ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 12, December 2014

Refinement of search results using cross lingual reference technique

Monika Sharma[1], **Sudha Morwal**[2]

Department of Computer Science, Banasthali University Jaipur, India[1,2]

Abstract: In this paper we present a fully implemented system which supports cross-lingual search of World Wide Web. Cross-language information retrieval (CLIR) is a subfield of information retrieval (IR) that deals with the retrieval of the information in a language different from the language of the user's query. Google is a powerful tool to search information on the World Wide Web. Google is a keyword based search engine. Google is a widely used search engine that uses the keyword matching to find the documents that are important & relevant to user's query. It will retrieve the documents as a search result that contains query terms. Search words have multiple meanings or appear in multiple contexts or in multiple languages. Semantically similar pages that are desirable by the user are often not retrieved .CLIR is helpful for refining the search result of Google. It allows the search engine to not only retrieve the documents containing the query terms but also the documents containing the semantic equivalent of query terms in the other languages. The system, which is named as "Refinement of Google Search Results by Cross Lingual Reference technique" will improve the quality of the search results of Google by providing the multilingual documents in response to a user query written in single language. The goal of this system is to improve the search result of Google by using the CLIR approach for search. This System will avoid the overhead involved in creating multiple requests in case of different languages document retrieval .This System can be used for NLP Linguistics, Govt Organizations such as Intelligence dept, Immigration dept.,Multilingual population regions etc.

Keywords*:* Information Retrieval, World Wide Web, CLIR

I. INTRODUCTION

"Refinement of Google Search results by Cross Lingual Reference Technique" helps to improve the search results of the Google by using Cross Language Information Retrieval. This System aims to perform semantic search that uses semantics or the equivalent word of the query terms in the other languages such as Hindi and Bengali to produce the highly relevant search results. Semantically equivalent pages in the other languages to the search query are also retrieved as a search result. It enables the user to enter the query in one language and then retrieve all the documents containing the equivalent words in the other languages . This system allows the users to retrieve documents in different languages rather than just their original query language.

A. What is Cross Lingual Information Retrieval (CLIR)

CLIR is the task of issuing a query in one language and retrieving the set of relevant documents in other languages. The aim of CLIR is to provide the benefit to the user in finding and assessing information without being limited by language barriers. Cross-lingual information retrieval has become more important in recent years. CLIR enables the users to retrieve documents of the other languages than their original query language.Potential users for CLIR are users who find it difficult to formulate a query in their non-native language and users who are multilingual and want to save time by entering a query in one language instead of entering the query in all languages in which they want to search, Cardeñosa.J.et al, [1].

Cross-Language Information Retrieval allows the users to search and read pages in the language different from the language of the search terms Swapna.N.et al, [2]. Cross-language information retrieval is a type of information retrieval in which the language of the query is different from the language of the documents retrieved as a search result. In CLIR system a user is not limited to his own native language, so the user can make his query in his native language but the system returns set of documents in another languages. CLIR involves the retrieval of documents in a language other than the query language. Since the language of query and the documents needs to be translated in CLIR. CLIR system simplifies the search process for multilingual users and enables those who know only one language to provide queries in their language and then get help from translators for using other languages documents, Alizadeh.H.et al, [3].

B. How Google Search works

Google is a very popular search engine on the World Wide Web, handling more than billion searches every day [6]. Google uses the keyword matching to find the documents that are most relevant to user's query. Google performs keyword matching by checking the keywords in database and retrieve the documents that contain keyword but it does not retrieve the equivalent words of the query terms in other languages. For e.g. If "Prime Minister" is a keyword to search, then the search result will contain the documents having "Prime Minister" but search result will not retrieve the documents having equivalent words in the

other languages like प्रधानमंत्री (Hindi) and প্রধানমন্ত্রী (Bengali).Multiple language documents containing the equivalent word to the search query
are not retrieved.

Google checks each of keyword in its database and then retrieve the list of documents that contains keyword. Google runs on a distributed network of the multiple computers and it can perform fast parallel processing on multiple computers. Parallel processing is a method of computation in which many calculations can be performed simultaneously, significantly speeding up data processing [7]. Google has three distinct parts:

1.) Googlebot i.e. a web crawler which finds and fetches web pages from the World Wide Web.
2.) The indexer sorts words on every page alphabetically and stores the index of words in a Google's index database.
3.) The query processor compares the search terms to the index of words and then provides the set of documents that are most relevant to the search terms.

The Keyword based search does not provide the relevant information to the User. There is a requirement of "Cross Language Information Retrieval" that enables the user to search the query terms in multiple languages. Search result of the Google can be refined by using CLIR.

C. Limitations of Google Search

Google is keyword based search engine and it uses the keyword matching to retrieve the documents that contain the keywords in query. Google will retrieve the document that matches with the search query and it does not retrieve the documents containing the semantic equivalent to the query terms in the other languages. This will give irrelevant search results to the user query. Semantically similar pages to the user's query in other languages will not be retrieved. This Limitation can be overcome by using CLIR technique. Using Cross Language Information Retrieval allows the user to make query only in one language and retrieve the list of documents having equivalent words in the other languages. Search result of the Google has been refined by using CLIR with search engine.

D. CLIR in the Refinement of Google Search Result

CLIR is helpful for refining the search result of Google by providing the search result beyond the linguistic barriers. CLIR does not limit the search result to keyword matching but it will also provide the multiple language documents having equivalent word to the search query as a search result. The main focus is to perform semantic and intelligent search that improves the accuracy of the search results by understanding searcher intent and the contextual meaning of terms to generate more relevant results. Semantically similar pages to the query in the other languages will also be retrieved. When user searches a query then the multiple languages documents containing the same meaning of the query will also be retrieved as a search result.

II. QUERY TRANSLATION APPROACH FOR CLIR

Search strategies are continuously improving their techniques to provide more relevant, accurate information for a given query. The Search can be refined by providing the intelligent search in the unrestricted domain. Multilingual information search becomes important due to increasing the amount of online information available in non-English languages and multiple language document collections. This can be achieved by Query translation. Query translation using CLIR became the widely used technique to access documents of the different languages from the language of query [7]. For translating the query, we can use an online translation i.e. Google Translate, train a Statistical Machine Translation system using parallel corpora , employ Machine Readable Dictionaries to translate query terms or use of large scale multilingual information sources like Wikipedia . CLIR using the Google Translate achieve the accuracy of the search results up to 90 %, Herbert.B.et al [8].

In this system we use Google Translate query translation approach. Translation can be applied to the query terms online. Online query translation can be achieved by using one of the Google Translate API which will convert the query into the other languages. Online query translation will help the user to translate his query in the other languages [6].

III.PROPOSED SYSTEM

A system has been designed using the python programming language for refining search result of Google.The purpose of this System is to deliver the meaningful or knowledgeful information queried by a user rather than have a user sort through a list of related keyword results. Semantic search is the closest a search engine can get to "reading the minds" of internet users. The goal is to access semantically similar pages to search query. Documents relevant to the User's query in the other languages will also be retrieved as a search result. Cross-Language information retrieval enables users to enter queries in languages in which they are fluent and uses the language translation methods to retrieve the documents written in the other languages as a search result. This System will enable the user to enter the query in one language and retrieve the set of documents of multiple languages as a search result. This type of system is useful in refining the search result of Google, it provides the more relevant and meaningful information in the multiple languages to the user. Instead of providing search result in one language it will provide the equivalent words in all other languages. If user searches a keyword in English language then the set of documents containing equivalent words in the other languages such as Hindi and Bengali language.

A System has been designed to improve search result of Google using Cross Lingual Cross Language Information Retrieval. The Proposed System allows the user to enter query in one language and then provide the information

about equivalent words in the other languages as shown in Fig.1.
Search result can be refined in the three phases.

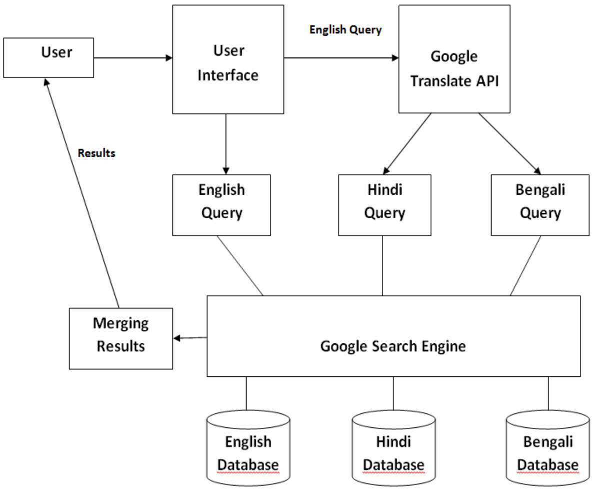
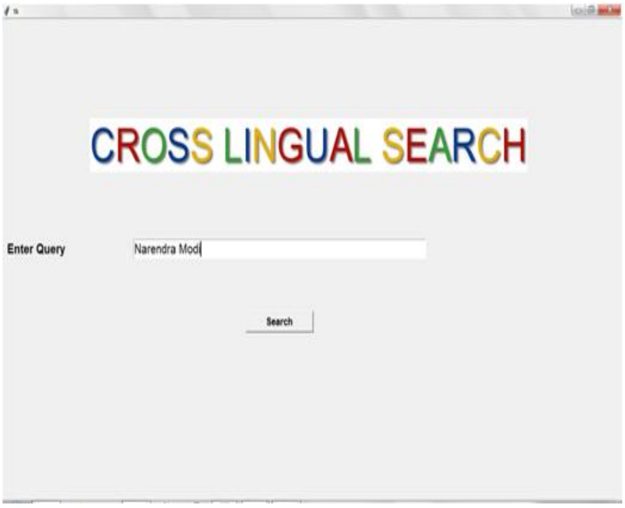


Fig.1. Block diagram of the proposed system

A. Query generation

User enters the query through the user Interface in English Language. Query submitted by the user is then passed to the Google search engine.User Interface can be designed in python programming using Tkinter tool.
For Eg. User search a Query "Narendra Modi" as shown in Fig. 2


CROSS LINGUAL SEARCH

Enter Query Narendra Modi

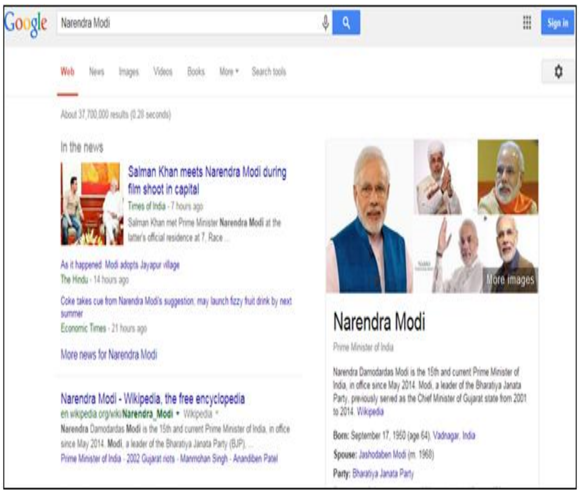
Fig. 2. User Interface

B. Query Translation

Query Translation can be applied to the query entered by user. This System involves translation of query written in English language to other Indian languages so that the semantically similar pages in all the languages will be retrieved as search result. The query translation can be achieved by using the Google Translate API, the query entered by the user in English language can be translated to the other languages like Hindi and Bengali. Google translate API will convert the query in one language into the equivalent words to the query in other languages.

C. Refined Search Result

Refinement of the search result of the Google will provide the intelligence to search. By passing the translated query to Google ,the search result of the Google can be improved by providing the most relevant information to the user's query. The refined search result will contain the documents of the multiple languages having similar word to the query entered by User. Fig.3 shows the Search Result of the Google that retrieve the search results in English language. Google checks its database for keyword to search like "Narendra Modi" and then it retrieves the set of all documents containing the Narendra Modi.



Narendra Modi

About 37,700,000 results (0.20 seconds)

In the news

Salman Khan meets Narendra Modi during film shoot in capital
Times of India - 7 hours ago
Salman Khan met Prime Minister **Narendra Modi** at the latter's official residence at 7, Race ...

As it happened: Modi adopts Jayapur village
The Hindu - 14 hours ago

Coke takes cue from Narendra Modi's suggestion, may launch fizzy fruit drink by next summer
Economic Times - 21 hours ago

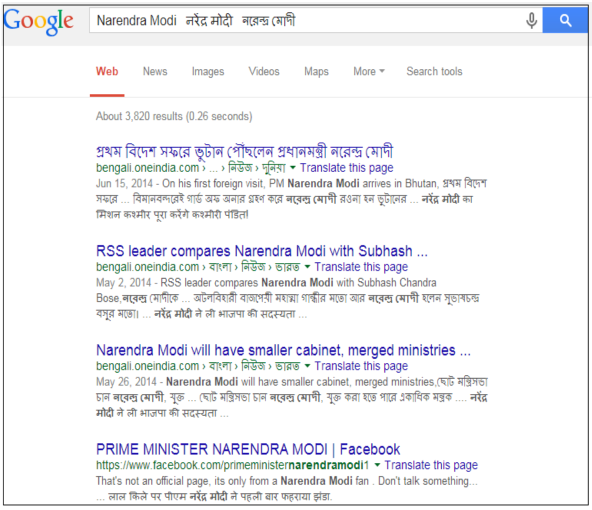
More news for Narendra Modi

Narendra Modi - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Narendra_Modi** ▾ Wikipedia ▾
Narendra Damodardas **Modi** is the 15th and current Prime Minister of India, in office since May 2014. **Modi**, a leader of the Bharatiya Janata Party (BJP) ...
Prime Minister of India - 2002 Gujarat riots - Manmohan Singh - Anandiben Patel

Narendra Modi
Prime Minister of India
Narendra Damodardas Modi is the 15th and current Prime Minister of India, in office since May 2014. Modi, a leader of the Bharatiya Janata Party, previously served as the Chief Minister of Gujarat state from 2001 to 2014. Wikipedia
Born: September 17, 1950 (age 64), Vadnagar, India
Spouse: Jashodaben Modi (m. 1968)
Party: Bharatiya Janata Party

Fig. 3. Search Result of the Google

Refined Search result will contain the documents having "Narendra Modi" and also the documents in the other languages having the equivalent words such as नरेंद्र मोदी (Hindi), নরেন্দ্র মোদি (Bengali) as shown in Fig.4



Narendra Modi नरेंद्र मोदी নরেন্দ্র মোদী

About 3,820 results (0.26 seconds)

প্রথম বিদেশ সফরে ভুটান পৌঁছলেন প্রধানমন্ত্রী নরেন্দ্র মোদী
bengali.oneindia.com › ... › নিউজ › দুনিয়া ▾ Translate this page
Jun 15, 2014 - On his first foreign visit, PM **Narendra Modi** arrives in Bhutan, প্রথম বিদেশ সফরে ... বিমানবন্দরেই গার্ড অফ অনার গ্রহণ করে **নরেন্দ্র মোদী** রওনা হন ভুটানের ... **नरेंद्र मोदी** का मिशन कश्मीर पूरा करेंगे कश्मीरी पंडित!

RSS leader compares Narendra Modi with Subhash ...
bengali.oneindia.com › বাংলা › নিউজ › ভারত ▾ Translate this page
May 2, 2014 - RSS leader compares **Narendra Modi** with Subhash Chandra Bose,**নরেন্দ্র** মোদীকে ... অটলবিহারী বাজপেয়ী মহাত্মা গান্ধীর মতো আর **নরেন্দ্র মোদী** হলেন সুভাষচন্দ্র বসুর মতো। ... **नरेंद्र मोदी** ने ली भाजपा की सदस्यता ...

Narendra Modi will have smaller cabinet, merged ministries ...
bengali.oneindia.com › বাংলা › নিউজ › ভারত ▾ Translate this page
May 26, 2014 - **Narendra Modi** will have smaller cabinet, merged ministries,ছোট মন্ত্রিসভা চান **নরেন্দ্র মোদী**, যুক্ত ... ছোট মন্ত্রিসভা চান **নরেন্দ্র মোদী**, যুক্ত করা হতে পারে একাধিক মন্ত্রক .... **नरेंद्र मोदी** ने ली भाजपा की सदस्यता ...

PRIME MINISTER NARENDRA MODI | Facebook
https://www.facebook.com/primeminister**narendramodi**1 ▾ Translate this page
That's not an official page, its only from a **Narendra Modi** fan . Don't talk something... ... लाल किले पर पीएम **नरेंद्र मोदी** ने पहली बार फहराया झंडा.

Fig. 4. Refined Search Result of the Google

IV. APPLICATIONS

1.This System can be helpful for immigration department. For eg. Immegration department interact with thousands of the Indian native Language speakers which are not able to understand English Languages ,So this project will be beneficial for all those peoples by providing the search result in their native languages.

2.This System can be used for multilingual population regions so that the peoples having different native languages retrieve documents in their native languages.

3.This system can also be used for intelligence departments.

V. BENEFITS

1. This System will improve the quality and relevancy of the search results of the Google.
2. Semantically equivalent pages in other languages to the search query are also retrieved as a search result.

VI. LIMITATIONS

"Refinement of Google Search Result using Cross Lingual Reference technique" works well for retrieving documents of multiple languages in a single query. However this project having several challenges that are given below

- This project uses Google Translate API for language translation hence it only support the languages provided by the Google API. If we want to add a new language that is not supported by Google then it is not possible.

- This System supports only the two or three language combination to pass as a query to a Google Search Engine.

VII. FUTURE WORK

This System will support the information retrieval related to limited languages i.e. three languages .In future this can be extended to multiple languages so as to retrieve the documents of multiple languages as a search result of a query in single languages.

VIII. CONCLUSION

Semantic based search can include the concept or idea expressed by the user in his query and can thus provide more accurate results than the traditional keyword search. CLIR refines the Google search result by retrieving the semantically similar pages of multiple languages. This system allows the search engine to not only retrieve the documents having query terms but also documents having semantic equivalent of the query in other languages. This System allows the search engine to provide search result more relevant and more closer to the User query. Refinement of search result of the Google using CLIR will make the Google search intelligent and achieve semantic search criteria.

REFERENCES

[1] J. Cardeñosa, C Gallardo, Adriana Toni," Multilingual Cross Language Information Retrieval A new approach".
[2] N.Swapna, N.Hareen Kumar, B.Padmaja Rani,"Information Retrieval in Indian Languages: A case study on Cross-Lingual and Multi-Lingual", International Journal of Research in Computer and Communication technology ,ISSN 2278-5841,Vol 1,Issue,September 2012
[3] H. Alizadeh, R. Fattahi "Applying Natural Language Processing Techniques for Effective Persian- English Cross-Language Information Retrieval", International Journal of Information Science and Management
[4] M.V Reddy, Dr. M. Hanumanthappa ,Manish Kumar,"Cross Lingual Information Retrieval using Search Engine and Data Mining",ACEEE Int. J. on Information Technology, Vol. 01,No. 02,Sep 2011.
[5] F. Ahmed,A. Nurnberger ,"Literature Review of Interactive Cross Language Information Retrieval Tools",The International Arab Journal of Information Technology ,Vol. 9,No. 5,2012.
[6] http://en.wikipedia.org/wiki/Google_Search
[7] http://www.googleguide.com/google_works.html
[8] B.Herbert,G. Szarvas,I Gurevych "Combining Query Translation Techniques To Improve Cross-Language Information Retrieval",2011
[9] http://en.wikipedia.org/wiki/Python_(programming_language)
[10] http:// pythonhosted.org/goslate/
[11] http://effbot.org/tkinterbook/tkinter-index.htm
[12] https://docs.python.org/2/library/tkinter.html