DCPAPERS
Proc. Int'l Conf. on Dublin Core and Metadata Applications 2014

Interlinking Cross Language Metadata Using Heterogeneous Graphs and Wikipedia

Xiaozhong Liu
School of Informatics and Computing
Indiana University, USA
liu237@indiana.edu

Miao Chen
School of Informatics and Computing
Indiana University, USA
miaochen@indiana.edu

Jian Qin
School of Information Studies
Syracuse University, USA
jqin@syr.edu

Abstract

Cross-language metadata are essential in helping users overcome language barriers in information discovery and recommendation. The construction of cross-language vocabulary, however, is usually costly and intellectually laborious. This paper addresses these problems by proposing a Cross-Language Metadata Network (CLMN) approach, which uses Wikipedia as the intermediary for cross-language metadata linking. We conducted a proof-of-concept experiment with key metadata in two digital libraries and in two different languages without using machine translation. The experiment result is encouraging and suggests that the CLMN approach has the potential not only to interlink metadata in different languages with reasonable rate of precision and quality but also to construct cross-language metadata vocabulary. Limitations and further research are also discussed.

Keywords: metadata; linked data; cross language; heterogeneous graph

1. Research Problem

Subject categories and keywords in metadata descriptions are primary subject access points for information discovery whether for English- or non-English-speaking users. While many non-English speaking users can read and understand English, it is often not the same for the opposite. To bridge the gap between languages, digital libraries such as Europeana (http://europeana.eu) offer cross-language metadata so that users can search by any language. The cross-language search function is valuable and enables information discovery in languages that users would have otherwise unable to reach due to the language barrier.

Cross-language subject tools for Asian languages, however, have been lagging behind the increase in Asian Internet users and research output. Although the Internet has created a global village, the lack of cross-language metadata prevents information from flowing bi-directionally between English and Asian languages and creates language silos of information. Take CiNii (http://ci.nii.ac.jp/) as an example: even though both Japanese and English resources are indexed in the CiNii database, cross language retrieval and recommendation is unavailable. The same problem exists in Google Scholar, a giant scholarly retrieval engine. In addition, current tools are often limited to standardized human or machine translation, which is not suitable for high quality information retrieval and recommendation. One contributing factor for the lack of cross-language information discovery and recommendation is the difficulty in constructing a multi-language metadata vocabulary.

It is well known that the construction of any vocabulary tool is time consuming and intellectually laborious. The Chinese language version (AAT-Taiwan) of the Art and Architecture Thesaurus ("AAT", 2014) for example, is translated and mapped with its English version. It contains 34,961 concepts, 26,813 translated concepts, 12,668 archived records, and 6,564 edited records and took multiple years and professionals and domain experts to complete. The maintenance and updating has been ongoing since its first release in 2009. Building cross-language counterparts is a huge endeavor and costly in both time and personnel.

https://doi.org/10.23106/dcmi.952136493

The usefulness/lack of cross-language subject vocabularies calls for new approaches to developing such vocabularies at a large scale while maintaining a reasonable level of quality and low cost. To address this conundrum, we propose a cross-language metadata network approach that will generate cross-language vocabularies on the fly by leveraging existing vocabulary resources. This paper reports a preliminary experiment as a proof of concept that uses metadata from four elements – publication, author, keyword, and venue – to construct cross-language metadata network graphs, which will then be linked through the language counterparts in Wikipedia concepts and subject categories. This approach will allow for searches in a user's native language to return results in multiple languages without machine translation.

2. Relevant Research

Developing cross-language metadata network graphs is motivated not only by the need for such tools but also by the issues in cross-language information retrieval that previous research has ignored or unable to address (Oard & Diekema 1998; Nie 2010; Ye, Huang, He & Lin 2012). Cross-language retrieval algorithms and methods are well documented in research publications. Most of these algorithms and methods, however, focused on translation rather than linking. They employed statistical models, i.e., latent semantic indexing (Littman, Dumais & Landauer 1998), parallel corpuses mining (Nie et al., 1999), and n-gram (AbdulJaleel & Larkey 2003) to construct bilingual translation models. As such, the translations rely on the source text and are limited to matching terms for translating the query from its original language to the target language in order to perform searches, rather than for linking relevant concepts cross languages. The translations have nothing to do with the metadata describing the source, much less creating both content and language linkages between metadata descriptions.

Machine translation plays an important role in constructing cross-language vocabularies (Dumais et al., 1997; Vossen 1998). Research literature in this field exhibits two paradigms of translating approaches: dictionary/rule based and parallel/comparable corpus based (Potthast, Stein, and Anderka, 2008). The first approach relies heavily on corpora and dictionaries while the second one uses the human-built cross-language links in knowledge bases such as Wikipedia. Cross-language links in Wikipedia explicitly connect concepts in different languages together and have proved to be useful sources for text mining across languages by navigating between the links. Studies show that same language pairs have a high ratio of cross-lingual links in Wikipedia. For example, the ratio of English-German links is as high as 95% (Sorg & Cimiano, 2008).

The method used by Sorg and Cimiano (2008) and Potthast et al. (2008) is called CL-ESA (Cross-Language Explicit Semantic Analysis). By projecting documents/queries to a vector space of concepts via Explicit Semantic Analysis (ESA) in one language, the vector space of concepts is mapped to a vector space of another language via cross-language links in Wikipedia. Potthast et al. (2008) used cross-language links in Wikipedia for cross-language information retrieval and showed a reasonably good performance in cross-language ranking and bi-lingual correlation ranking. Ye, Huang, He and Lin (2012) also employed Wikipedia as a graph-based bi-lingual resource for constructing a cross-language association dictionary (CLAD). They also found CLAD can be useful to enhance the cross-language information retrieval performance.

The studies mentioned above provide encouraging results for using Wikipedia as the bridge in developing cross-language metadata vocabularies. Although unforeseen factors may affect the precisions and coverage of concepts cross languages, it is nonetheless a worthwhile attempt in experimenting with the cross-language metadata linking approach using Wikipedia.

3. A Case Scenario in Cross-Language Vocabulary Linking

To demonstrate how cross-language vocabulary might be interlinked, we present a case scenario of metadata for scholarly publications. The DBLP Computer Science Bibliography (http://dblp.uni-trier.de/db/) contains metadata descriptions primarily for computer science publications written in English. The C-DBLP ("Chinese DBLP", n.d.) serves same goal for

https://doi.org/10.23106/dcmi.952136493

computer science publications written in Chinese. The metadata schemas for both DBLP and C-DBLP are comparable but do not communicate to one another, nor can users conduct searches across both databases. While different ownerships for each of these two databases is a primary factor for their inability to communicate to one another, it is also true that the metadata in two databases represent two completely different sets of publications and are in two different languages. Similarly, large search engine players such as Google Scholar and OCLC WorldCat index resources in multiple languages, but the metadata descriptions (e.g., keywords in different languages) in these systems are not related within their own system.

Figure 1. Wikipedia concepts and language links

Over the past decade, Wikipedia has become an increasingly important resource for the world knowledge. It provides two unique features that can potentially solve the aforementioned problems for cross-language information discovery. The first feature is that Wikipedia provides concept definitions in multiple languages. An example is the concept definition for "Semantic Web": this entry has been written in 39 languages (see Figure 1). In each language, the concept name is defined by the title of the article (entry). The Chinese counterpart for this concept is defined by the title "语义网", a term used in most publications for this topic in Chinese.

The other important feature is that all concepts in Wikipedia are inter-connected via Wikipedia hierarchical categories and hyperlinked among Wikipedia pages. For instance, the concepts "Semantic Web" and "metadata" are connected via the path

[Wikipedia Concept: Semantic Web] →
[Wikipedia Category: Knowledge Engineering] →
[Wikipedia Category: Knowledge Representation] ←
[Wikipedia Concept: Metadata]

In other words, all concepts in Wikipedia are inter-connected through topic links (Wikipedia categories) and cross-language equivalents.

https://doi.org/10.23106/dcmi.952136493

For the purpose of generating cross-language metadata vocabularies, the interconnectedness across multiple languages between concepts and knowledge categories in Wikipedia makes it an ideal source to leverage. If Wikipedia can be used as the intermediary vocabulary, we may be able to design algorithms to "ask" it to translate metadata between different languages. This means that digital libraries and repositories in different languages may use the intermediary tool to construct cross-language metadata vocabularies for information discovery and recommendation. It will be possible then for cross-language vocabulary tools to automatically select and recommend most relevant cross-language publications *without* having to rely on machine translation. In the cases of DBLP and C-DBLP, it is possible to use Wikipedia as the intermediary nodes to interlink publications, venues, and authors in these two digital libraries, no matter which language is used to search, via the *[Keyword] →[Wikipedia Concept]* link. As each Wikipedia concept is written in both Chinese and English, this step does not need to involve machine translation.

We are aware of the limitation of Wikipedia resource, and the sparseness of Wikipedia definitions in certain languages may limit the generalizability of the proposed method. For instance, if there is only a small amount of Wikipedia concepts defined in a language, the keyword projection performance can be understandably low.

4. Methodology

Using Wikipedia to create Cross-Language Metadata Networks (CLMN) involves two steps. In the first step a Single-Language Metadata Network (SLMN) is built for a monolingual digital library or repository. In the second step, the SLMN will be mapped to Wikipedia concepts and subject categories to create Cross-Language Metadata Networks (CLMN). Through this two-step method, cross-language metadata vocabularies are constructed and then used to connect metadata and resource objects in digital libraries/repositories across different languages. In the section below we will first discuss the method for generating metadata networks for an individual repository and then describe the CLMN through which SLMNs are interconnected via Wikipedia's bridge nodes, i.e., Wikipedia pages and subject categories.

4.1 Step 1: Creating Single-Language Metadata Networks (SLMN)

We assume that there are four types of resource objects – publications (papers, reports, webpages, and books), venues (journals, conference proceedings, and domain names as embodied by websites), subjects, and authors – in a single-language digital library. Between the four types of resource objects, there exist various types of linkages: citation linkages between publications, authorship linkages between authors and publications, and venue linkages between publications and venues. We also assume that in a single-language digital library (or repository), a list of subject terms and values (keywords or controlled vocabulary) is available to represent publications and venues and that metadata and publications share the same language.

Using the network theory, each resource object is a network node (vertex) and the links between nodes (vertices) are edges. Metadata in a single-language digital library are considered as a single-language metadata network (SLMN) in which the nodes are connected by edges. This network is heterogeneous by nature in the sense of network node types, because the same network contains multi-types of nodes: author (A), publication (P), venue (V), and keyword (K), which are what this study focuses.

For each digital library (a commercial database or an institutional repository), there exists a local SLMN. All four types of nodes mentioned above can be connected by any of the 7 types of edges: 1) $P \rightarrow A$, a paper is written by an author; 2) $P \rightarrow V$, a paper is published in a venue; 3) $P \rightarrow K$, a paper or publication is relevant to a keyword; 4) $P \rightarrow P$, a publication cites or links to publications; 5) $K \rightarrow P$, a keyword (topic) is assigned to publications; 6) $K \rightarrow A$, a keyword (topic) is assigned by authors; and 7) $K \rightarrow V$, a keyword (topic) is assigned to venues. Edge types 1, 2, 3 and 4 are implemented by using metadata in a single-language digital library. Keywords

https://doi.org/10.23106/dcmi.952136493

derived from publications, author names, and venues are labeled as topic and represented by edge types 5, 6, and 7, which are calculated by using PageRank with Prior algorithms (Liu, Zhang, and Guo, 2013) on the homogeneous citation graphs (publication-citation graph, author-citation graph, and venue-citation graph). Note that, as this network can be potentially used for resource recommendation, all edges are associated with an edge weight, $P(v|u)$, which indicates the transitioning probability (weight) from node u to node v.

4.2 Step 2: Creating Cross-Language Metadata Networks (CLMN)

The goal of this step is to generate cross-language metadata networks using computational methods. The CLMNs generated from using Wikipedia and the PageRank Prior algorithms will function as a linking mechanism to interconnect metadata silos of single language into a global network with the capability of performing cross-language information discovery and recommendation. In the CLMN approach, a collection of digital libraries or repositories are represented by k Metadata Networks (MNs). Figure 2 visualizes the CLMN creation progress. There are four layers in a CLMN and k SLMNs connect to the Wikipedia concept and Wikipedia category nodes on the CLMN, in which Wikipedia nodes function as the bridge to interconnect different SLMNs. Meanwhile, all Wikipedia nodes (Wikipedia concepts and Wikipedia categories) also connect with the incoming/outgoing links (between Wikipedia concepts), concept-category relations, and the hierarchical relations between categories.

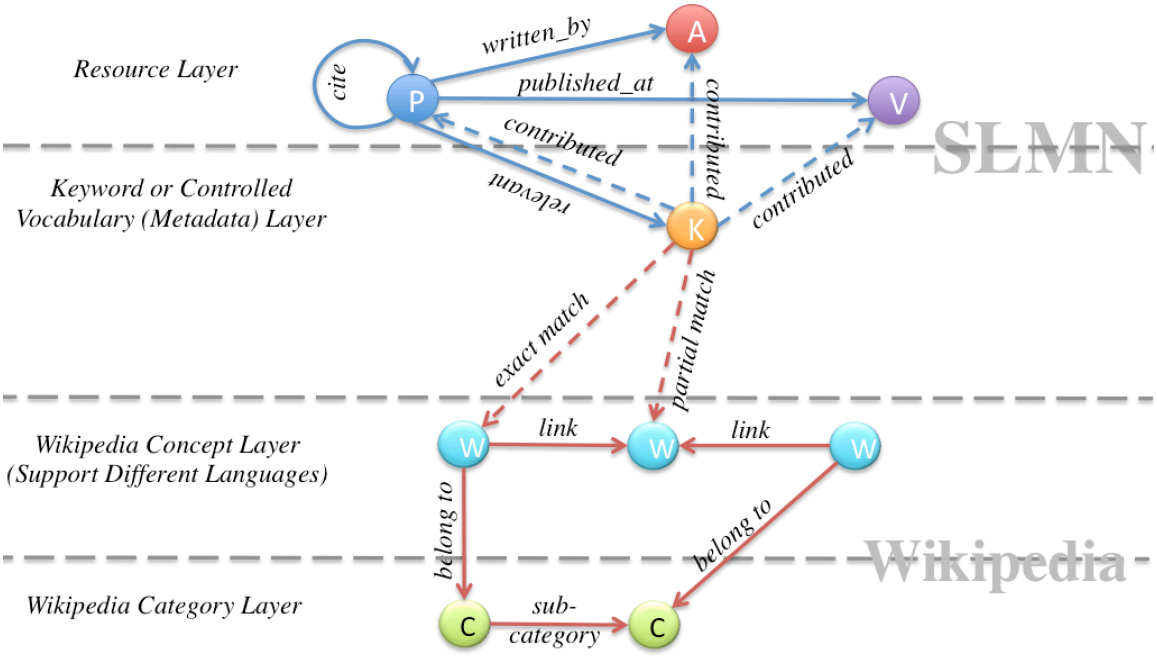

Figure 2. Cross-Language Metadata Networks (CLMN)

In Figure 2, dotted lines indicate the calculated or inferred relationships and the solid lines indicate the relationships physically exist in the repository or Wikipedia database. It depicts how one SLMN typically connects to Wikipedia nodes, which is also how other SLMNs will connect to the Wikipedia nodes. The middle section is where automatic pairing and linking of the concepts in different languages takes place. All keywords or controlled vocabularies (node *K)* connect to Wikipedia concepts via two kinds of edges: exact match edge and partial match edge. The former edge type indicates that the string represented by node K is exactly the same as Wikipedia concept title. Note that K on different SLMNs may be in different languages, while Wikipedia concept is also indexed by multiple languages. The latter edge type is generated by using information retrieval algorithms, e.g., language model or vector space model, which means that the target keyword or controlled vocabulary is part of the content of the Wikipedia concept's

https://doi.org/10.23106/dcmi.952136493

content. Similarly, the content of Wikipedia concept may also be in different languages. Similar as the edge types in SLMN, all edges between Wikipedia nodes and keywords nodes are associated with the edge weight.

5. Preliminary Experiment

As a proof of concept for the proposed method, we construct a CLMN by using the ACM Digital Library (English computer science publications + metadata, http://www.acm.org/) and WanFang Digital Library (Chinese computer science publications + metadata, http://www.wanfangdata.com.cn/). All four types of nodes in publications' metadata across both libraries – authors, venues, papers, and keywords – were connected by using the intermediary layer Wikipedia as shown in Figure 2. For this experiment, we used Wikipedia Chinese and English 2014 April dumps.

Due to the space limit, we present only the metadata layer and Wikipedia layer in this section. The CLMN constructed in this preliminary experiment contains 1,481 English keywords and 121 Chinese keywords (English keywords 10 times more than Chinese keywords because of the data limitation). Connected to these keywords were 1,719 Wikipedia page nodes and 1,146 Wikipedia category nodes.

Two exemplar Chinese keywords, "机器学习" (Machine Learning) and "信息抽取" (Information Extraction) , were used as query terms to find the related English keywords by using two types of paths: 1. *[Chinese Keyword] → [Wikipedia Concept] ←[English Keyword],* and 2. *[Chinese Keyword] → [Wikipedia Concept]→ [Wikipedia Category] ← [Wikipedia Concept] ←[English Keyword]* (Edge direction was ignored). The first path used only one intermediary Wikipedia node between the query and target keywords in Chinese and English. The second one was more complicated because the Chinese query keyword and English target keyword may link to different Wikipedia concepts and these Wikipedia concepts may share the same Wikipedia category.

Given the space limitation, we investigated only the first example in more details. Figure 3 displays the paths through which the results for "机器学习" were generated. Different types of nodes are represented by different colors on the CLMN graph. This graph example shows that Wikipedia page and category nodes function as intermediary nodes to link together the same concept Machine learning in English and Chinese.

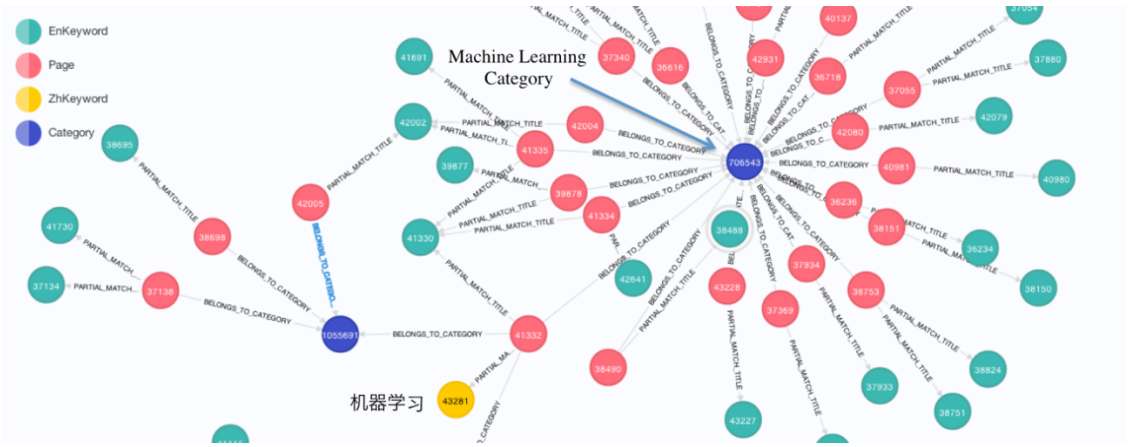

Figure 3. Related English Keywords for "机器学习" on the CLMN (via Wikipedia nodes)

https://doi.org/10.23106/dcmi.952136493

The specific paths for query "机器学习" on the CLMN are listed below (CK = Chinese keyword, WP = Wikipedia page, WC = Wikipedia category, and EK = English Keyword):

Result for path [Chinese Keyword] → [Wikipedia Concept] ←[English Keyword] (1 result)

1. CK:机器学习→WP:machine_learning←EK:machine_learning

Results for path [Chinese Keyword] → [Wikipedia Concept]→ [Wikipedia Category] ← [Wikipedia Concept] ←[English Keyword] (26 results)

1. CK:机器学习→ WP:machine_learning→WC:Machine_learning←WP:cluster_analysis←EK:cluster_analysis
2. CK:机器学习→ WP:machine_learning→WC:Machine_learning←WP:expectation_maximization_algorithm ←EK:em_algorithm
3. CK:机器学习→ WP:machine_learning→WC:Cybernetics←WP:complex_systems←EK:complex_systems
4. CK:机器学习→ WP:machine_learning→WC:Machine_learning←WP:reinforcement_learning←EK:reinforc ement_learning
5. CK:机器学习→ WP:machine_learning→WC:Machine_learning←WP:pattern_recognition←EK:pattern_rec ognition
6. CK:机器学习→ WP:machine_learning→WC:Machine_learning←WP:formal_concept_analysis←EK:concep t_analysis
7. CK:机器学习→ WP:machine_learning→WC:Machine_learning←WP:unsupervised_learning←EK:unsuper vised_learning
8. CK:机器学习→ WP:machine_learning→WC:Machine_learning←WP:hidden_markov_model←EK:hidden_ markov_model
9. CK:机器学习→ WP:machine_learning→WC:Machine_learning←WP:expectation_maximization_algorithm ←EK:expectation_maximization
10. CK:机器学习→ WP:machine_learning→WC:Machine_learning←WP:supervised_learning←EK:supervise d_learning
11. CK:机器学习→ WP:machine_learning→WC:Machine_learning←WP:pattern_recognition←EK:pattern_de tection
12. CK:机器学习→ WP:machine_learning→WC:Machine_learning←WP:artificial_neural_network←EK:neura l_networks
13. CK:机器学习→W P:machine_learning→WC:Machine_learning←WP:artificial_neural_network←EK:artificia l_neural_network
14. CK:机器学习→ WP:machine_learning→WC:Cybernetics←WP:genetic_algorithm←EK:genetic_algorithm
15. CK:机器学习→ WP:machine_learning→WC:Machine_learning←WP:nearest_neighbor_search←EK:neare st_neighbor_search

https://doi.org/10.23106/dcmi.952136493

16. CK:机器学习→
    WP:machine_learning→WC:Machine_learning←WP:principal_component_analysis←EK:principal_component_analysis
17. CK:机器学习→
    WP:machine_learning→WC:Cybernetics←WP:artificial_intelligence←EK:artificial_intelligence
18. CK:机器学习→WP:machine_learning→WC:Cybernetics←WP:system←EK:systems
19. CK:机器学习→WP:machine_learning→WC:Cybernetics←WP:autonomy←EK:autonomy
20. CK:机器学习
    →WP:machine_learning→WC:Cybernetics←WP:control_theory←EK:control_theory
21. CK:机器学习→
    WP:machine_learning→WC:Machine_learning←WP:support_vector_machine←EK:support_vector_machine
22. CK:机器学习→
    WP:machine_learning→WC:Cybernetics←WP:information_theory←EK:information_theory
23. CK:机器学习→
    WP:machine_learning→WC:Machine_learning←WP:discriminative_model←EK:discriminative_model
24. CK:机器学习
    →WP:machine_learning→WC:Machine_learning←WP:perceptron←EK:perceptron
25. CK:机器学习→
    WP:machine_learning→WC:Machine_learning←WP:formal_concept_analysis←EK:formal_concept_analysis
26. CK:机器学习→
    WP:machine_learning→WC:Machine_learning←WP:conditional_random_field←EK:conditional_random_field

Specific paths for query "信息抽取" are listed below:

Results for path [Chinese Keyword] → [Wikipedia Concept] ←[English Keyword] (1 result):

1. CK:信息抽取→WP:information_extraction←EK:information_extraction

Results for path [Chinese Keyword] → [Wikipedia Concept]→ [Wikipedia Category] ← [Wikipedia Concept] ←[English Keyword] (13 results):

1. CK:信息抽取→
   WP:information_extraction→WC:Artificial_intelligence←WP:artificial_intelligence←EK:artificial_intelligence
2. CK:信息抽取→
   WP:information_extraction→WC:Artificial_intelligence←WP:computer_vision←EK:computer_vision
3. CK:信息抽取→
   WP:information_extraction→WC:Artificial_intelligence←WP:description_logic←EK:description_logics
4. CK:信息抽取→
   WP:information_extraction→WC:Artificial_intelligence←WP:fuzzy_logic←EK:fuzzy_logic

https://doi.org/10.23106/dcmi.952136493

5. CK:信息抽取→ WP:information_extraction→WC:Artificial_intelligence←WP:game_theory←EK:game_theory
6. CK:信息抽取→ WP:information_extraction→WC:Artificial_intelligence←WP:intelligent_agent←EK:intelligent_agent
7. CK:信息抽取→ WP:information_extraction→WC:Artificial_intelligence←WP:markov_random_field←EK:markov_random_field
8. CK:信息抽取→ WP:information_extraction→WC:Natural_language_processing←WP:cross-language_information_retrieval←EK:cross_language_information_retrieval
9. CK:信息抽取→ WP:information_extraction→WC:Natural_language_processing←WP:information_retrieval←EK:information_retrieval
10. CK:信息抽取→ WP:information_extraction→WC:Natural_language_processing←WP:latent_semantic_analysis←EK:latent_semantic_analysis
11. CK:信息抽取→ WP:information_extraction→WC:Natural_language_processing←WP:natural_language_processing←EK:natural_language_processing
12. CK:信息抽取→ WP:information_extraction→WC:Natural_language_processing←WP:natural_language←EK:natural_language
13. CK:信息抽取→ WP:information_extraction→WC:Natural_language_processing←WP:question_answering←EK:question_answering

The specific results shown above demonstrate that the path *[Chinese Keyword]* → *[Wikipedia Concept]* ←*[English Keyword]* can find accurate translation, while the path *[Chinese Keyword]* → *[Wikipedia Concept]*→ *[Wikipedia Category]* ← *[Wikipedia Concept]* ←*[English Keyword]* can locate a number of high quality related (linked) keywords in a different language. The experiment results suggest that CLMN is promising as a means to link metadata across languages and digital libraries. The metadata used in this experiment are relatively specialized with reasonable level of quality, hence whether the method can be applied to other domains and accomplish a comparable level of performance will need further study and evaluation.

6. Discussion and Conclusion

The resulting CLMNs have a number of potentials for metadata representation and resource discovery. The four sets of results presented in the last section are structured data with path and node information attached. They can be parsed into the format suitable for building cross-language vocabularies using computer programs. Such cross-language vocabularies can be then encoded in the Linked Data formats and shared through vocabulary services. Another application is to recommend resources (i.e., publication, author or venue) across repositories and languages. For example, given an author ID (on a SLMN), the system can recommend publications potentially relevant to users' interest in a different language. Given a keyword (on a SLMN), we can recommend top related venues (venue recommendation) or expert (author recommendation) in a different language.

Unlike classical machine translation methods that use homogeneous data sources, this study employed heterogeneous graph mining and text mining methods to connect all the metadata via

https://doi.org/10.23106/dcmi.952136493

Cross-Language Metadata Networks (CLMN), in which Wikipedia is used as the intermediary nodes to link local repositories. We took metadata from ACM and WanFang digital libraries to run our experiment. The results suggest that CLMN as a novel approach was able to find not only accurate translations but also locate related metadata in different languages. This is especially encouraging for developing a low cost and effective method for automatic cross-language vocabulary construction.

The reliability and validity of CLMN method need further study and experiment to verify. We plan to conduct further experiment with other sources of metadata, e.g., those available in open repositories where metadata are crowd-sourced and in disciplines other than computer science. As our next step research, we are keen on developing a bilingual vocabulary linked data set using this method in a humanities domain by leveraging data from public digital libraries.

References

AAT (ART & Architectural Thesaurus). Retrieved Aug 1, 2014 from http://www.getty.edu/research/tools/vocabularies/lod/

AbdulJaleel, Nasreen, and Leah S. Larkey. (2003). Statistical transliteration for English-Arabic cross language information retrieval. In Proceedings of the twelfth international conference on Information and knowledge management (pp. 139-146). ACM.

Chinese DBLP. Retrieved Aug 1, 2014 from http://cdblp.cn/index.php

Dumais, Susan T., Todd A. Letsche, Michael L. Littman, and Thomas K. Landauer. (1997) Automatic cross-language retrieval using latent semantic indexing. In AAAI spring symposium on cross-language text and speech retrieval (Vol. 15, p. 21).

Littman, Michael L., Susan T. Dumais, and Thomas K. Landauer.(1998). Automatic cross-language information retrieval using latent semantic indexing. In Cross-language information retrieval (pp. 51-62). Springer US.

Nie, Jian-Yun, Michel Simard, Pierre Isabelle, and Richard Durand. (1999, August). Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval(pp. 74-81). ACM.

Nguyen, D., A. Overwijk, C. Hauff, D.R. Trieschnigg, D. Hiemstra, and F. De Jong, (2009). WikiTranslate: query translation for cross-lingual information retrieval using only Wikipedia. In Evaluating Systems for Multilingual and Multimodal Information Access (pp. 58-65). Springer Berlin Heidelberg.

Nie, Jian-Yun. (2010). Cross-language information retrieval. Synthesis Lectures on Human Language Technologies, 3(1), 1-125.

Potthast, Martin, Benno Stein, and Maik Anderka. (2008). A Wikipedia-based multilingual retrieval model. In Advances in Information Retrieval (pp. 522-530). Springer Berlin Heidelberg.

Sorg, Philipp, and Philipp Cimiano. (2008a). Cross-lingual information retrieval with explicit semantic analysis. In Working Notes for the CLEF 2008 Workshop.

Sorg, Philipp, and Philipp Cimiano. (2008b). Enriching the crosslingual link structure of Wikipedia-a classification-based approach. In Proceedings of the AAAI 2008 Workshop on Wikipedia and Artificial Intelligence (pp. 49-54).

Vossen, Piek. (1998). A multilingual database with lexical semantic networks. Kluwer Academic Publishers, Dordrecht.

Ye, Zheng., Huang, Jimmy X., He, Ben, and Hongfei Lin (2012). Mining a multilingual association dictionary from Wikipedia for cross-language information retrieval. Journal of the American Society for Information Science and Technology, 63(12), 2474-2487.

https://doi.org/10.23106/dcmi.952136493