2010 Second International Conference on Computer Modeling and Simulation

Cross-lingual Information Retrieval Model for Vietnamese-English Web sites

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

***Abstract*—Up to now, there are so many CLIR systems has been researched and built. Generally, These CLIR systems are built upon some search engine to skip building a crawler, an indexer and a searcher component. By this way, these CLIR systems do not have enough documents gathered for identifying pairs of similar content documents in languages and they have to send and receive too much data to and from the search engine and web sites while processing user queries. This is a big disadvantage which makes the CLIR systems inefficient. In this paper, we would like to introduce a model of Cross-lingual information retrieval system for Vietnamese-English web sites which include a crawler, an indexer and a searcher and show how gathered documents are processed to efficiently identify and retrieve the similar documents in languages.**

Keywords*—*cross-lingual information retrieval, bilingual information retrieval, Vietnamese-English web sites.

I. INTRODUCTION

THERE are many researches on CLIR and many efficient CLIR systems had been built to support searching information in different languages. The common system architecture of CLIR is summarized as follow [1][2][3][4][5][6][7]:

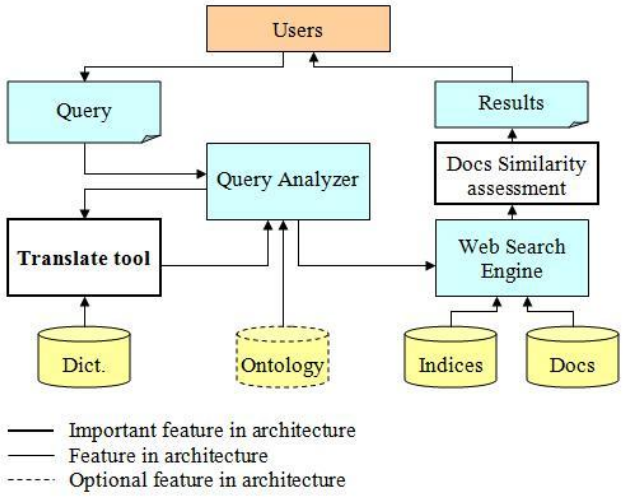

Figure 1.Common architecture of current CLIR

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

According to the above architecture, whenever user search information by using a certain CLIR system, it has to translate the search query into supported languages and sends the translated search queries to a certain Web Search Engine to get lists of document links containing some keywords of the search query in supported languages. After that, the CLIR system has to download document following the lists of document links for similarity assessing process. There are two disadvantages come from the model of the above CLIR system:

- The CLIR system has to send a very large number of queries to a certain Web Search Engine to get document links and then download them for similar assessing process. This makes the bandwidth for the CLIR system increased and slows the network.
- The CLIR system has to do the same tasks to process user queries regardless some similar queries has been processed. This disadvantage comes from the model that the CLIR does not have its own storage for internet gathered documents and its analyzed results. Thus, the CLIR system has to translate the query into its supported languages, send all translated queries into a certain Web Search Engine, download documents following the lists of links the Web Search Engine returned, identify similar documents in languages and then return them to a user every time it receives a query. In above steps, identify similar documents in languages is an important and time-consuming step. If this step repeats many times, the CLIR becomes inefficient.

The idea to increase the efficient of searching document in a CLIR system is to build it as a whole entity which contains components greatly combining to process user requests. These components are also carefully designed for CLIR systems to resolve their problems. Besides, the CLIR system also concentrates on Vietnamese-English web sites. That means the collection of documents the CLIR system crawled for find similar or translated documents is limited in separate sites.

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 Vietnamese-English web sites.

II. MODEL OF SYSTEM

Model of CLIR system in this paper works on Vietnamese-

978-0-7695-3941-6/10 $26.00 © 2010 IEEE
DOI 10.1109/ICCMS.2010.298

English web sites. A Vietnamese-English web site is defined as follow:

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

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:

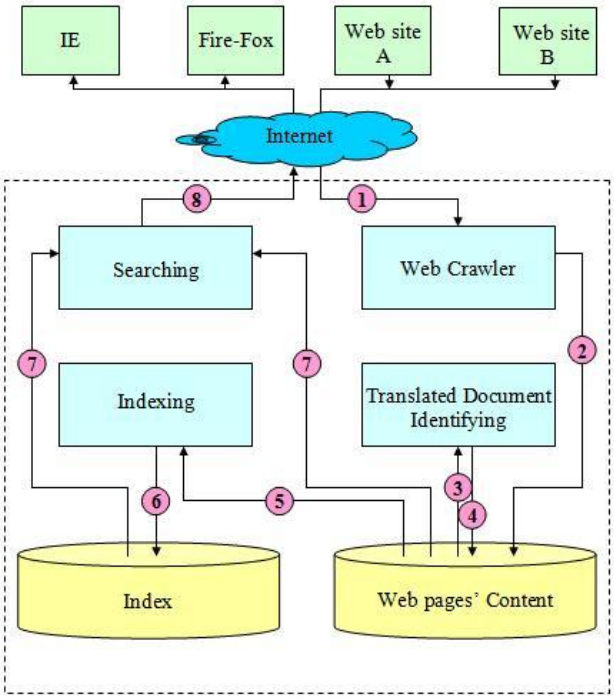

Figure 2.Proposal model of English-Vietnamese bilingual IR system

The system model composes 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 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.

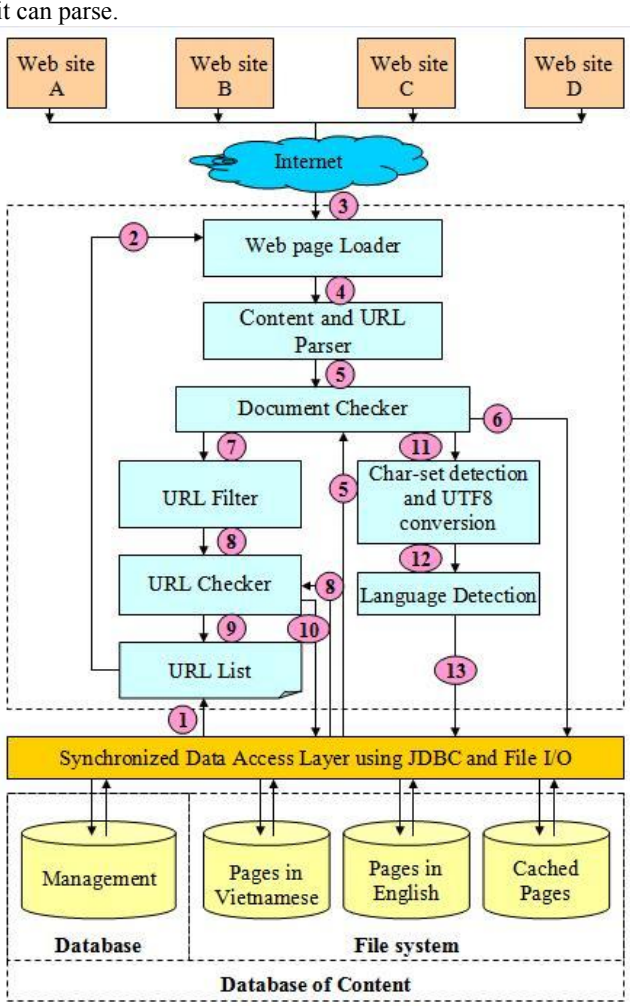

Figure 3.Web Crawler

- 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. 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].

C. 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.
- 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.

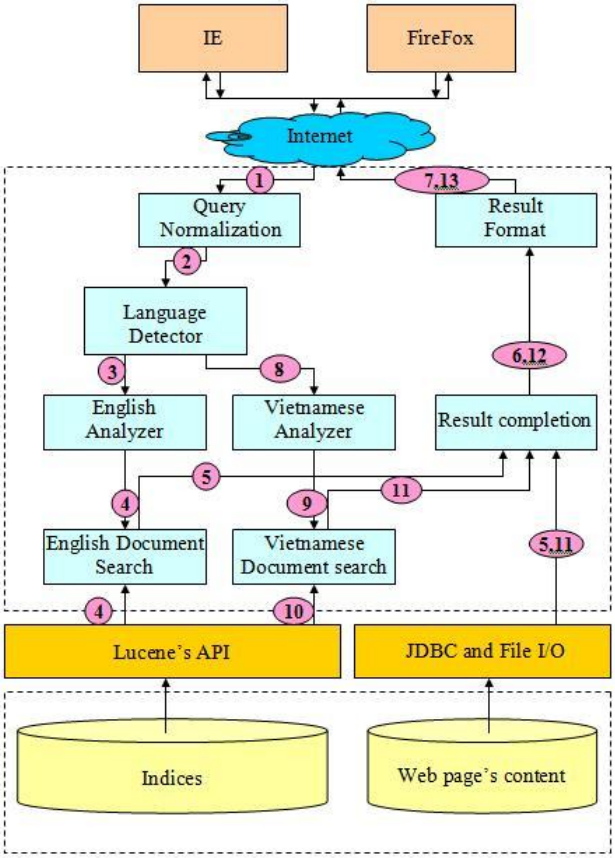

Figure 4.Searching component

III. IDENTIFICATION OF TRANSLATIONS

Translated document identifying component is designed as follow:

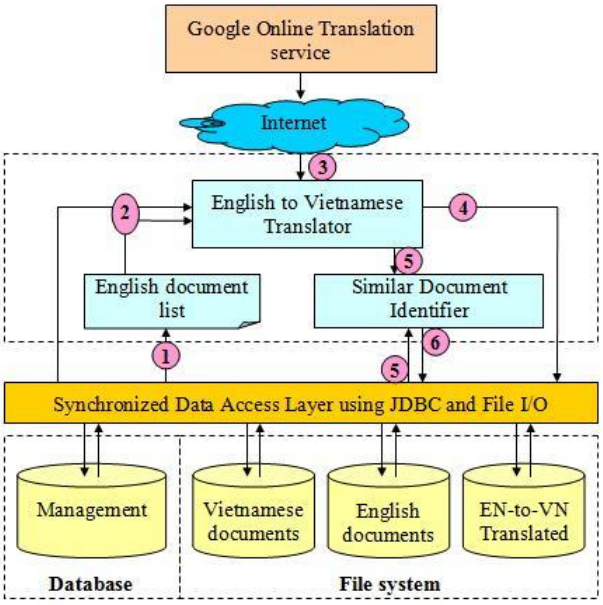

Figure 5.Translated document Identifying component

- 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.

IV. Conclusion and Further works

The CLIR model proposed for Vietnamese-English web sites takes some advantages:

- The CLIR systems based on the model is complete systems include their own storage for documents and analyzed results so that they work more efficient than others which have to download and process documents every time users request.
- The 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 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 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.

However, the model 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 Vietnamese-English web sites.

The model is designed to work as Extended Boolean Search model and uses the Lucene library [10] for Indexer and Searcher components. The system based on the model has been built for experiment and has some pretty good initial achievements such as 95% similarity identified documents are correct, 100% documents are exactly categorized into appropriate language categories.

After evaluating the system completely, we will extend the system model to work as Semantic Search model on Vietnamese and English queries.

References

[1] Ranbeer Makin, Mikita Pandey, Prasad Pingali and Vasudeve Varma, "Experiments in Cross-lingual IR among Indian Languages", 2008.

[2] Jeanine Liileng and Stein L. Tomassen, "Cross-lingual Information Retrieval by Fearture Vectors, 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 Retrieval". 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", 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/, 2009.