International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:3, No:10, 2009

A New Model of English-Vietnamese Bilingual Information Retrieval System

Chinh Trong Nguyen, Dang Tuan Nguyen
Faculty of Computer Science
University of Information Technology, Vietnam National University of HCMC

***Abstract*—**In this paper, we propose a new model of English-Vietnamese bilingual Information Retrieval system. Although there are so many CLIR systems had been researched and built, the accuracy of searching results in different languages that the CLIR system supports still need to improve, especially in finding bilingual documents. The problems identified in this paper are the limitation of machine translation's result and the extra large collections of document to be found. So we try to establish a different model to overcome these problems.

***Keywords*—**Bilingual Information Retrieval, Cross-lingual Information Retrieval, Bilingual Web sites.

I. Introduction

THERE are many researches on CLIR and many efficient CLIR systems had been built to serve the need of finding information in different languages. However, the accuracy of results of many CLIR systems is not very high, especially in finding bilingual documents. After making some surveys on current CLIR models, the common architecture of CLIR is summarized as follow [1][2][3][4][5][6][7]:

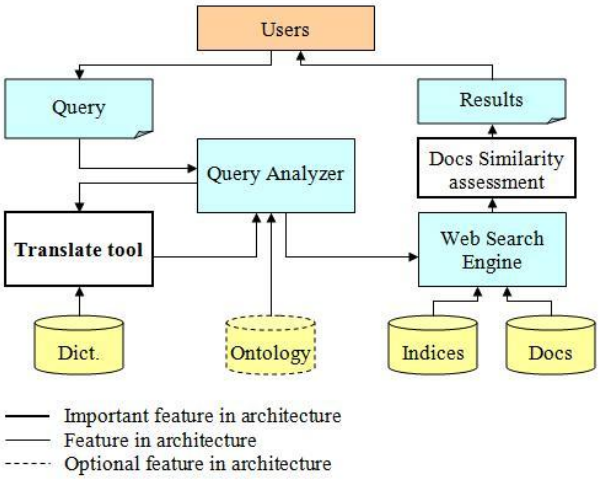

Fig. 1: Common architecture of current CLIR

In architecture showed in Fig. 1, the Translation tool and Docs Similarity assessment are features which make CLIR systems different.

According to the above architecture, there are two processes using machine translation: query analyzing and document similarity assessing. These are points which make the accuracy of retrieval results decreased because of two reasons:

Firstly, in query analyzing process the input query has to be translated into others languages for searching. Because of limitation of machine translation, the translated queries do not have the same meaning as the original query. So, the returned documents from sending these queries to the Web Search Engine will be more differ from expected documents.

Secondly, in document similarity assessing all returned documents in different languages have to be translated into a selected language to calculate the similarity. In this process, the translation makes the content of documents differ from their origin and then the limitation of similarity identification methods make results more inaccurate. One of reasons is the large collections of document in which the CLIR system works. In large collections, there are many documents whose contents are about the same subject or event but their meanings are quite different. After translating, these translated documents are more difficult to identify exactly which is similar to the others, especially to identify which is translated version.

The idea to increase the quality of searching document in CLIR system is to reduce the number of processes using machine and to narrow the collections in which the CLIR system search for the similar or translated documents. This means the CLIR system will search in separate sites instead of all collected web pages to find similar or translated documents.

The idea of limiting the collection to find similar or translated documents in separate sites comes from many surveys on web sites in Vietnam. There are more and more institutes or enterprises in Vietnam expose themselves to everybody by using their web sites in Vietnamese and in English. Some of them; such as People's Committee of provinces, have a special group translating Vietnamese web pages in their site into English. Their web sites are really English-Vietnamese bilingual web sites.

II. Proposal Model

Proposal model of English-Vietnamese bilingual IR system in this paper works on bilingual web sites. Bilingual web site is defined as follow:

Definition 1: *Given a web site W using two languages L_1 and L_2 to present. Assume $WP_1 = \{p_{11}, p_{12}, p_{13}, .., p_{1n}\}$ is set of W's web pages presented in L_1; $WP_2 = \{p_{21}, p_{22}, p_{23}, .., p_{2m}\}$ is set of W's web pages presented in L_2. W is a bilingual web site if for each p_{1i} in WP_1 there is a p_{2j} in WP_2 which is a*

translated version of p_{1i}.

All of web sites of People's Committee of province in Vietnam are bilingual web sites following definition 1.

The proposal model is as follow:

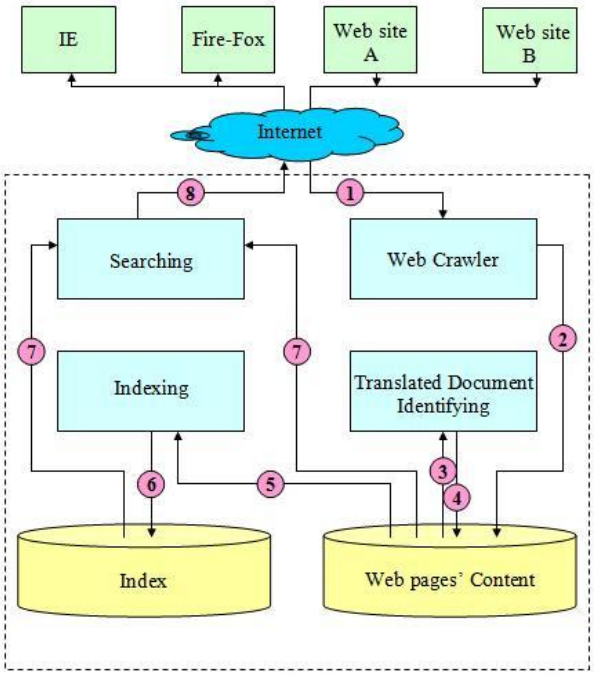

Fig. 2: Proposal model of English-Vietnamese Bilingual IR system

The proposal model has four components:

- Web Crawler: this web crawler is similar to other web crawlers in many search engines except that it crawls only on a specified site and gets only web pages presented in Vietnamese or English. All web pages will be stored by language within their domain name in database so that they can be analyzed as in a site and in a language later.
- Translated Document Identifying: All of web pages of a site crawled will be processed to identify the translated pages of each page. The results will be stored in database for searching.
- Indexing: All of web pages crawled in all site will be indexed. There are two index systems for English pages and Vietnamese pages.
- Searching: Users send query to this component and then get results from it.

A. Web Crawler

Web Crawler in the proposal model is designed to get only web pages in specified web site as follow:

- Web page Loader: downloads documents at specified URL. The document can be a HTML document or somewhat it can parse.

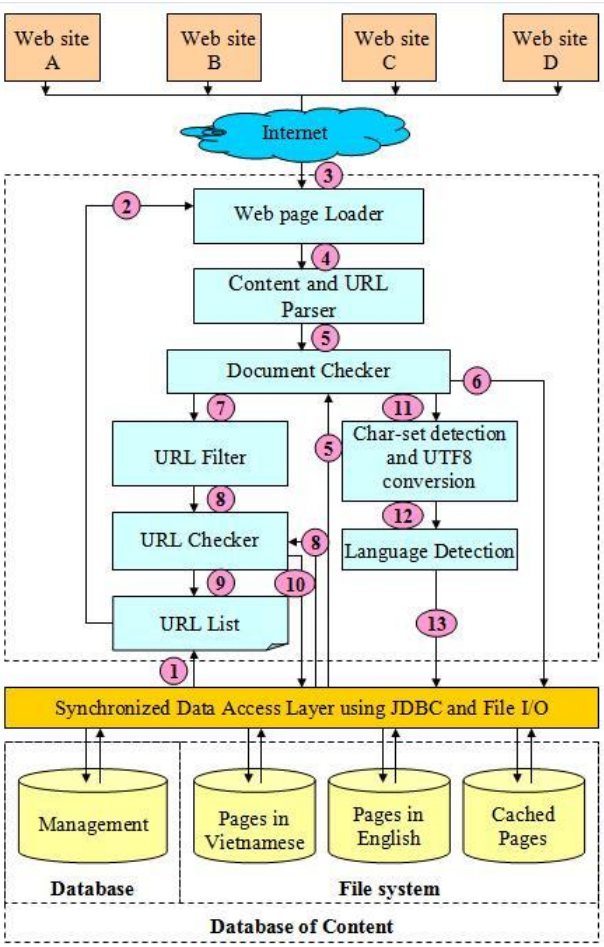

Fig. 3: Web Crawler

- Content and URL Parser: parses the document gotten from Web page Loader. The results are an URL list and a main content of the document.
- Document Checker: checks for existence of the main content in downloaded documents. If the main content exists, all of following step will be skip. Document checker use Rabin's fingerprinting method [8].
- URL Filter: removes all URLs in different domain name, URLs to document cannot be parsed by Web page Loader, and invalid URLs.
- URL Checker: checks for existing URLs that are downloaded or being downloaded.
- Char-set detection and UTF8 conversion: detects char-set of the document and converts to UTF8. All documents in proposal model are stored and processed in UTF8.
- Language Detection: Detects if the document is written in Vietnamese or in English. Detecting method is based on W. B. Cavnar and J. M. Trenkle's Text categorization method [9]. The document will be stored in Vietnamese collection or English collection under domain name of the site.

B. Translated Document Identifying

Translated document identifying is very simple. It is designed as follow:

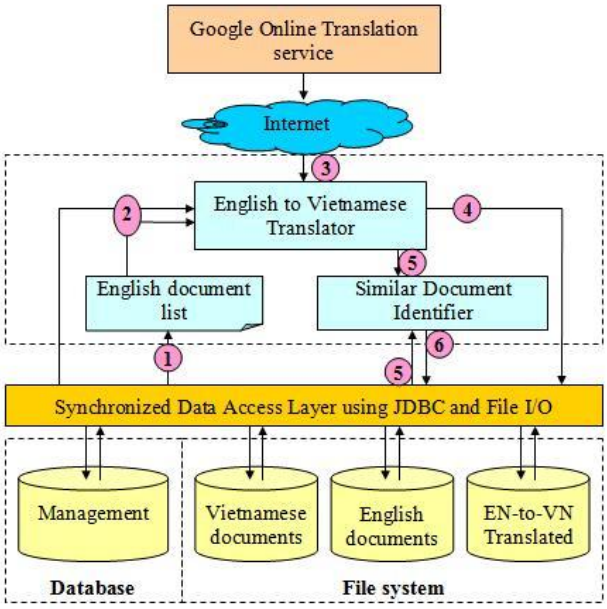

Fig. 4: Translated Document Identifying component

- English to Vietnamese Translator: translates all English documents in each web site into Vietnamese using Google Translation service. Google Translation service is a very good free online translation service. It is based on statistical machine translation method, so the word phrases in translated results are similar to expected word phrases. Although it still has some problems in translating, it is excellent at translating abbreviations. English to Vietnamese translation process has been selected because the NLP methods applying in English are better than Vietnamese. So the results of translation are better and stored in separate place under domain name of the web site.
- Similar Document Identifier: identifies the similarity between Vietnamese documents and each translated document. A pair of a Vietnamese document and an English document of which the translated result has the highest similarity with the Vietnamese document will be identified as bilingual documents. Bilingual documents can be identified like this because they are in bilingual web site.

C. Indexing

There are two collections to be indexed. They are Vietnamese collection containing all Vietnamese documents and English collection containing all English documents in all crawled web sites. Each collection has it own index system. Indexing processes is based on Lucene's API library [10].

D. Searching

Searching process in the proposal model differs from others in CLIR systems. It uses the result of Translated Document Identifying component for searching. It is designed as follow:

- Query Normalization
- Language Detector: detects if query language is Vietnamese or English to have the appropriate analyzer because each language has different stop words, keywords. If the query's language is other than English or Vietnamese, it will be skipped.

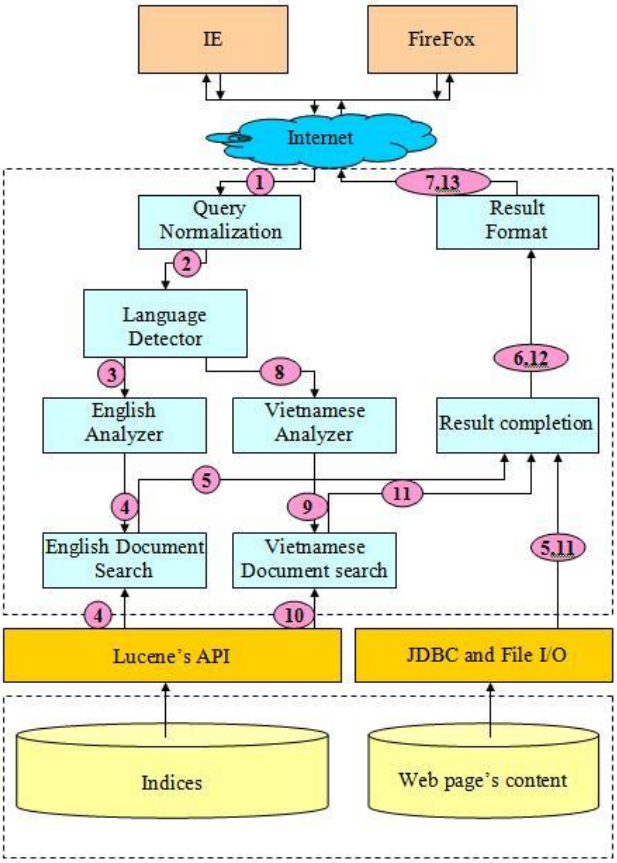

Fig. 5: Searching component

- English Analyzer: removes all stop words in the English queries, extracts keywords from them and them stems keywords using Porter algorithm.
- Vietnamese Analyzer: removes all stop words in Vietnamese queries and extracts keywords from them using a dictionary.
- English Document Search: uses Lucene's API library to search on index system of English collection.
- Vietnamese Document Search: uses Lucene's API library to search on index system of Vietnamese collection.
- Result completion: completes search results. After searching by English Document Search or Vietnamese Document Search, the returned documents are all in language of query. Therefore, result completion uses the results of Similar Document Identifier stored in database to get the translated documents of returned documents.
- Result Format: prepares the results to show in browsers.

III. Advantages and Disadvantages of Proposal Model

A. Advantages

The proposal model is designed for English-Vietnamese bilingual web site so it takes some advantages from bilingual web site:

- The proposal model utilizes the translated results of translating groups in institutes and enterprises, especially in People's Committee of provinces in Vietnam.
- The translated document identifying method is no need to be a high precise method. Because the set of documents to identify is small and they are in a bilingual web site, the precision of the method used in this set will be higher than used in the set of all documents on Internet; even though, the method can be a plagiarism detection method. In experiment, that the custom plagiarism detection for Vietnamese documents has been used for examining bilingual documents has quite good results.
- In the proposal model, English documents are translated only once, and then each of them is identified if it is the translated version of some Vietnamese document. The results are stored in database for using later. This mean all documents crawled is required to translate and identify translated version only once for all searching operation later. This is an advantage in comparison with other CLIR systems.
- Another advantage of the proposal model is that the query does not have to be translated in the other language to search over two collections. Because of limitations in machine translation, the searching only in collection whose language of documents are the same as language of query reduces the incorrectness in translated queries. Therefore, searching results are more accurate.

B. Disadvantages

Completely based on bilingual web sites, the proposal has some disadvantages:

- The results of searching will be poor in subject and small in number of results if the number of bilingual web pages in bilingual web sites is small. Although there is quite small number of bilingual web sites now in Vietnam, it is increasing because more and more institutes and enterprises in Vietnam need to introduce themselves to people in the world while they have to keep their information updated. Moreover, that there are many new free, powerful, easy to use CMS supporting multilingual will encourage more institutes and enterprises establish and keep their multilingual web sites updated.
- The system built upon proposal model cannot identify the translated version of a document if they are in two separate web sites. This means that there are a document D_1 and a translated version of it, T_1; D_1 is in web site W_1, T_1 is in web site D_2; and there is no translated version of D_1 in W_1. In this situation, the system cannot result both D_1 and T_1. This disadvantage comes from the model which completely based on bilingual web sites.

IV. Conclusion

This paper proposes a new model of English-Vietnamese bilingual IR system. The proposal model is designed to utilize the translated results of translating groups in institutes and enterprises which need to introduce themselves through their English-Vietnamese bilingual web sites. The proposal model reduces the incorrectness of results in CLIR system by reducing the number of translating processes and narrow the size of collections in which the system identifies the translated versions.

The proposal model is in experiment to evaluate the applicability of it.

References

[1] Ranbeer Makin, Mikita Pandey, Prasad Pingali and Vasudeve Varma. "Experiments in Cross-lingual IR among Indian Languages". Advances in Multilingual and Multimodal Information Retrieval. Springer Berlin/Heidelberg, 2008.

[2] Jeanine Liileng and Stein L. Tomassen. "Cross-lingual Information Retrieval by Fearture Vectors". Natural Language Processing and Information Systems. Springer Berlin/Heidelberg , 2007.

[3] Jagadeesh Hagarlamudi and A Kumaran. "Cross-Lingual Information Retrieval System for Indian languages". 8th Workshop of CLEF, 2007.

[4] Manoj Kumar Chinnakotla, Sagar Ranadive, Om P. Damani and Pushpak Bhattacharyya. "Hindi to English and Marathi to English Cross Language Information Retrieval Evaluation". 8th Workshop of CLEF, 2007.

[5] Martínez-Santiago, A. Montejo-Ráez, and M.A. García-Cumbreras. "SINAI at CLEF Ad-Hoc Robust Track 2007: Applying Google Search Engine for Robust Cross-Lingual Retrival". 8th Workshop of CLEF, 2007.

[6] Aitao Chen, Hailing Jiang and Fredric Gey. "English-Chinese Cross-Language IR using Bilingual Dictionaries". 2001.

[7] Atsushi Fujii and Tetsuya Ishikawa, "Japanese/English Cross-Language Information Retrieval: Exploration of Query Translation and Transliteration". 2001.

[8] Andrei Z. Broder. "Some applications of Rabin's fingerprinting method". Sequences II: Methods in Communications, Security, and Computer Science. Springer-Verlag, 1993.

[9] W. B. Cavnar and J. M. Trenkle. "N-gram-based text categorization". Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, 1994.

[10] http://lucene.apache.org/