An Experimental Analysis on the Influence of English on Hindi Language Information Retrieval

Kumar Sourabh
Department of Computer Science and IT,
University of Jammu
Jammu, J&K 180001. INDIA

Vibhakar Mansotra
Department of Computer Science and IT,
University of Jammu
Jammu, J&K 180001. INDIA

ABSTRACT

With the internet growing at an exponential rate the web is increasingly hosting web pages in different languages. Hence it becomes essential for the search engines to be able to search information stored in a specific language. The native users also tend to look for any information on web nowadays. This leads to the need of effective search mechanisms to fulfill native user's needs and provide them information in their native languages. The most preferred language for web information access is English Language but the growth and popularity of internet in countries like China, India, Russia Middle East and south Asia has lead to the importance of IR in Languages other than English. There has been a great increase in Indian content particularly Hindi on the Internet but retrieval of Hindi information is a complex task because of the various reasons such as Phonetic nature of Hindi language, morphology, word synonyms, ambiguous words etc. Influence of regional and foreign languages especially English on Hindi has been observed as a major factor hindering the performance of IR in Hindi language. The objective of the paper and present research is to study the effect of English on Hindi IR. This paper covers the comprehensive analysis of the influence of English Language on Hindi IR and its effect on performance of search engines supporting Hindi Language.

Keywords: Search engines, morphology, synonyms, precision, Hindi, Information retrieval (IR).

1. INTRODUCTION

The growth of web as a source of unlimited information has lead to a revolution in the area of IR. The internet has become a popular source of information not for the technocrats only but a common man too. The most preferred language for web information access is English Language but the growth and popularity of internet in countries like China, India, Russia Middle East and south Asia has lead to the importance of IR in Languages other than English. The increasing growth of non English content on the web needs a proper mechanism to be accessed so that such content becomes noticeable and available whenever wherever necessary. Looking from a global perspective India is the world's third largest user base behind China and the United States. There has been a great increase in Indian content on the Internet and this has made the companies like Google, Yahoo, and MSN to offer contents in Indian languages particularly in Hindi which is the official Indian language and spoken worldwide.

Even though there is an availability of huge content in Hindi and other Indian regional languages on web, a very few search engines like Google, Raftaar, Webkhoj, Guruji are available for retrieval of such information but performance of these search engines is not up to the mark as they are not able to provide quality information particularly in Hindi Language. This is because Hindi Language has a very complex and varied structure and vast language diversity. About 22 languages are spoken across the country (India) and Hindi being the National and official language gets influenced by the regional languages.

Major factors that affect the performance of search engines for Hindi searching on web are Phonetic nature of Hindi language, morphology, word synonyms, ambiguous words and Influence of regional and foreign languages especially English on Hindi. The influence of English language on Hindi has been observed as one of the major factors that affect the performance of Hindi IR which has been studied experimentally and discussed in this paper.

2. INFORMATION RETRIEVAL AND HINDI LANGUAGE

Hindi language is spoken by the major population of India. About 5% of population understands English as their second language. Hindi is spoken about 30% of the population. A wide variety of Hindi Data and Literature is now available on web. The number of users who want the information in Hindi language is increasing [1]. Hindi is the language of dozens of major newspapers, magazines, radio and television stations and of other media. Major Hindi newspapers and TV channels have their websites in Hindi which are used by wide section of society. [2]. A recent survey by a Delhi based research organization - Juxt Consult - says that 44 % of existing Internet users in India prefer Hindi over English, if made available. Similarly 25% existing Internet users prefer other regional languages. Many big companies like Google, Yahoo and Sify are also taking big steps in Hindi and other regional languages [3]. Hindi IR is still in a very nascent stage. As mentioned above problems like Phonetic nature of Hindi Language, morphology, word synonyms and ambiguous words affects the performance of the search engines in Hindi language information retrieval. These factors are briefly discussed as:

- ***Morphological Factors***: Morphology is the branch of linguistics that studies patterns of word formation within and across languages, and attempts to formulate rules that model the knowledge of the speakers of those languages. [4]
- ***Phonetic nature of Hindi Language***: Many different languages are spoken in India, each language being the mother tongue of tens of millions of people. While the languages and scripts are distinct from each other, the grammar and the alphabet are similar to a large extent. One common feature is that all the Indian languages are phonetic in nature [5]. For example; Following are the possible spelling variations for the Hindi word अंग्रेजी (angrējī): (means English) अँग्रेजी, अंगरेजी, अन्ग्रेजी, अँगरेजी, अंग्रेजी, अंग्रेज़
- ***Words Synonyms***: India has rich diversity in languages, culture, customs and religions. But, the language structure and variation in dialects is making hindrances in the advantages of Information retrieval revolution in India. For example: we know God is named as "भगवान" in Hindi but we can also call "भगवान"as "प्रभु"इश्वर" or "देवता" and more. It is difficult to decide that which one is to choose?
- ***Ambiguous Words***: Many words are polysemous in nature. Finding the correct sense of the words in a given context is an intricate task. One word has more than one meaning and meaning of word is depends on context of sentence. Example कर (Tax) having synonyms ब्याज, शुल्क, सूद, महसूल, टैक्स in one context and in another context कर (Hand or arms) हस्त, बाँह, आच, शबर and कर (to do) करना in another context.

An experimental analysis has been done on these parameters by Dr. S.K Dwivedi and Rajesh Kr. Goutam. [2] Their experimental study reveals that not only the quantity but the quality of the results is improved by the inclusion of these parameters in Hindi web searching.

Another important factor that affects the performance of Hindi IR on web is "Influence of English language on Hindi" which is the objective of the present experimental study.

3. INFLUENCE OF ENGLISH ON HINDI INFORMATION RETRIEVAL

The English language has influenced Indian languages in many ways: it affected the pronunciation of Hindi words. So many English words have been localized in India. Some of the words appear as if they were native Hindi words. Indians, sometimes, are unable to get the equivalent word for that of the English. For instance, the words, such as, road, bus, pen, television, radio, please, rail, email, password, insurance, internet, director, department etc are used even by the uneducated Indians without being aware of the language of those words. Most of the Indians use these words in English than in their native language. [7]

English language has its influence over Hindi not only in speaking but in writing too. When we talk about especially Hindi literature on web it becomes more evident. Influence of English on Hindi language has been observed as one of the very important parameters for Hindi Information retrieval which is more clearly explained in the example as.

Example: In English the word *exercise* is written in Hindi as (एकसरसाइज़). Due to the phonetic nature of Hindi language the word *exercise* (एकसरसाइज़) has following phonetic variations एकसरसाइज़, एक्सरसाइज , एक्सेरसाइज़ , एक्सरसाइज in Hindi. As per various phonetic variations mentioned above in the example a variety of popular keywords and queries have been tested for experiments from various domains.

In order to carry out the study software has been developed by the authors. The tool has been designed and developed as an interface to the search engines with a motive of generating possible variations of a single/multiple word query. These variations are generated based on the various factors like morphology, phonetics, word synonyms and Hindi equivalents of English words. A large scale database has been developed as a backend of the interface.

4. EXPERIMENTAL RESULTS

In this section experimental results and analysis on the effect of English words written in Hindi for IR process has been done. In the following table a sample set of common and popular English keywords along with their phonetic variants written in Hindi can be seen.

Fig. 1

From the above table it can be observed that search engine does return documents for single keyword query, documents for all phonetic variants of the keywords are also returned which are huge in number. From the table above it can easily be concluded that people have their own way of representing the Hindi words and no standard is followed for storing Hindi data on web. Also the documents are retrieved for every phonetically variant English Keyword written in Hindi script.

In the above table the column with bold Hindi entries shows the keywords which are obtained by using Google transliteration tool. The table shows that transliteration does not provide correct Hindi word in most of the cases. For example the correct transliteration for word *University* should be यूनिवर्सिटी whereas Google transliteration provides the word उनिवेर्सित्य which is completely wrong. It is clearly evident from the table above that 1540 documents have been retrieved for the wrong keyword उनिवेर्सित्य (University) and the same follows for other single word queries Insurance इन्सुरांस, Parliamentपर्लिअमेंट,Corruptionकोर्रुप्तिओन,Policy पोलिस्य, Specialistस्पेसिअलिस्ट. It can be concluded that Hindi website developers make use of unchecked and non standard transliteration which makes the Hindi IR process a difficult task.

In the next experiment Multiword Hindi queries are selected to test the effect of English influence on Hindi IR on precision and quantity of documents retrieved. The Hindi query is transformed into it variants by the software by replacing the Hindi keywords with English keywords written in Hindi without changing the meaning of the query. The queries are converted into two levels. In first level one Hindi word is replaced by its English equivalent and in second level more than one word/s is replaced by their English equivalent words without changing the meaning of the original Hindi query. Example:

An English query “***Foreign investment in India***” can be written in Hindi as “विदेशी निवेश भारत में ” where Hindi keyword विदेशी means “Foreign” “फारेन” and निवेश means “investment” “इन्वेस्टमेंट”. The query for the two levels is transformed as:

फारेन निवेश भारत में (**Foreign** nivesh bharat mein).

फारेन इन्वेस्टमेंट भारत में (**Foreign investment** Bharat mein).

Therefore the original Hindi query “विदेशी निवेश भारत में ” “videshi nivesh Bharat mein” supplied by the user is transformed into two equivalent senses containing a mixture of both English and Hindi language where meaning of the query remains same. From the sample set of one hundred queries some randomly selected queries are presented below in table.

Fig. 2 and Fig. 3

From the above table it is evident that documents are returned for original as well as transformed Hindi query and the quantity of the retrieved documents is quite considerable. In case of search engines the quality of results is more important than the quantity therefore a table and graphs are presented below for the analysis of precision values. Three popular search engines namely Google, Bing and Alta Vista are used for retrieving web results.

Fig. 4 and Fig 5

From the above table it can be clearly seen that Hindi data of similar nature can be mined out against Hindi queries by transforming them into their variants by including English keywords written in Hindi. The transformation of queries resulted in an increase of retrieved data. The relevance of the retrieved data can also be seen in the precision column. For every Hindi query and its transformed variation/s the degree of relevance of documents is very close or equal or improved e.g. for the Hindi query स्वास्थ और रक्तदान; 9 out of first 10 documents are relevant and for transformed queries which are of similar nature हेल्थ और रक्त दान and हेल्थ और ब्लड_डोनेशन; 9 of the first 10 documents are relevant and the same repeats for rest of the Hindi queries as shown in the table above. Without transformation of Hindi queries the user may miss the chance of retrieving the relevant information as the Hindi user may not be aware of the presence of such information on web and may be unable to formulate the variation query based on the factor of English influence. From the above table it can be concluded that, English influenced Hindi information is present and is increasing day by day on web. By the inclusion of the English keywords in Hindi script in the form of query, the scope of searching in Hindi and getting relevant information can be increased.

5. DISCUSSION AND CONCLUSIONS

In the above section we have observed that English Language has its impact on Hindi Information retrieval. A wide section of society use Hindi as its first language to search information on web. Although wide variety of Hindi literature is available on the web but the retrieval of relevant information in Hindi is a complex task because of the aforesaid problems. The process of Hindi IR becomes more difficult because of the structure of Hindi Language. Generally people do not follow the actual Hindi writing standard which widens the gap between Hindi web data and users.

It has been concluded that relevant information can be mined out by transforming the Hindi queries. Search engines neither make transformations of the query nor find keyword equivalents. Because they may have the performance and throughput problems if parameters like Hindi Phonetics, synonyms and English equivalent Hindi keywords are implemented at root level. However this problem can be solved at interface level. Therefore to lessen the efforts of a Hindi user to search such information a software has been developed (a brief description has been mentioned above in section 3) which acts like an interface between user and search engines. With the help of this tool user can widen the scope of search on web in Hindi language.

There is general feeling among few researchers that Hindi spelling checker like Google’s “Did you mean” tool can be used at the interface level to overcome the issues related to the phonetic nature of Hindi language. But our observation suggests that inclusion of such tool may not help to solve the problem as there is already a huge amount of improperly written Hindi data present on web which will remain unexplored by including the online spell checkers. However the inclusion such a tool may be incorporated as an option.

6. REFERENCES

[1] S.K.Dwivedi And Parul Rastogi Babasaheb Bhimrao Ambedkar University“*An Entropy Based Method For Removing Web Query Ambiguity In Hindi Language*” Journal Of Computer Science 4 (9): 762-767, 2008 ISSN 1549-3636 © 2008 Science Publications

[2] S.K. Dwivedi And Parul Rastogi Rajesh Kr. Gautam “*Impact Of Language Morphologies On Search Engine Performance For Hindi And English Language”* (IJCSIS) International Journal Of Computer Science And Information Security, Vol. 8, No. 3, June 2010.

[3] Ranjan Srivastava, Chief Of Bureau, Prabhat Khabar, (Friday, April 28, 2006) “*The Future Of Hindi On The Internet*” Http://Www.Raftaar.In/Thehoot.Htm.

[4] Rajeev Rathor Master Of Engineering *Thesis Thapar University*.” Patiala Morphological POS Tagger For Hindi Language” URL: -Http://Dspace.Thapar.Edu:8080/Dspace/Bitstream/10266/554/1/Rajeev+Thesis+Report.Pdf

[5] GANAPATHIRAJU, Madhavi,Balakrishnanmini, BALAKRISHNAN N. REDDY Raj “*Om: One Tool For Many (Indian) Languages*” Journal Of Zhejiang University SCIENCE ISSN 1009-3095

[6] Influence Of English On Indian Languages {Article Source} Http://Www.Indiachatforum.Net/Thread-Influence-Of-English-On-Indian-Languages

English Words	Hindi Words Google Transliteration	Standard Hindi Keywords	Phonetic Equivalent/s		Search Engine Google			
Woman	**वोमन**	वुमन	वूमैन	वूमेन	42,200	101,000	3,520	34,800
Insurance	**इन्सुरांस**	इंश्योरेंस	इंश्योरंस	इन्श्योरेन्स	1,220	246,000	1,390	3,710
Cancer	**कैंसर**	केंसर	कैन्सर	केन्सर	1,880,000	35,700	6,490	1,820
Hospital	**हॉस्पिटल**	हास्पिटल	होस्पिटल	होस्पीटल	404,000	263,200	2,440	,100
Corruption	**कोरुप्तिओन**	करप्शन	करपशन	कोरप्शन	1,890	368,000	1,110	58
Computer	**कंप्यूटर**	कम्प्यूटर	कम्प्युटर	कंप्युटर	4,450,000	1,040,000	2,610,000	537,000
University	**उनिवेर्सित्य**	यूनिवर्सिटी	युनिवेर्सिटी	युनिवर्सटी	1,540	1,070,000	1,420	3,270
Director	**डिरेक्टर**	डायरेक्टर	डाइरेक्टर	डायरैक्टर	4,600	735,000	97,000	8,300
Accident	**एक्सिडेंट**	एक्सीडेंट	एक्सीडेन्ट	ऐक्सिडेंट	26,300	125,000	2,510	5,160
Parliament	**पर्लिअमेंट**	पार्लियामेंट	पार्लिमेंट	पार्लियामैंट	3,240	38,000	639	1,120
specialist	**स्पेसिअलिस्ट**	स्पेशियलिस्ट	स्पेशलिस्ट	स्पेशालिस्ट	143	1,440	51,300	9,820
Expert	**एक्सपर्ट**	एक्स्पर्ट	ऐक्सपर्ट	ऐक्स्पर्ट	269,000	8,730	182	8
Policy	**पोलिस्य**	पॉलिसी	पोलिसी	पालिसी	609	395,000	69,700	289,000

Fig. 1 keywords along with their phonetic variants written in Hindi

		Influenced Hindi Query		Google		
In English	**Hindi Query**	**Level 1**	**Level 2**	**Search Results**		
Health and blood donation	स्वास्थ और रक्तदान	**हेल्थ** और रक्तदान	**हेल्थ** और **ब्लड_डोनेशन**	1,520	8,360	1,020
Treatment for Blood pressure	रक्तचाप का इलाज	**ब्लडप्रेशर** का इलाज	**ब्लडप्रेशर** का **ट्रीटमेंट**	95,800	37,200	2,150
Cardiologist	हृदय चिकित्सक	**हार्ट डॉक्टर**	**कार्डियोलाजिस्ट**	241,000	99,100	1,770
Government Employment Policy	सरकार द्वारा रोज़गार योजना	सरकार द्वारा रोज़गार **स्कीम**	सरकार द्वारा **एम्प्लॉयमेंट** स्कीम	1,840,000	51,600	93
Foreign investment in India	विदेशी निवेश भारत में	**फारेन** निवेश भारत में	**फारेन इन्वेस्टमेंट** भारत में	658,000	527	66
Corruption free India.	भ्रष्टाचार मुक्त भारत	**करप्शन** मुक्त भारत	**करप्शन फ्री इंडिया**	181,000	19,700	47,000

Fig. 2 Transformed queries into two equivalent senses containing a mixture of both English and Hindi: Tabular representation

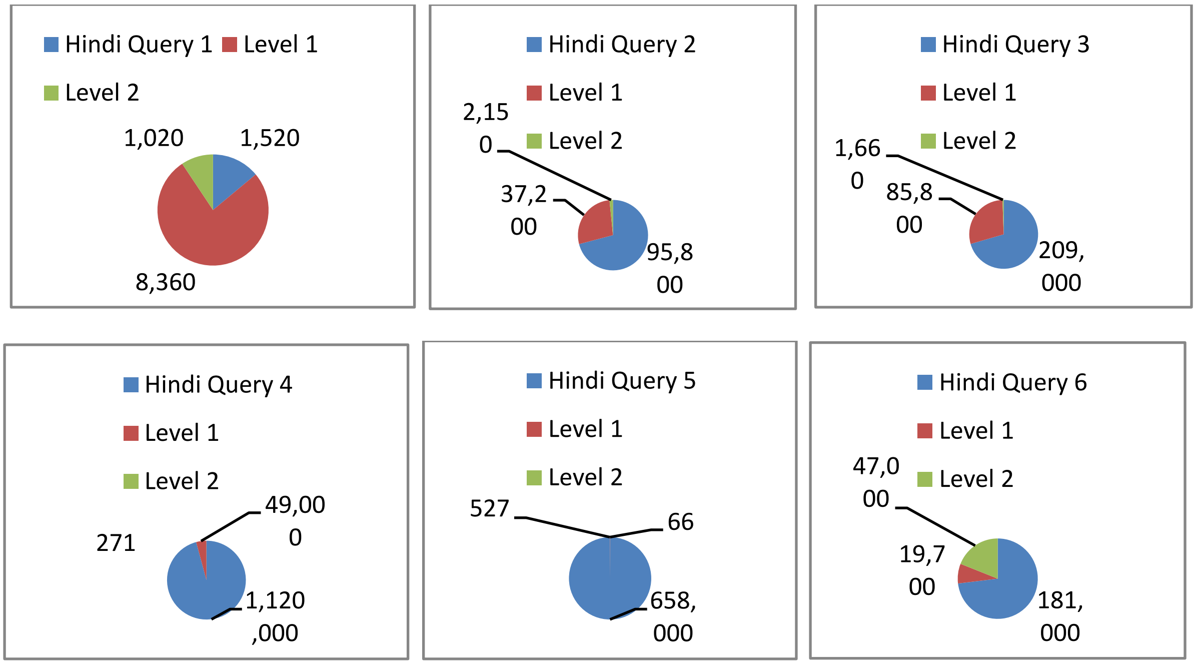

Fig. 3 Transformed Queries into two equivalent senses containing a mixture of both English and Hindi: Graphic representation

Hindi Query	**Influenced Hindi Query**		Precession @ 10								
	Level 1	**Level 2**	Google			Bing			AltaVista		
स्वास्थ और रक्तदान	हेल्थ और रक्तदान	हेल्थ और ब्लड_डोनेशन	0.9	0.9	0.9	0.9	0.8	0.8	0.9	0.9	0.8
रक्तचाप का इलाज	ब्लडप्रेशर का इलाज	ब्लडप्रेशर का ट्रीटमेंट	0.8	0.9	0.7	0.8	0.7	0.5	0.8	0.7	0.5
हृदय चिकित्सक	हार्ट चिकित्सक	कार्डियोलाजिस्ट	0.9	0.8	1	0.9	0.7	1	0.9	0.7	1
सरकार द्वारा रोज़गार योजना	सरकार द्वारा रोज़गार स्कीम	सरकार द्वारा एम्प्लॉयमेंट स्कीम	1	0.9	0.8	0.8	0.8	0.8	0.6	0.7	0.8
विदेशी निवेश भारत में	फारेन निवेश भारत में	फारेन इन्वेस्टमेंट भारत में	0.9	0.8	0.8	0.9	0.8	0.4	0.9	0.9	0.4
भ्रष्टाचार मुक्त भारत	करप्शन मुक्त भारत	करप्शन फ्री इंडिया	1	1	1	1	1	1	1	1	1

Fig. 4 Analysis of precision values: Tabular representation

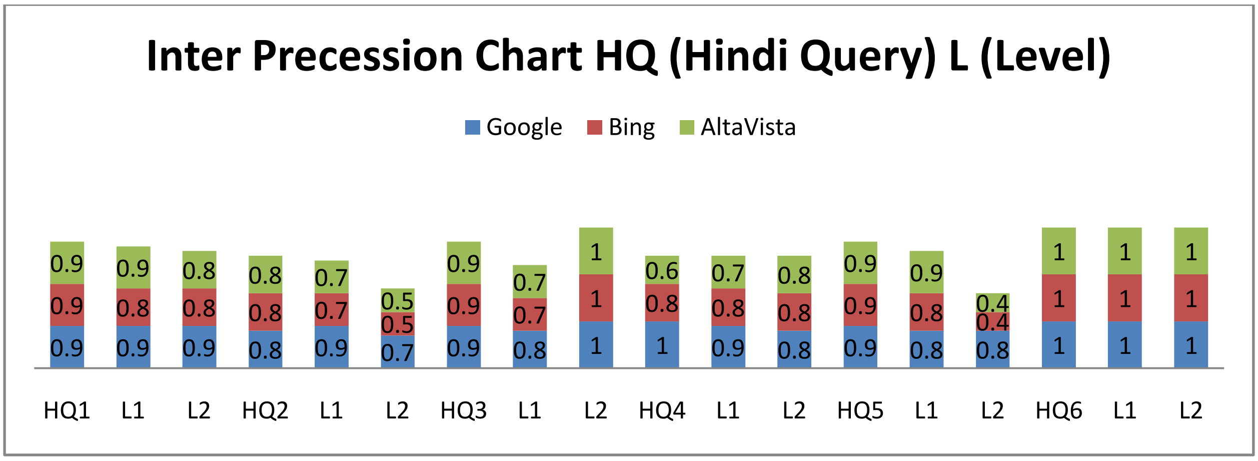

Fig. 5 Analysis of Inter precision values: Graphical representation