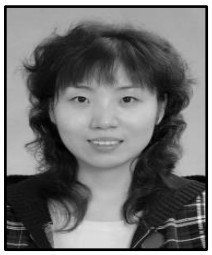
Our work is supported by the National nature science foundation of China (No. 61103161), the Program for New Century Excellent Talents in University (NCET-12-0579) and the "985" special funds in School of information engineering, Minzu university of China.

References

[1]. N. A. L. Khac, L. M. Aouad and M. -T. Kechadi, "Distributed Knowledge Map for Mining Data on Grid Platforms", IJCSNS International Journal of Computer Science and Network Security, vol. 7, no. 10, **(2007)**, October.
[2]. K. Suyeon, E. Suh and H. Hwang, "Building the Knowledge map: an industrial case study", Journal of

Knowledge Management, vol. 7, no. 2, **(2003)**, pp. 34 – 45.
[3]. K. Chan and J. Liebowitz, "The synergy of social network analysis and knowledge mapping: a case study", International Journal of Management and Decision Making, vol. 7, no. 1, **(2005)**, pp 19-35.
[4]. C. Lin, J. –M. Yeh and S. –M. Tseng, "Case study on knowledge-management gaps", Journal of Knowledge Management, vol. 9, no. 3, **(2005)**, pp 36-50.
[5]. G. Palla, I. Dernyi and I. Farkas, "Uncovering the Overlapping Community Structure of Complex Network in Nature and Society", Nature, vol.435, no. 7043, **(2005)**, pp. 814-818.
[6]. K. M. Fisher and M. Kibby, (Eds.), "Knowledge acquisition, organization and use in biology", Heidelberg, Germany, Springer Verlag, **(1996).**
[7]. A. Charu and H. Wang, "Managing and Mining Graph Data", Springer-Verlag New York Inc, **(2010)**.
[8]. B. A. Schwendimann, "Making Sense of Knowledge Integration Maps", Digital Knowledge Maps in Education, Chapter 2, **(2014)**, pp. 17-40.
[9]. S. Ebener, A. Khan, R. Shademani, L. Compernolle, M. Beltran, M. A. Lansang and M. Lippman, "Knowledge mapping as a technique to support knowledge translation", Bulletin of the World Health Organization, **(2006)** August.
[10]. X. Zeng, D. F. Wong, L. S. Chao and I. Trancoso, "Graph-based Semi-supervised Model for Chinese Word Segmentation and Part-of-Speech Tagging", In proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, **(2013)**, pp. 770-779.
[11]. D. Das and N. A. Smith, "Graph-based lexicon expansion with sparsity-inducing penalties", In proceedings of the North American Chapter of the ACL, **(2012)**, pp. 677-687.
[12]. A. Alexandrescu and K. Kirchhoff, "Graph-based learning for statistical machine translation", In proceedings of the North American Chapter of the ACL, **(2009)**, pp. 119-127.
[13]. X. Jiang and L. Qiu, "A Tibetan ontology concept acquisition method based on HowNet and Chinese Tibetan dictionary", In proceedings of International Conference on Asian language, **(2013)**, pp. 189-192.
[14]. **(2014),** http://www.wolframalpha.com/.

Authors



Lirong Qiu she received her Ph.D. degree in Computer Science from Chinese Academy of Science (2007). Now she is an associate professor of Information Engineering Department, Minzu University of China. Her current research interests include natural language processing, artificial intelligence and distributed systems.