An Approach to Cross-Age and Cross-Cultural Information Access for Digital Humanities

Akira Maeda and Fuminori Kimura

College of Information Science and Engineering, Ritsumeikan University
1-1-1 Noji-higashi, Kusatsu, Shiga 525-8577, Japan
amaeda@is.ritsumei.ac.jp and fkimura@is.ritsumei.ac.jp

1. Introduction

Since libraries have collection of documents across age and culture, and even language, the libraries are inherently multi-age, multi-cultural, and multi-lingual. In the digital age, more and more historical documents are being digitized to preserve contents written in deteriorating papers.

Given this background, libraries, governments, and major internet providers such as Google, Yahoo, MSN, are forming consortiums to massive preservation of historical documents stored in libraries. (e.g. Google Book Search, Open Content Alliance, MSN Search Books, World Digital Library, etc.). It means that more and more old text contents will be accessible on the internet in the near future. Obviously, huge amount of knowledge in old documents is as important as recent born-digital documents typically available on the web, because old documents are the collection of wisdom from B.C. Thus, it might be very useful to be able to access such old documents.

However, it is not always easy to retrieve old documents, mainly due to the substantial change in language and culture over time. Therefore, we need a method to access old documents written in ancient language using modern language. We call this method "Cross-Age Information Retrieval". Moreover, we should consider the cultural difference over time, even for the same language. For this, we need a method of "Cross-Cultural Information Retrieval".

Most of the researches on information retrieval and information access focus on documents written in modern language, but we believe that knowledge and wisdom written in old documents provide rich and valuable information which are not available in modern language documents, especially in web contents.

We propose a "Cross-Age and Cross-Cultural Information Retrieval" methods in order to tackle these problems. It will discover hidden knowledge and wisdom written in old documents.

2. Related Work

Many researches on Cross-Language Information Retrieval have been conducted in the last 10 years, with the background of the rapid growth of the Web around the world since the middle of 1990's. Various approaches, including query translation, document translation, and the use of intermediate language have been studied, and for certain language pairs (e.g. between European languages), adequate retrieval effectiveness has been achieved [1].

On the contrary, there are very few researches on information retrieval method for historical documents, and most of which is based on simple keyword matching. Recently, some approach has been proposed to access historical documents, it could be regarded as a kind of Cross-Age Information Retrieval [2][3]. Our goal is to establish a more effective and sophisticated retrieval method that considers not only language difference over time, but also cultural difference between languages and ages.

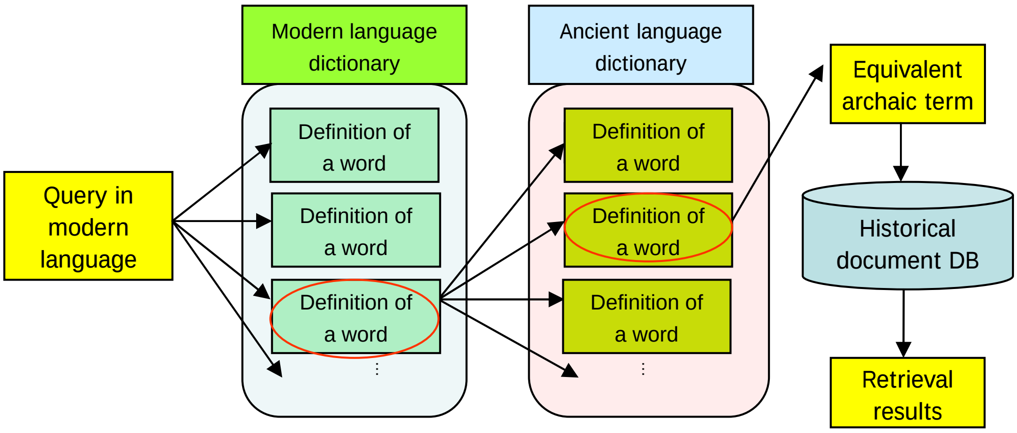

Figure 1. Overview of the proposed method for Cross-Age Information Retrieval.

3. The Proposed Method

Of course, it is not very easy to realize such retrieval methods. We adopt dictionary-based query translation approach, since it is proven to be the most effective method for Cross-Language Information Retrieval. In order for dictionary-based methods to be effective, we need to use precise and comprehensive dictionaries for both modern and ancient languages. From these two dictionaries, we try to discover relationships between entries in those dictionaries, and to "translate" the query terms in modern language into equivalent terms in ancient language. For this translation process, we propose the following method:

1. For each entry in the modern dictionary, we look for an equivalent entry in the archaic word dictionary by calculating the similarities between the definition of the modern word and all the definitions of the archaic words. For this process, we can use standard text similarity measure based on vector space model and tf-idf term weighting scheme.

2. Then, we take the most similar definition in the archaic word dictionary, and this entry (headword) is regarded as the equivalent of the modern word.

3. If more than one equivalent entry exists, we disambiguate the translation candidates using the term association measure such as mutual information to find the most equivalent archaic term for the modern language term.

The overview of the proposed method is illustrated in Figure 1.

4. Document Collections

Until recently, it was not easy to obtain historical documents in a text format. However, some digital libraries (e.g. Google Book Search, Open Content Alliance, etc.) are ready to provide their collection of historical documents in text format for research purposes. Moreover, there are numerous existing old documents available online. In Japan, there is a volunteer-based effort called "Aozora Bunko" to digitize and to make accessible over 6,000 copyright-expired classic literatures online. Also, many universities and institutions have already been providing collections of old documents in text format. We can use these huge collections of old documents for our proposed method.

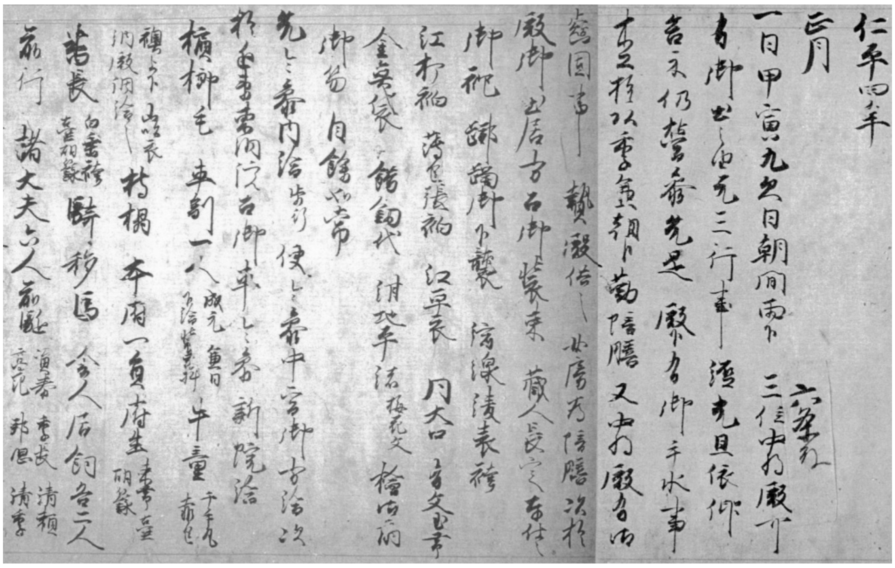

Figure 2. Example of the original copy of a historical Japanese document “Hyohanki”.

For now, we are focusing on a Japanese historical document called “Hyohanki”, which was written in late Heian era (12th century) in Japan. It is a valuable resource for the research of Japanese culture of that time period. An example of its original copy is shown in Figure 2. Although some part of it has been deteriorated and missing, all of the existing pages are digitized into the text format.

5. Language Resources

As described in Section 3, we need dictionaries in order to translate the modern language query into archaic terms. In the case of “Hyohanki”, we can use some existing electronic dictionaries available on CD-ROM. For Japanese modern language, we use “Kojien”, one of the most famous and comprehensive Japanese language dictionaries. For ancient language, we use “Shogakukan Kokugo Daijiten”, which covers both ancient and modern Japanese.

6. Preliminary Experiments

6.1. Experiment on translation accuracy

We conducted a preliminary experiment to test the accuracy of “translation” from modern language to ancient one. In the experiment of obtaining equivalent archaic term from modern language term in Japanese using the document collection and dictionaries described in Sections 3 and 4, we found that our proposed method archives about 72% of “translation” accuracy. Obviously, we have to improve our method substantially for this method to be really useful. One of the ideas to improve our method is to ignore some terms and symbols that are specific to a dictionary in the similarity calculation of dictionary definitions.

6.2. User survey

We conducted a user survey in order to assess the usefulness of the proposed method. In this experiment, we divide the users into two classes. One is “novice users”, which consists of 8 university students who are not familiar with ancient documents. The other is “expert users”, which consists of 8 university students and professors whose specialty is Japanese literature and especially in the age of the document used in this experiment.

The results of the user survey are shown in Table 1. In the ratings, “1” means the worst and “5” means the best. From the table, we can observe that our proposed method is more favored by novice users than expert users. It suggests that our method has a possibility to bridge the gap between ancient documents and novice users. On the other hand, it suggests that the translation accuracy of our method is obviously still far from the knowledge of the expert users. Nevertheless, the ratings for "practicability of cross-age IR" suggests that our proposed method might become a promising method for improving access to ancient documents both for novice and expert users.

Table 1. Results of the user survey.

	Ratings of novice users					**Ratings of expert users**					**Avg. of ratings**	
	1	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**	**novice**	**expert**
Usability of cross-age IR		1	2	3	2		2	3	3		3.8	3.1
Accuracy of "translation"			2	4	2		2	2	2	1	4.0	3.3
Practicability of cross-age IR		1	1	3	3		1	4	1	2	4.8	3.5

7. Conclusion

In this paper, we proposed a novel information retrieval technique called "Cross-Age Information Retrieval", which can be used to access old documents written in ancient language using a query in modern language. Although our proposed technique is still in an early stage, we believe that we can achieve adequate retrieval effectiveness by incorporating techniques used for Cross-Language Information Retrieval.

Our future work include conducting experiments of retrieval effectiveness, consideration of cultural difference over time, and thus extending our technique to realize Cross-Age, Cross-Cultural, and Cross-Language Information Access.

References

[1] Jianqiang Wang and Douglas W. Oard. Combining bidirectional translation and synonymy for cross-language information retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR2006)*, pp. 202-209, 2006.

[2] Andrea Ernst-Gerlach and Norbert Fuhr. Retrieval in text collections with historic spelling using linguistic and spelling variants. In *Proceedings of the 7th ACM/IEEE Joint Conference on Digital Libraries (JCDL 2007)*, pp. 333-341, 2007.

[3] Garmaabazar Khaltarkhuu and Akira Maeda. Retrieval Technique with the Modern Mongolian Query on Traditional Mongolian Text. In *Proceedings of the 9th International Conference on Asian Digital Libraries (ICADL2006)*, pp. 478-481, 2006.