Language Independent Information Retrieval from Web

R.Seethalakshmi, Ankur Agrawal, Ranjit Ranjan
Shanmugha Arts,Science, Technology and Research Academy,
Tirumalaisamudram, Thanjavur- 613 402
rseetha_in@yahoo.co.in , ankuragr.engg@gmail.com , ranjit.sastra@gmail.com

ABSTRACT

Language independent information retrieval is one of the major issues in the web access by the regional population of any kind. This paper addresses the design and implementation of such information retrieval system. In this system the user is allowed to pose the query in any language and also he can retrieve the information in any other specified language. This approach encounters the design and implementation of a software_morph_parser which encompasses the natural language processing principles and retrieves the information efficiently. The software_morph_parser divides the input search text into individual words and keywords are identified. The keywords are converted into their root forms by removing all their inflexion forms and the corresponding root words are translated into the target language. The multi-lingual web database is dynamically indexed by a dyn_crawler and a search engine is invoked which searches the indexed database and ranks the pages as per the relevance to the keyword. The links are displayed to the user in the priority order of relevance. The user can click on the link and access the web page pertaining to the required information. This system is aimed to breach the language difficulties that the regional population faces in accessing the web.

Keywords

Software_morph_parser, Multi-lingual web, dyn_crawler, search engine, Language-independent, web-page- rank, translator database, indexed database.

1. INTRODUCTION

In this highly intellectual world, internet is the main source of information for the human population. The growth of web as a complete reservoir of knowledge has ushered in an era of Information Revolution. Since its initiation most of the web access is in English language which is the most dominated and preferred one. In recent times the rapid growth in the popularity of computers and the internet in non-English speaking countries like India have increasingly made the need and importance of reaching out to the non-English speakers globally to be felt. With the rapid increase in content written in non-English languages and in multiple languages on the internet, a proper mechanism is needed to make this content noticeable and available wherever and whenever necessary. The "Language Independent Information Retrieval from Web" (LIIRW) is an attempt in this direction. By providing the user with the independence of typing the query in any language of his choice and getting the results in any language or any combination of languages, it is intended to make the multilingual content of the web easily available and more noticeable. The figure 1 below shows the growth of internet in the multilingual domain. As shown in the figure English language dominates with a share of 52% in its individual share. The share of regional languages is very small in comparison. Now with the multilingual web development, the usage of regional languages is increasing very rapidly (Chinese has gone from 5 % to 21% of its own share) and it provides more clarity to the information retrieved.

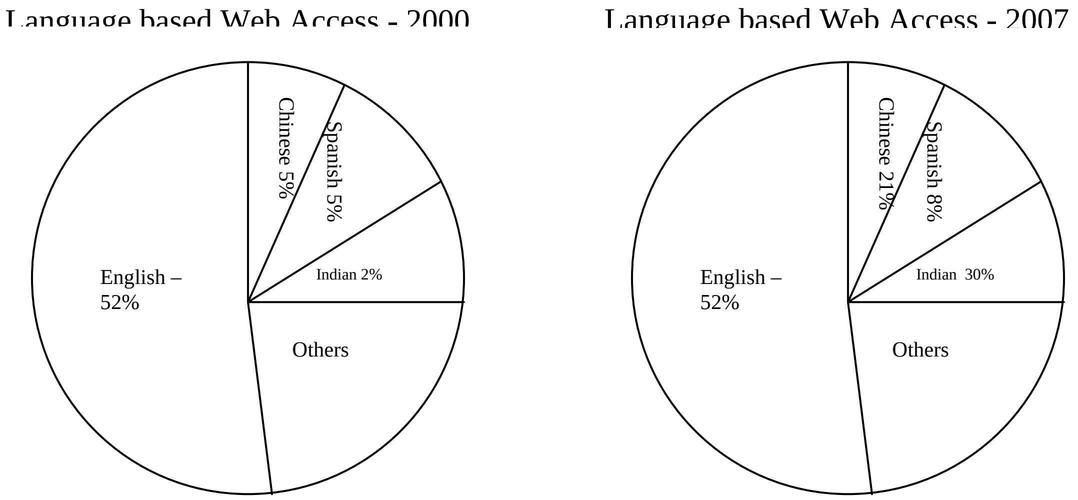


Figure 1: Multilingual Web Access growth

2. Multilingual Approach

Language is a system of arbitrary symbols and the rules used to manipulate and translate across. It is a common approach used by humans for communication. Languages are all of varying nature and type. It ranges from mother tongue, which is the native language of the person, English which is the standard common language to the foreign languages. An important issue in human life is learning. During the course of learning, information flows in different forms. Everyday learning occurs in some form or the other. Knowledge can be derived from newspapers, radio, TV, internet and so on. Of all the forms, Internet is the most dominated efficient and prioritized approach which contains a large repository of information in various languages. As the need for multilingual web access has increased, the various documents are either translated or created new and launched onto the web sites. Multilingual access improves business, transaction processing, entertainment and so on. Further it improves the better understanding of the information particularly when conflict occurs in the contexts of getting the information. Hence this paper addresses the issue of **L**anguage **I**ndependent **I**nformation **R**etrieval from **W**eb **(LIIRW).** This paper addresses the initial implementation of the LLIRW concept in three Indian languages: Hindi, Tamil and English. However the algorithms and techniques used can be easily extended and applied to other languages as well.

LIIRW Functional Block Diagram

The LIIRW is divided into three functional modules which are defined as follows.

1. A Translator named software_morph_parser
2. A Crawler and Indexer called   dyn_crawler and
3. A Search Engine.

Figure 2 pictorially represents the functional block diagram.

2.1.1 Software_morph_parser

The software_morph_parser (which acts as the translator module) takes the search text entered by the user and translates it onto the target language. Target language is the language in which the search results are required. The dyn_crawler acts as the crawler and indexer to dynamically crawl the web pages making an index of the web pages in the LIIRW database. The search engine uses the indexed database to search for the pages relevant to the search text. The indexed database is divided into three, one each for Tamil, English and Hindi. The search is made in the database of the target language. If it is desired to obtain the results from all the three languages a search is made in all the three databases and the results are sorted in the increasing order of relevance.

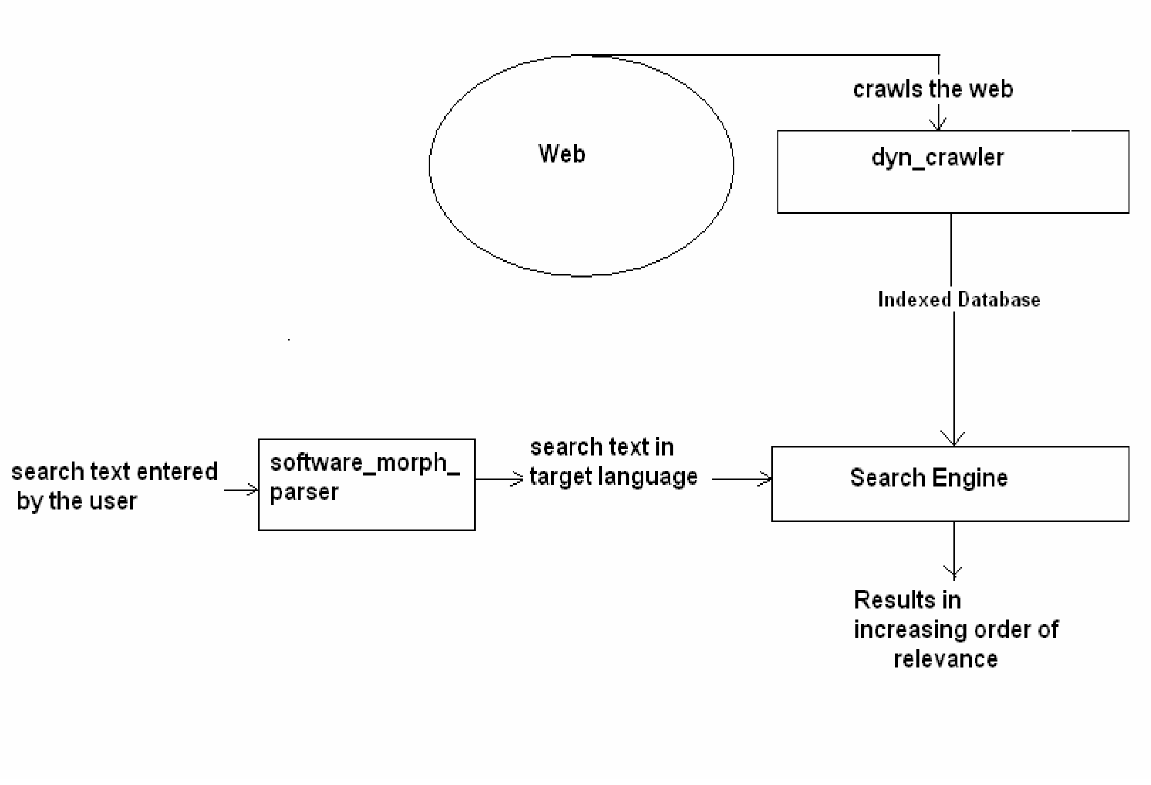

Web
crawls the web
dyn_crawler
Indexed Database
search text entered by the user
software_morph_ parser
search text in target language
Search Engine
Results in increasing order of relevance

Figure 2: Block Diagram of LIIRW

2.1.2 Translator

The translator module uses the principles of morphological parsing and feature structure grammar for parsing the individual words as well as the phrases entered by the user as the search text. It uses two approaches in parallel for analyzing and translating the search text. One approach translates the individual words and the other one is used for translating phrases. Highest priority is given to the results obtained when the whole keyword is taken as the phrase. The next priority is given to the results containing one or more combination of the words in the search text.

In the first approach the search text is segmented into individual words. The individual words are identified as being demarcated by space(s). Each of these individual words is then checked for their grammatical form i.e. for the presence of any prefix and/or suffix and for the part of speech to which they belong. For example suppose the search text entered by the user is "*computerization of Indian villages".* The words are segregated as "*computerization", "Indian", "of*", "*villages".* The word *computerization* is further parsed as computer + ize + tion (noun + ize + ation, noun to verb to noun). Similarly the word "*villages"* is parsed as village + es(noun +es, noun + PLURAL). The parsing of the word *Indian* yields Indian (proper noun). The words like *of, in* etc are used only for the phrase searching and are not considered for translation as independent words. Each of the root words so obtained (in the given examples the root words are *computer, Indian, and villages)* are then searched in the translator database for their meanings in the target language. A separate database is maintained for the meanings of root words of each combination of languages (i.e. for Hindi to English, Hindi to Tamil and English to Tamil), as a part of the translator database. These translated root words are then taken as keywords for search in the indexed database of the target language.

In the second approach which always works in parallel with the first one, the whole phrase is translated into the target language using feature structured grammar and then the translated text is used as keyword as a whole. Highest priority is given to the results obtained by using the whole phrase as the keyword.

2.1.3 Crawler and Indexer: dyn_crawler

The crawler crawls the web and pages are indexed along with their corresponding keywords as given in their meta tag. This indexing is a dynamic process and goes on continuously. The crawler and indexer form the heart of the search engine. The search engine takes as input the result of the *software_morph_parser* and searches the indexed database. The results are prioritized according to the search text and the web pages are set relevant to the whole phrase or to the one or more combinations of the words.

The various databases used in LIIRW are as described below.

1. **The translator database**

The translator database is used for storing the translations of root words in different languages. A database for all the combinations is maintained which includes English to Hindi and vice-versa, English to Tamil and vice-versa and Hindi to Tamil and vice-versa.

2. **The indexed database**

The indexed database contains the index of web pages along with their corresponding keywords of relevance. Thus these databases are efficiently organized and maintained. These databases play a major role in the information retrieval. Since, the goal is to search web pages of specific languages, context focused crawlers are designed. The crawler instead classifying the text does the language focused crawling. The language identification module returns the name of the language for a given web page. This module is aware of all the proprietary encodings and also uses a bag of words to recognize unknown encodings from meta-tag information that might be found in an HTML page. In many cases, web pages contain more than one language, especially one of the languages being English. This happens since many of the website organizes certain information such as menu items, or disclaimers and other such formatting information in English. In some websites such as blogs or forums majority of the content might be English, with Indian language e-content being a major focus. The language identifier module returns a language only if the number of words in a web page is above a given threshold or rank is of appropriate value.

3. Working Principle of LIIRW

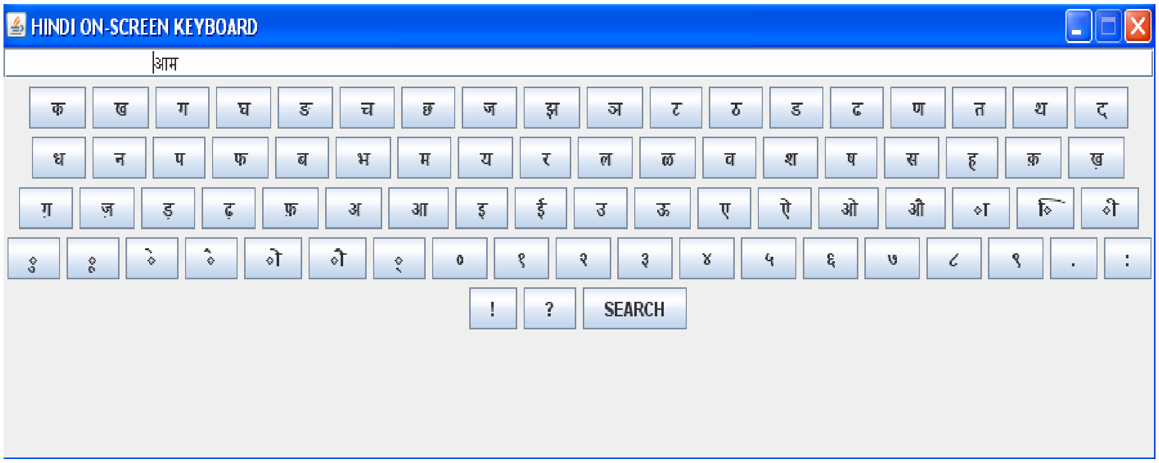
An information access process assumes an interaction cycle consisting of query specification, receipt and examination of retrieval results, and then either stopping or reformulating the query and repeating the process until a perfect result set is found. The steps are as defined below.

1. Start with an information need.
2. Select a system and collections to search on.
3. Formulate a query.
4. Send the query to the system, translate the query and search.
5. Receive the results in the form of information items.
6. Scan, evaluate, and interpret the results.
7. Either stop, or,
8. Reformulate the query and go to step 4.

Users scan information structure, be it titles, thesaurus terms, hyperlinks, category labels, or the results of clustering, and then either select a displayed item for some purpose (to read in detail, to use as input to a query, to navigate to a new page of information) or formulate a search text (either by recalling potential words or by selecting categories or suggested terms that have been scanned). In both cases, a new set of information is then made viewable for scanning. Queries tend to produce new, ad hoc collections of information that have not been gathered together before, whereas selection retrieves information that has already been composed or organized. Navigation refers to following a chain of links, switching from one view to another, toward some goal, in a sequence of scan and select operations. Browsing refers to the casual, mainly undirected exploration of information structures, and is usually done in tandem with selection, although queries can also be used to create sub collections to browse through. An important aspect of the interaction process is that the output of one action should be easily used as the input to the next.

4. Analysis of LIIRW

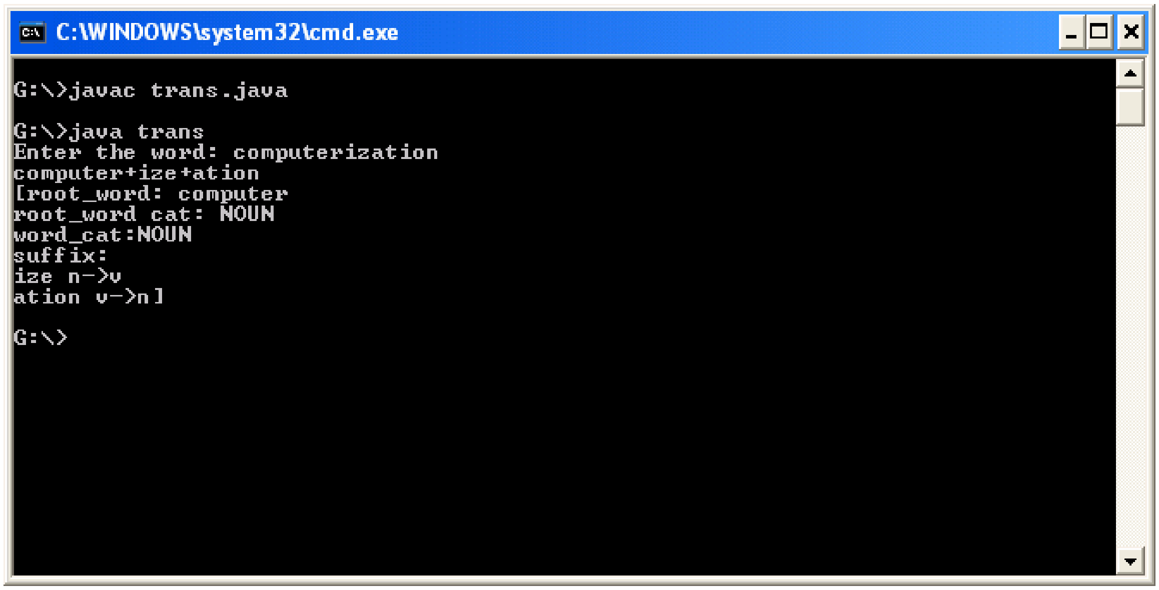
LIIRW is a very efficient technique. It is designed as above and is implemented in LAMP (Linux – Apache-MySQL-Php) architecture. First the input search text is obtained from the keyboard as shown below in figure 3 and it is processed and the results are shown as figure 4 .


HINDI ON-SCREEN KEYBOARD
क ख ग घ ङ च छ ज झ ञ ट ठ ड ढ ण त थ द
ध न प फ ब भ म य र ल ळ व श ष स ह क़ ख़
ग़ ज़ ड़ ढ़ फ़ अ आ इ ई उ ऊ ए ऐ ओ औ ा ि ी
ु ू े ै ो ौ ् ० १ २ ३ ४ ५ ६ ७ ८ ९ . :
! ? SEARCH

आम

Figure 3 Snapshot of Hindi keyboard

If the user desires to enter the search text in a language other than English, he will be provided with a similar onscreen keyboard in the required language. The entered text will be taken as input and will be broken into individual words. Each word will then be analyzed by the parser module of the software_morph_parser and an output will be produced as shown below. As a sample we have taken the analysis of the word ***computerize.***



```
G:\>javac trans.java

G:\>java trans
Enter the word: computerization
computer+ize+ation
[root_word: computer
root_word cat: NOUN
word_cat:NOUN
suffix:
ize n->v
ation v->n]

G:\>
```

Figure 4 : Morphological analysis of keyword

5. Conclusion

Recently internet is populated with rich content. The digitization of documents is the key aspect and it is attempted to do it in the regional languages. Millions of web pages are available in regional languages and the web has exponentially grown as a multilingual domain. This paper is an attempt to make such a multilingual interaction possible by allowing the user to query in a different language and retrieve information in another language or a combination of languages. This improves the knowledge domain of the users and makes the Internet accessible by any novice user irrespective of their linguistic skills.

6. References

1. Thorsten Brants , "Natural Language Processing in Information Retrieval",Google Inc, 2003
2. Prasad Pingali, Jagadeesh jagarlamudi, Vasudeva Varma, " WebKhoj: Indian language IR from Multiple Character Encodings", ACM 1-59593-323, Edinberg, Scotland, 2006
3. Ramdasi Nagnath Ramchandra1and Patil Suresh K," Nature of Complexities in a Document: Content Digitization Aspects with Special Reference to Indian Heritage Knowledge Domain**",**2006
4. Paul Buitelaar, Klaus Netter, Feiyu Xu," Integrating Different Strategies for Cross-Language Information Retrieval in the MIETTA Project",Germany.
5. Tim Gollins ,Mark Sanderson, "Improving Cross Language Retrieval with Triangulated Translation", Sigir'01,2001

6. Natalia V. Loukachevitch, Boris V. Dobrov,” Cross-Language Information Retrieval Based on Multilingual Thesauri Specially Created for Automatic Text Processing”, Russia,2002
7. Martin Volk, Spela Vintar, Paul Buitelaar,” Ontologies in Cross-Language Information Retrieval”, Switzerland, 2002
8. Prasenjit Majumder, Mandar Mitra, Kalyan Datta, ”Multilingual Information Access: an Indian Language Perspective**”,** SIGIR, 2006
9. Ricardo Baeza-Yates,_Berthier Ribeiro-Neto, ”Modern Information Retrieval”, AWL,1999
10. James Allen, “Natural Language Understanding”, Pearson Education, Inc. 1995