

2009

Cross Lingual Information Retrieval Using Data Mining Methods

R. Shriram

BS Abdur Rahman University, shrionsong@yahoo.com

Vijayan Sugumaran

Oakland University, sugumara@oakland.edu

Follow this and additional works at: <http://aisel.aisnet.org/amcis2009>

Recommended Citation

Shriram, R. and Sugumaran, Vijayan, "Cross Lingual Information Retrieval Using Data Mining Methods" (2009). *AMCIS 2009 Proceedings*. 180.

<http://aisel.aisnet.org/amcis2009/180>

This material is brought to you by the Americas Conference on Information Systems (AMCIS) at AIS Electronic Library (AISEL). It has been accepted for inclusion in AMCIS 2009 Proceedings by an authorized administrator of AIS Electronic Library (AISEL). For more information, please contact elibrary@aisnet.org.

Cross Lingual Information Retrieval Using Data Mining Methods

R. Shriram

Department of Computer Science and Engineering
BS Abdur Rahman University
Vandalur, Chennai 600 048, India
shrionsong@yahoo.com

Vijayan Sugumaran

School of Business Administration
Oakland University
Rochester, MI 48308
sugumara@oakland.edu

ABSTRACT

One of the challenges in cross lingual information retrieval is the retrieval of relevant information for a query expressed in a native language. While retrieval of relevant documents is slightly easier, analyzing the relevance of the retrieved documents and the presentation of the results to the users are non-trivial tasks. A method for information retrieval for a query expressed in a native language is presented in this paper. It uses insights from data mining and intelligent search for formulating the query and parsing the results. It also uses heuristic methods for the categorization of documents in terms of relevance. Our approach compliments the search engine's inbuilt methods for identifying and displaying the results of queries. A prototype has been developed for analyzing Tamil-English corpora. The initial results have shown that this approach is suitable for on the fly retrieval of documents.

Keywords

Cross Lingual Information Retrieval, Heuristic method, Text categorization, Searching the Web in Tamil.

INTRODUCTION

The rapid growth in our capabilities for data processing has triggered the explosion of information available on the Internet. However, majority of the information available on the Web is in English rendering it out of reach for non-English speaking users. Cross-language information retrieval (Ballesteros and Croft, 1997; Dumais, Landauer and Littman, 1996) enables users to enter queries in languages they are fluent in, and uses language translation methods to retrieve documents originally written in other languages.

The scope for Cross-Language Information Access (CLIA Report, 2008) goes beyond the Cross-Lingual Information Retrieval (CLIR) paradigm by incorporating query disambiguation as well as post search processing. The key emphasis is on the relevance of the results. The cross lingual information access paradigm may take the form of machine translation of snippets, summarization and subsequent translation of summaries and/or information extraction from the target language. CLIR has three basic approaches (Maeda et al., 2000): a) document translation - where the queries are posed to existing document repositories, b) query translation - where the queries are translated into the target language and results displayed, and c) inter lingual translations - where queries and results are translated. Our approach is a variation of the third category.

Thus, the specific objectives of this research are to

- develop an approach for cross lingual information retrieval for queries expressed in the native language, Tamil,
- use data mining techniques to cluster the results and retrieve a resultant set closest to the user's query, and
- present the results in various display methods to the user.

The key aspect of the proposed approach is as follows. It is composed of two distinct aspects: preprocessing the query to identify the query's meaning and post-processing the results for relevance match. In the preprocessing stage, the query will be expanded and the expanded query is presented to the search engine. In the post processing stage, based on the relevance match of the retrieved content, the resultant documents will be reordered and presented to the user. The initial feedback from the users seems to indicate that the relevance of the retrieved documents is higher than the conventional approach. However, the time needed to perform the processing is a significant factor.

The rest of the paper is organized as follows: The next section describes the proposed approach. Following that, the implementation of the model and the results of the experiments are discussed. Related work and conclusion are presented in the subsequent sections along with future work.

PROPOSED APPROACH

The proposed approach (Figure 1) is composed of two distinct and complementary stages, namely, preprocessing and post processing. In the preprocessing stage, the search query in an Indian language (Tamil) is parsed and disambiguated using the lexicons available for that Indian language and the initial search terms are translated into English. These English terms may be further disambiguated using WordNet and other ontologies and the expanded query is submitted to the search engine. In the post processing stage, the results from the search engine are summarized, mapped to the target language and the results presented to the user. The target language that we want to focus on is Tamil. The results in English need to be summarized for relevance and displayed to the user.

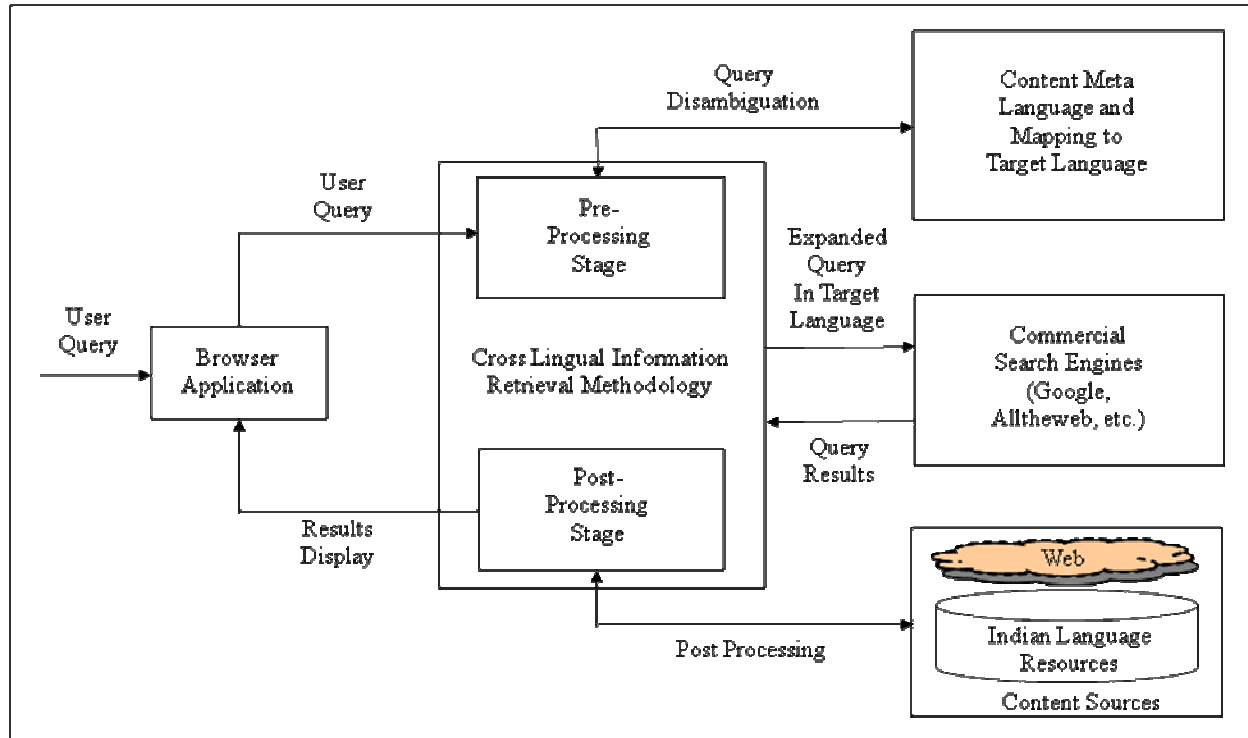


Figure 1. Overall Approach

Preprocessing stage

The preprocessing stage is composed of four distinct steps. These steps have been adapted from the work of Storey et al. (2008).

Step 1 - Query Parsing: The first step involves parsing the natural language query specified by the user in his or her own language. First the query is segmented and the words are disambiguated using an ontology and/or other lexicons available in that language. Then the initial query is translated into English. This translation may be prone to ambiguities if the context is not clearly specified and the words are not chosen carefully from a lexicon.

Step 2 - Query Expansion: The output of the parsing step is a set of initial translated query terms which become the input to the query expansion step. The query expansion process involves expanding the initial query using lexicons and ontologies. It also includes adding appropriate personal information as well as contextual information. For each query term, the first task is to identify the proper semantics of the term, given the user's context. To do so, the word senses from a lexicon such as WordNet are used. For each term, synonym sets are extracted. The appropriate word sense is determined based on the context and other query terms (may also need user input) and a synonym from that synset is added to the query. To ensure precise query results, it is important to filter out pages that contain incorrect senses of each term. Thus, a synonym from the unselected synset with the highest frequency is added as negative knowledge to the query. Since ontologies contain domain specific concepts, appropriate hypernym(s) and hyponym(s) are added as mandatory terms to improve precision. In this step, personal information and preferences relevant to the query is added. For example, user may limit the search to certain geographical area or domain. Such information helps narrow down search space and improve precision.

Step 3 – Query Formulation: The output of the query expansion step is the expanded set of query terms that includes the initial query terms, synonyms, negative knowledge, hypernyms, hyponyms, and personal preference information. This expanded set becomes the input to the query formulation step. In this step, the query is formulated according to the syntax of the search engine. For each query term, the synonym is added with an OR operator (e.g. query term OR synonym). Hypernym and hyponym are added using the AND operator (e.g. query term AND (hypernym OR hypernym)). The negative knowledge is added using the NOT operator. The first synonym from the highest remaining synset not selected is included with the NOT operator (e.g. query term NOT synonym).

Step 4 – Search Knowledge Sources: This step submits the query to one or more search engines (in their required syntax) for processing using the API provided by them. The query construction heuristics work with most search engines. For example, AltaVista allows queries to use a NEAR constraint, but since other search engines such as Google and AlltheWeb do not, it is not used. Likewise, query expansion techniques in traditional information retrieval systems can add up to 800 terms to the query with varying weights. This approach is not used in our methodology because web search engines may limit the number of query terms. In this case, we use Yahoo API for our approach.

To illustrate the process, consider the sample query ‘Anna palkalai kazhagam’ expressed in Tamil language. In the query parsing stage, the query is checked against the lexicons in Tamil. For the term ‘Anna’, the meanings elder brother, brother, and person are retrieved in Tamil. For the term ‘palkalai’ the meaning ‘multiple arts’ is found. ‘Kazhagam’ refers to organization, institution or society. These terms are the results of the disambiguation in the native language. In the query expansion stage, based on the expansion from lexicons and ontologies, the meaning ‘palkalaikazhagam’ is derived. This is then confirmed by the user. The meaning for palkalaikazhagam is ‘University or institution of fine arts’. After interaction with the user, the term University is confirmed. After the translation to English, the system suggests that that previous phrase (Anna) may refer to a proper noun – name, place, or origin. Now, the phrase in the target language is derived as ‘Anna University’ and the query is formed as: Anna + University. Thus, the query formulation, translation and assembly are done only at the last possible instance after inferring the context of the user’s query. The personal preference parameter is now added as the place, namely, Chennai. So, the overall query is formed as ‘Anna University Chennai’ and sent to the search engine for processing. Thus, the disambiguation and processing takes place twice in the system – once in the native language and once after translation into English.

Post processing stage

In this stage, the results from the search engine (URLs and ‘snippets’ provided from the web pages) are retrieved, and translated to the target language. Available lexicons and ontologies are also used in the translation. Further, a heuristic result processing mechanism described below is used to identify the relevance of results retrieved with respect to the source language. Finally, these are aggregated and the resultant summarized content is presented to the user. The user can either accept the results or rewrite and resubmit the query to get more relevant results. This stage also integrates the search results from multiple heterogeneous sources and takes care of the differences in formatting. The results are organized based on the knowledge sources used, target language, or domain. Overall, the objective of the post processing stage is to parse and re-organize the retrieved results. While the query undergoes transformation in finding appropriate equivalents, the query results also need to be processed using similar transformation methods. The approach used is from the information retrieval perspective, which also integrates the insights gained from data mining.

Our approach (depicted in Figure 2) visualizes the problem of categorizing the results akin to the results merging approach discussed in Si et al. (2008). The approach starts with the set of retrieved result snippets or documents for each user query which is given by the search engine. These results, though relevant, need another stage of post processing to bring the results closer to the user’s query. The objective of the post processing is to aggregate the results for the given query, rank and reorder the results, and present the results in a new sorted order based not only on the output of the search engine but also on the content of the retrieved documents. These steps are analogous to the strategies used in Feature subset selection and hence some of the metrics such as information gain, pair-wise relationship and cluster relationship are applied for the problem after appropriate modification. Thus, these downloaded documents are indexed and comparable document scores for the downloaded documents are calculated. The metrics utilized are word relevance, word to document relationship and clustering strategies. Finally, all the returned documents are sorted into a single ranked list along with display mechanisms for helping the user view and decide on the results.

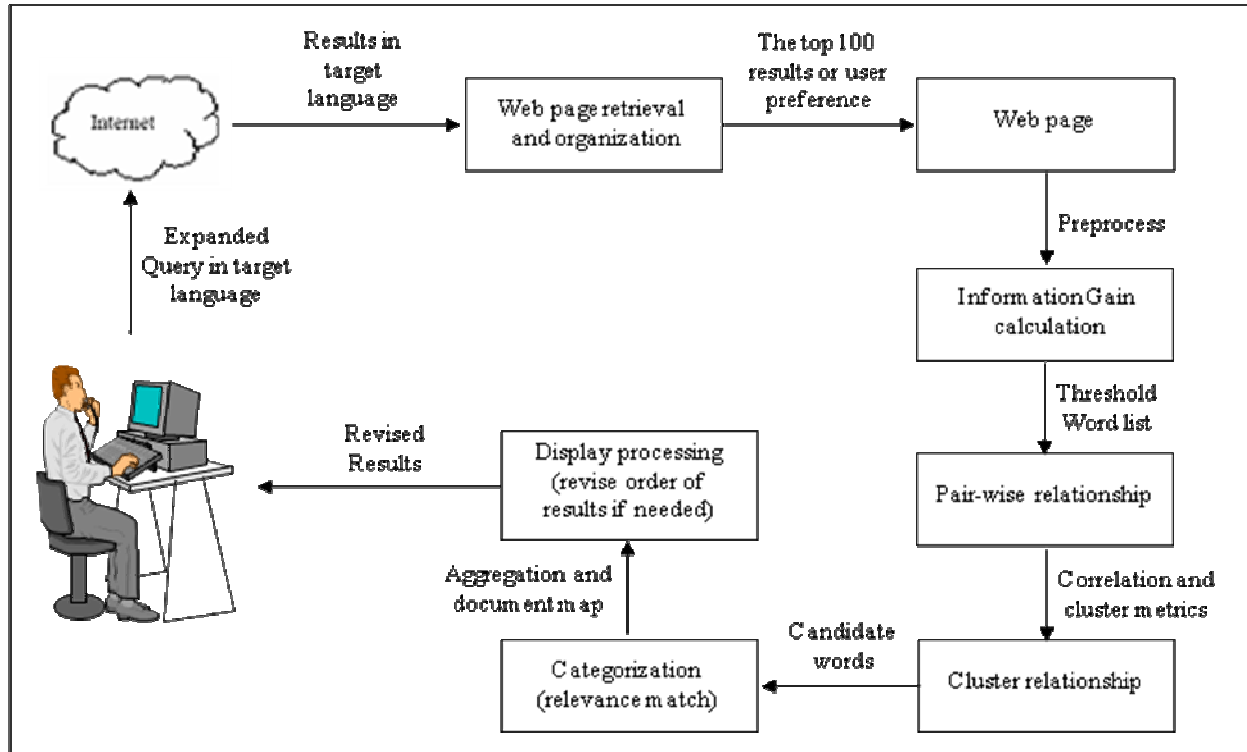


Figure 2. Post-processing Methodology

The information retrieved (results obtained through the search process) needs to be organized appropriately for effective processing. The following measures are used for this purpose. These measures have different connotations in traditional data mining. This paper applies the semantics of these parameters in the context of cross lingual information retrieval.

- Candidate word: words that have high information gain in a given document.
- Information gain: the parametric importance of a word in the retrieved document.
- Pair-wise measure: the correlations between the candidate word and the input keywords are found using the pair wise measure. They also give the relationship between words and help disambiguate similar words.
- Cluster relationship: this gives a measure of the degree of clustering of candidate words in a given document.

The post processing stage (Figure 2) of the proposed information retrieval method has the following five distinct steps:

- Parameter estimation:** This step scans the document and identifies at a high level the various keywords that are present. Two methods have been applied in the pre-processing stage 1) removal of stop words (stop word list) and 2) stemming algorithm: porter's stemming algorithm. After the pre-processing,
 - The information gain of keywords is calculated using statistical methods. Information gain is defined as the number of times the word (meaning) occurs in the document. This means that for each word, its meaning and related words are taken into account when the importance is found. This is accomplished by using ontologies.
 - For words with maximum information gain, the pair wise relationship of each keyword is also observed. While the information gain for each word will give the measure of importance of each word in isolation, the pair-wise measure gives a broad measure of the inter-relationships between the word and the input key words. The pair wise measure is found using correlation based methods. If two words are not related the correlation between them is 0. If they are dependent, the values are +1 or -1. Based on the above two pairs, for words which satisfy the threshold limit, the pair wise measure is obtained. The pair wise relationships help guide the search for candidate words.
 - The mutual correlation between each word and all the other candidate words are found. This is continued till the features with maximum correlations are identified and sorted. The cluster relationship is the measure of distribution of the word in the document. This gives a measure of whether the word is restricted to a narrow area in the document or is uniformly spread throughout the document. The threshold limits for each parameter are derived dynamically based on the distribution of the document.

- b) **Categorization:** The words, which satisfy the threshold limits for each of the three parameters, are selected as candidate words. These three measures help establish relationships between the words and give a measure of the relationships in the documents. Thus, the values found not only represent the values, but also help the system discover previously unknown relationships between various candidate words. The advantage of having an incremental and heuristic method is that it allows the categorization to iterate in a thorough manner without losing any aspect of the document. Based on the parameters a document map is formulated. The document map (intermediate structure) is an intermediate representation displaying the key candidate words at the place of occurrence. The document maps help in characterization, differentiating and grouping content.
- c) **Aggregation:** The same method (steps *a* and *b*) is applied for all the documents and the results are aggregated. Based on this, the documents (Figure 3) are re-ordered using the parametric results and the categorization stage. The aggregation problem is similar to the results merging problem in information retrieval. In this approach, the contents of documents retrieved by the search engine for the given query are validated. This stage eliminates irrelevant documents, aggregates different versions of the same document and organizes the overall display to be of documents that are relevant and closest to the user query. The users are shown the document map and the documents. As these three stages are iterated for every query, this method is suitable for on the fly and run time information retrieval.
- d) **Display processing:** At this stage, along with the document map, a word-by-word translated map in user's language is also shown. Thus the users can select documents that they feel are relevant from the map in their own language. A further level of display processing is performed for the output. For the selected document, when the user takes the mouse over a word, an equivalent word in native language is retrieved and displayed. This method (applied in rikai.com) helps the users by enabling them to understand the drift of the contents. In the future, translation mechanisms will be applied for the retrieved text.
- e) **Learning:** Based on the results of the above stages, the learning algorithm assigns weights to the parameters and learns (using machine learning methods) from the selection options made by the users. The feedback and suggestions from the users are collected for document mapping. Simultaneously, the above method is applied for documents in the native language where parallel corpora are available.

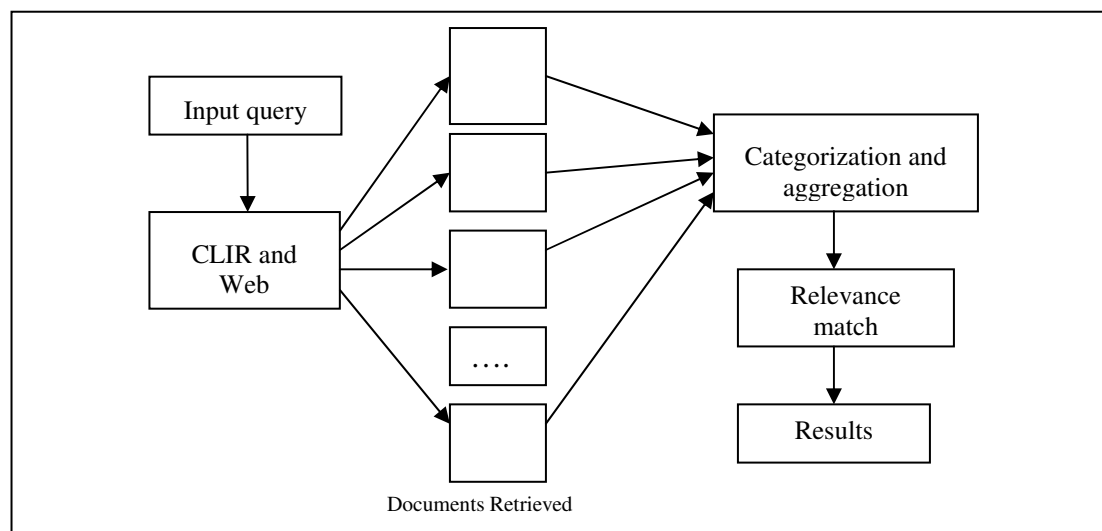


Figure 3. Results Transformation Method

The key aspects in this system are the mapping mechanism between words in different languages, aggregation method at document/repository level, structure of document maps and the learning system. For each candidate word the pair wise measure gives a measure of correlation. However, these correlations are not available in dictionary representations and must be generated by use of appropriate ontological systems. Thus, a search and traversal method that navigates the ontology for distance measure is crucial for finding the pair wise and relationship measures.

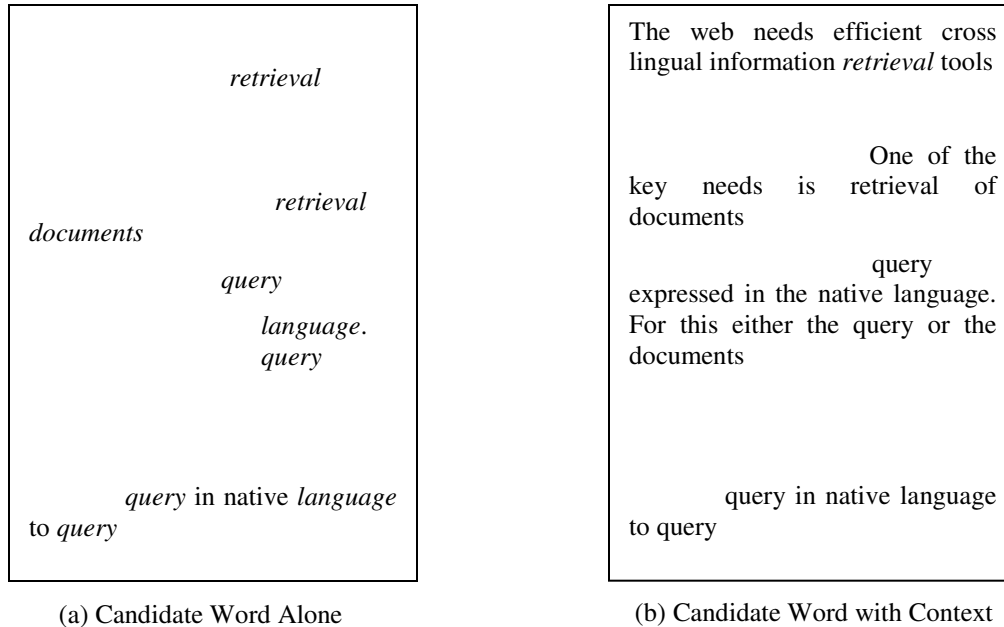
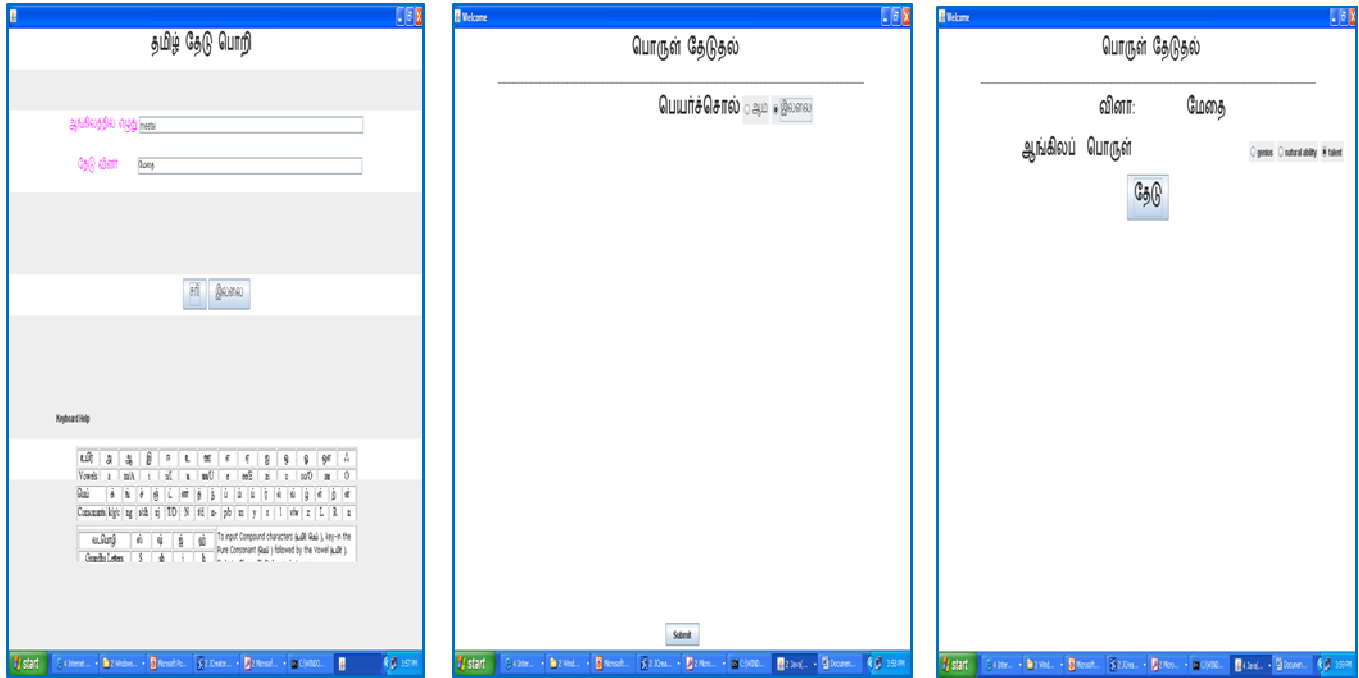


Figure 4. Document Map Methods

The languages used are Tamil (native language) and English. The aggregation method at the document and repository levels takes the candidate words and prepares a document map that is abstract, but with sufficient level of detail to ensure that users get a sense of what is being conveyed. The display (Figure 4) is in two levels, candidate word alone and candidate word with context. At this stage the candidate word level document map has been designed to show the distribution of the clustering of the top-level three candidate words (users can customize the number of words). The candidate word context level helps the user in understanding the detail. Based on the design experiences, the users have always opted for viewing the context of the three top-level candidate words. The learning system learns from user selections and adjusts itself for future selections. A discussion on the relevant literature is given in the next section to place our approach in the proper context.

IMPLEMENTATION AND RESULTS

A prototype model using the proposed method has been implemented using the Yahoo Search API methods in Java. The prototype is a work in progress. A multidimensional ontological structure has been constructed using Data cubes for the language based processing. The system processes the input for understanding the context of the query. It was observed that this stage of query processing requires interaction with the users on the exact context of their query. While personalization was not attempted here, the interaction in most cases improved the content of the results. Some screen shots of the system is given in Figure 5. For example the query “poovizhi” in Tamil was categorized as a noun with the meaning “flowery eyes”. After the suggestions provided by the querying system were taken into account, the noun was further categorized as “Film name” or “song name”. After this, the expanded query is sent to the search engine. The results are retrieved and categorized appropriately. The documents are downloaded individually and mapped using the post processing algorithm. The candidate word maps are calculated for each document and ranked. The ranked list is reordered appropriately and presented to the users. In the document map, when the user places the cursor over the keywords, the meaning of the words is shown. This helps the users identify the relevance by themselves and make appropriate decisions. The crux of the post processing algorithm is in the fact that it organizes the results according to the content of the ‘query’ expressed in the native language and eliminates any results to the contrary if presented by the search engine. Thus, if the search engine retrieves results corresponding to flowery eyes, then these are eliminated as the context specified by the users are limited to film names or song names alone. At this stage, the system works only for html documents and not for content in pdf or word formats. There is a latency associated with the effort. To overcome that in the post processing of the query, results are narrowed to a limited set of query context results of each word tense. This helps the post processing system have a manageable domain of inputs for further processing. In case the query results are not enough, the domain is expanded stage by stage.



(a) Entering the Tamil Query “Medhai”

(b) Proper Noun Checking

(c) Dictionary Matching

Figure 5. Screenshots from the Prototype System

The prototype was pilot-tested by ten users, consisting of under graduate/post graduate students and faculty. They had knowledge of the local language and were familiar with using Tamil Search Engines. However, the respondents were not familiar with the underlying processes involved in the proposed approach. They were given a set of tasks which involved retrieval of certain information. Altogether 30 tasks, which involved multiple words in a query (1-5), were given to the users. The user’s task was to test the traditional system and the prototype system. To model the normal processing, the search results without any pre or post processing were presented to the user using the same set of interfaces. The participants did not interact with each other during the validation process. The tasks ranged from simple keywords involving nouns (genius, leader), to proper nouns (anna, crescent), multiple related words (anna university, exam results) to finally multiple unrelated words (leader serial test). The users were asked to mark the relevance of the results on a ten point Likert scale. The time taken was estimated by using interfaces in the prototype. The results were generated using simple summary statistics and interviews with the users. The results (Figure 6) indicate that the relevance of the documents obtained through our approach is slightly higher compared to those obtained without our approach. The methodology used to validate the search results is to display and compare the unprocessed results of the search from the traditional approach against the results from our approach.

In each session, the time taken to display the search results (unprocessed set versus processed set of documents) was also measured. The delay times (Figure 7) experienced by the users during the use of our approach was found to be high. During the post processing stage considerable amount of work is being done in running the scripts and displaying the results in different formats for the users. Thus, there is a tradeoff between access relevance and access time and users have to decide which is more important and strike a balance. However, it must be stressed that the users did not feel uncomfortable with the delay and were not turned off. Considerable amount of work is being done now to reduce the delay. The operation of the system during high loads is also being studied to check the impact of the system on user satisfaction.

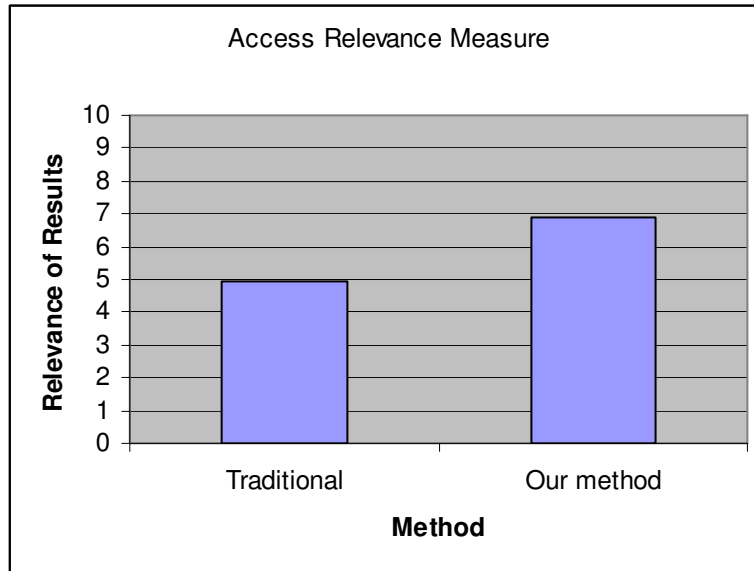


Figure 6. Document Relevance Comparisons

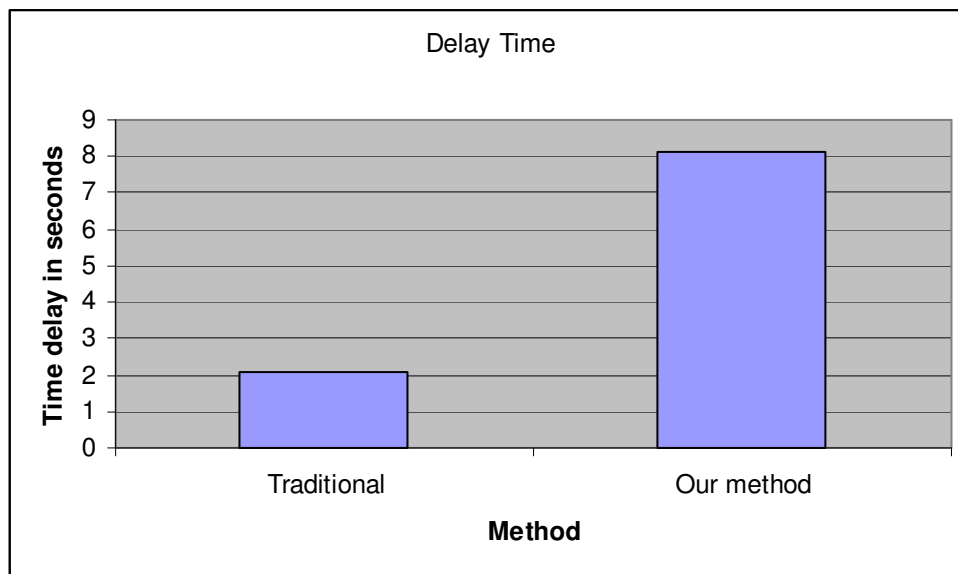


Figure 7. Delay Time Comparisons

When given a choice, the users used the context based document map more for their results display which validated our specific post-processing display approach. The use of mouse over option on the results displayed by the system was also high, leading us to the conclusion that these strategies can be applied until on the fly translation systems are developed.

RELATED WORK

There are four Web-based common methods that are relevant to this research. The first method is where either the query or the documents are mapped into a common representation system. For the transition from query in native language to query results in English, a dictionary based system and interaction with users have been used to disambiguate the query. The next step is to search the web using appropriate key word equivalents in English, summarize the results and display them to the users. This method relies on the dictionary look up and user interaction for proper translation, which may not be always feasible as the query given may contain proper nouns (Lu et al., 2002). The second method is to collect information from parallel corpora on the web (Li and Yang, 2006). These methods are suitable for information extraction where the resources

are available, and not for run-time and on the fly access where the queries are short and diverse. The third method is for closely related languages and some intrinsic relationships (Jarvelin et al., 2008). While this approach is fairly advanced, this is not applicable for all languages or corpora. The fourth method is using Anchor Text Mining (Lu et al., 2004). In this approach the anchor text sets are used as bilingual corpora to reduce the existing difficulties in translation. Anchor text mining needs mechanisms (Cheng et al., 2004) for training and validation, which take up considerable time and effort. As discussed above, existing methods have severe limitations and have limited success. Hence, there is a great need for developing a systematic approach that provides results back to the user that is relevant to the cross lingual query.

A multilingual multi-stage approach is described in Capstick (2000). It involves user interaction to restrict the domain of interest and understand the context of the query similar to our approach. Our work, while not expressive enough to support multiple languages, uses more comprehensive categorization and ranking approaches. The use of Yahoo Search APIs for Question Answering has been explored by Gomez, Rosso and Sanchis (2007) where the snippets obtained in the Yahoo Search Engine are re-ranked. Our approach, while similar, goes beyond their N-gram approach and adopts a three stage processing. The use of clustering and document re-formatting for retrieved documents has been described in Steinberger, Pouliquen & Ignat (2005). The methods used in our algorithm are more restricted in nature than the multilingual approach described. The other key difference is that our approach is geared towards dynamic information access from the web as opposed to searching in an existing collection. We also adopt a different result display method. An approach towards CLIR based on query expansion of bi-directional dictionaries and ontologies is described in Rao and Sobha (2008). Our work utilizes a more complex query expansion and post result processing approach based on interaction with the users.

CONCLUSION

This paper has outlined an approach for Cross Lingual Information Retrieval which emphasizes pre and post processing strategies for the queries entered in a source language. Mechanisms for displaying the results have been outlined which give the users a better idea of what the documents contain. The approach works on top of existing search engines and helps refine the search process further. Initial results are encouraging. The drawbacks of the system at present are the high delay times experienced by the users, which we hope to alleviate in due course. The system needs further refinement and detailed testing before being deployed. As part of future work, the system will be deployed in a business context with content both in Tamil and English. Controlled experiments will be used to evaluate the potential impact of the system on sales. The current prototype has been designed as a stand-alone application. However, the deployment of the system as a part of proxy servers is envisaged in the near future. The initial results from the relevance point of view show that the methodology has sufficient promise for further development.

REFERENCES

1. Ballesteros, L. and Croft, W. B. (1997) Phrasal Translation and Query Expansion Techniques for Cross-Language, Information Retrieval, *Proceedings of SIGIR'97*.
2. Capstick, J., Diagne, A. K., Erbach, G., Uszkoreit, H., Leisenberg, A., Leisenberg, M. (2000) A system for supporting cross-lingual information retrieval, *Information Processing and Management*, Vol. 36, No. 2, January, pp. 275 – 289.
3. CLIA Research Report, "Development of Cross Lingual Information Access (CLIA) System" funded by Government of India, Ministry of Communications & Information Technology, Department of Information Technology (No. 14(5)/2006 – HCC (TDIL) Dated 29-08-2006) retrieved from www.mt-archive.info/IJCNLP-2008-CLIA.pdf on Feb 21, 2009.
4. Dumais, S. T., Landauer, T. K. and Littman, M. L. (1996) Automatic Cross-linguistic Information Retrieval Using Latent Semantic Indexing, *SIGIR'96, Workshop on Cross-Linguistic Information Retrieval*, pp. 16-24.
5. Gomez, J. M., Rosso, P., Sanchos, E. (2007) Re-ranking of Yahoo Snippets with the JIRS Passage Retrieval System, *Proceedings of Workshop on Cross Lingual Information Access, CLIA-2007, 20th International Joint Conference on Artificial Intelligence, IJCAI-07, Hyderabad, India, January 6-12*.
6. Järvelin, A., Talvensaaari, T., Järvelin, A. (2008), Data Driven Methods for Improving Mono- and Cross-lingual IR Performance in Noisy Environments, *Proceedings of the second workshop on Analytics for noisy unstructured text data*, July 24, Singapore.
7. Li, K. W. and Yang, C. C. (2006) Conceptual Analysis of Parallel Corpus Collected From the Web, *Journal of the American Society for Information Science and Technology*, 57, 5, 632–644.
8. Lu, W. H., Chien, L. F., Lee, H. J (2002) A transitive model for extracting translation equivalents of web queries through anchor text mining, *Proceedings of the 19th international conference on Computational linguistics*, pp.1-7, August 24-September 01, Taipei, Taiwan.
9. Lu, W. H., Chien, L. F., and Lee, H. J (2004) Anchor Text Mining for Translation of Web Queries: A Transitive Translation Approach. *ACM Transactions on Information Systems*, 22, 2, pp. 242-269.

10. Maeda, A., Sadat, F., Yoshikawa, M., and Uemura, S. (2000) Query term disambiguation for Web cross-language information retrieval using a search engine, *Proceedings of the fifth international workshop on on Information retrieval with Asian languages*, pp.25-32, September 30-October 01, Hong Kong, China.
11. Rao, P. R. K. & Sobha. L (2008) AU-KBC FIRE2008 Submission - Cross Lingual Information Retrieval Track: Tamil-English, *Proceedings of the Forum for Information Retrieval Evaluation Workshop*, 12-14th December, Kolkata, India.
12. Si, L., Callan, J., Cetintas, S., Yuan, H. (2008) An effective and efficient results merging strategy for multilingual information retrieval in federated search environments, *Information Retrieval*, Vol. 11 , No. 1, February, pp. 1 – 24.
13. Storey, V.C., Burton-Jones, A., Sugumaran, V., Purao, S. “CONQUER: A Methodology for Context-Aware Query Processing on the World Wide Web,” *Information Systems Research*, Vol. 19, No. 1, March 2008, pp. 3 – 25.
14. Steinberger, R., Pouliquen, B., & Ignat, C. (2005) Navigating Multilingual News Collections Using Automatically Extracted Information, *Journal of Computing and Information Technology*, Vol. 13, pp. 257–264.