Factors Affecting the Performance of Hindi Language searching on web: An Experimental Study.

Kumar Sourabh Vibhakar Mansotra

Abstract

With the internet growing at an exponential rate the web is increasingly hosting web pages in different languages. It is 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 engines to fulfill native user's needs and provide them information in their native languages. The major population of India use Hindi as a first language. The Indian constitution identifies 22 languages, of which six languages (Hindi, Telugu, Tamil, Bengali, Marathi and Gujarati) are spoken by at least 50 million people within the boundaries of the country—there are a large number of them living outside the country. The Hindi language web information retrieval is not in a satisfactory condition. The presence of Hindi on the World Wide Web is still limited and tentative because of attitudinal and technical factors. Besides the other technical setbacks the Hindi language search engines face the problem of morphology, phonetics, word sense disambiguation etc. The performance of search engines is affected by these problems. This paper covers the comprehensive analysis and also the comparison of the affect of language structure related factors (morphology, phonetics, WSD, synonyms,) on the performance of search engines supporting Hindi language.

Keywords: search engines, morphology, word sense disambiguation, precision, Guruji, Raftaar and Hindi Language.

1. Introduction

With the web content being written in different languages of the world, it has become important to have tools that can retrieve information from the documents written in different languages. In the context of Indian languages, Hindi language has been given much emphasis leading to the development of significant number of Hindi documents. In fact, of the top 100 languages in the world, English occupies the top position, with Hindi coming fifth. [1].

Hindi language information retrieval on the web is still in its nascent stage. The number of users who want the information in Hindi language is increasing. This leads to the demand of the Hindi information retrieval on the web. It is the fact that to date Internet is vigorously used in India by the people who are comfortable in English language. The under development of web in Indian regional languages is one of the important reasons behind the limited growth of Internet in India. Indians use 22 official languages and 11 written script forms and among all the languages 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 [2].

It is the language of dozens of major newspapers, magazines, radio and television stations and of

- *Kumar Sourabh is currently pursuing PhD degree program in Computer science University of Jammu, India, PH-9469163570. E-mail: kumar9211.sourabhe@gmail.com*
- *Vibhakar Mansotra is currently Associate professor in University of Jammu, India, PH-9419103488. E-mail:vibhakar20@yahoo.co.in*

other media. This generates the need of the development of the powerful tools for Hindi language information retrieval. [3].

2. Encoding Standards for Indian Languages

Majority of information seekers use a search engine to begin their Web activity. In this case, users submit a query, typically a list of keywords, and receive a list of Web pages that may be relevant, typically pages that contain the keywords. Today though considerable amount of content is available in Indian languages, users are unable to search such content because Indian language websites rely on unique encodings or proprietary extensions of existing standard encodings [4]

The two main standards in character representation of Indian languages are ISCII and Unicode.

2.1 Indian Standard Code for Information Interchange (ISCII).

2.2 Unicode

The Unicode standard provides with three encoding formats: UTF-8, UTF-16 and UTF-32. Any one of these forms can be used to represent the Unicode characters. Each of these is used in different environments. The default encoding form of Unicode is UTF-16. [5]

Most often it has been observed that the use of proprietary fonts of different standards in Central Government Offices in India, which are not compatible with each other, is causing serious problems in information exchange amongst these offices, In order to facilitate free exchange of information/files/documents. Department of Information Technology, Government of India has accepted Unicode encoding for fonts as Indian standard in this regard. [6]

3. Hindi language and web searching

The spread of Internet in India is today constrained by the fact that mostly the English knowing have been benefited by Internet which is a disappointing situation as the real benefit of internet does not reach to the common man having less/no knowledge of English language.

The under development of Web in Indian regional languages is one of the important reasons behind the limited growth of Internet in India. A recent survey by a Delhi based research organization - Juxt Consult - says that 44 % of existing Internet users in India prefer Hindi to 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. Despite the latent demand among Internet users for Hindi, if there is very dismal use of Hindi, it is due to certain constraints. These include technological, attitudinal and economic factors. The most important hardware used for internet surfing is the keyboard. Various Hindi keyboards are available in different varieties. Most of the keyboards are phonetically different. However, a detailed analysis of whether these are truly optimal or better arrangements exist, has not been done. Most of this research has been in two broad directions: *Normal Keyboards Ambiguous Keyboards* [7]

Another constraint in spread of Hindi over the WEB is that of limited content. Where there are more than 20 billion pages on web in English, this number is not more than 10 million in Hindi. This poverty of content is partly due to technological factors and partly due to attitudinal. It is a big dilemma that on the one hand the number of Hindi readers and number of Hindi-speaking people using mobile and computers is so large, and on the other hand the websites are very limited in their content and number.[8] This dilemma should be overcome as soon as possible.

4. Search engines supporting Hindi Language

Now a day's various search engines support information retrieval in different languages. Google, yahoo, Bing, AltaVista are popular worldwide for searching the web. Recently Hindi search engines like Guruji and Raftaar from India have emerged out for information retrieval in Hindi Language. These Indian search engines are new as compared to international search engines listed above. These search engines find their usage in India for Hindi IR. In this Paper we chose Google, Bing and Guruji for experiment purposes based upon their usage and popularity in India.

5. Factors affecting performance of Hindi language searching on web.

Factors which affect searching Hindi information on web are.

- **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. [9]

- **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. [10]

- **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.

In this paper we have focused on these four major and critical problems, details and experimental analysis are discussed below.

6 Experimental Study On

6.1 Morphological Factors

Hindi language is morphologically rich language. It has well defined morphological structure and well defined grammar. But the grammatical and language structural standard is least followed due to various reasons. One of the reasons is the language diversity in India. Including Hindi there are about 28 Languages spoken in India and Hindi being the National Language of India is influenced by the regional languages which results a change in dialects not only in speaking but writing also. Every language uses some markers like (English language uses s, es, ing and ऐं, यां, ◌ो, ओं *MAATRAAS* in hindi language) are used with a root word and new words are constructed . For ex. (Planning in English) *Yojnaaon* योजनाओं, *Yojnaayein* योजनाएं in Hindi, are the morphological variants of root word *Yojnaa* योजना. It is desirable to combine all the morphological variants of the words in a single canonical form. The process is called as word stemming and this canonical form is called as root word or base word. We have taken a sample set of 50 queries to test the affect of the root word. Following table (6.1) is the set of randomly selected queries from the set which throw light on the effect of the root word on the performance of Hindi language search engines. Table 6.1.1 shows the results of experiments for effect of morphological factors on Hindi queries.

(Space for Tables 6.1 and 6.1.1)

It has been observed that documents returned by all three search engines are more in number when query with root word is submitted. This justifies the searching of documents in the root word because in general we get better results with the keywords in their root form.

It has also been observed that only Google shows listing of morphological variants of root words, where as Bing and Guruji show only listing of root word supplied in almost all the sample queries listed above in the table.

From the above results it is evident that only Google indexes the documents keyword in their root form. Bing and Guruji do not index in that form that is the reason number of documents retrieved in their case is less in comparison to Google. The overall comparison of results from the three search engines in tables above show that in general the quantity of results retrieved increased when the keywords are used in their root form. In case of search engines the quality of results is more important than the quantity. Table 6.1.2 and Graph (6.1.3) shows the comparison of the precision values of the three search engines. The precision value is calculated by taking the top 10 results of the search engines. On closely observing the results we can say that precision value in case of Google is high in almost all queries. As mentioned above Google does its indexing in the root form of keywords it can be concluded from the table above that relevancy of the results is also high in Google in comparison to other two search engines which denotes that not only quantity but the quality of results is also affected by the morphological variations in the keywords.

6.2 Phonetic nature of Hindi Language and Spelling variations

The major reasons for spelling variations in language can be attributed to the phonetic nature of Indian languages and multiple dialects, transliteration of proper names, words borrowed from regional and foreign languages, and the phonetic variety in Indian language alphabet. The variety in the alphabet, different dialects and influence of regional and foreign languages has resulted in spelling variations of the same word [11]. For example; Following are the possible spelling variations for the Hindi word अंग्रेजी (angrējī): (means English)

अँग्रेजी, अंगरेजी, अन्ग्रेजी, अँगरेजी, अंग्रेजी, अंग्रेज़ी

There are numerous words which are phonetically equivalent but vary in writing.

The word *school* in hindi can be written in different ways (स्कूल, सकूल, स्कुल) When information is searched for a single standard keyword school स्कूल and non standard Hindi phonetic equivalent keyword स्कुल 6.9 million results are shown by Google for former and 1.4 million for later. Hindi Language is influenced by the other regional languages which results in phonetic variety of words for example the English word school (स्कूल in Hindi) is pronounced and written as *ISKOOL* इस्कूल by the majority of population of India in different states. For the Hindi word *ISKOOL* इस्कूल more than two thousand results are found. Search engines should be capable of retrieving the results against phonetically equivalent words of keywords entered to search. User may use any keyword for searching and search engines should be capable to support all phonetically equivalent words. Following are randomly selected queries from the set of 50 queries tested on Google search engine

Following table 6.2.1 and graph 6.2.2 below show the results and precision offered by Google.

(Space for Table and graph 6.2.1 6.2.2)

In the above table it can be clearly seen that search engines return a handful of documents on various Hindi phonetically equivalent queries. It is observed that no particular standard exists for writing the keyword to fetch Hindi web data. For every phonetically equivalent keyword/s in the query variation in the results exist. I.e. a different set of documents are retrieved with least repetition. From the precision chart it is clearly observed that the degree of relevance for queries containing phonetically equivalent keywords is almost same or nearly equal. The native Hindi user may not be aware of the Phonetic issues in Hindi IR and may miss the relevant information of his/her use.

6.3 Words Synonyms

A word can express a myriad of implications, connotations, and attitudes in addition to its basic "dictionary" meaning. And a word often has near synonyms that differ from it solely in these nuances of meaning. Choosing the right word can be difficult for people, as well as for the information retrieval system. For example the word (आभूषण) in Hindi (Ornament) in English, has following commonly spoken synonyms गहने, जेवर, अलंकार.

Table 6.3.1 and graph 6.3.2 has been presented below which shows the comparison of precision values against three search engines.

(Space for table 6.3.1 and graph 6.3.2)

From the experiments researchers observed that using Hindi keywords with their synonyms improves the information retrieval against a query in Hindi language.

Not only quantity of documents returned is affected but quality is also affected by using synonyms of Hindi keywords.

From the above table and graph it is be observed that documents returned by Google are more in quantity than other two search engines and least number of documents are returned by Guruji search engine the reason behind may be availability of less documents or poor indexing . However we are interested in quality of results than quantity, As far as quality of results is concerned it can be clearly seen that Google and Bing provide quality data than Guruji. And in the average case Google still stands first in the row that means precision values by Google are more than that of Bing and Guruji in this case. Thus it becomes clear that by changing a keyword into its synonym equivalent, results can be obtained. Therefore it is evident that synonyms of keywords play an important role in the process of Hindi information retrieval system.

6.4 Ambiguous words

Ambiguous words deflate the relevancy of the results. The examples mentioned below shows this aspect very clearly. Consider the following query.

(In English) →(Women like gold).

(In Hindi)→ (नारी को सोना पसंद है).

In this query the word सोना (Gold) is ambiguous as it has another meaning i.e. to sleep. In the context of above query the word सोना is gold. But it can be also interpreted as women like to sleep.

Another Query: (In English) →(The common people's choice).

(In Hindi) →(आम लोगों की पसंद).

Here the word आम is ambiguous. The word आम in above query means common. However, In Hindi it also means mango. So the above query can be interpreted as "mango is people's choice". Various experiments have been done on this issue on three search engines to check their performance on handling of ambiguous word in a particular context.

We experimented on a sample set of 50 ambiguous queries and below we present five randomly selected ambiguous queries. In table (6.4.1) second column contains five queries in Hindi, third column holds the ambiguous keyword in one context and fifth column holds the same ambiguous keyword in other context. Fourth and sixth columns hold the meaning of queries in English with respect to the ambiguous keyword in context.

(Space for Table (6.4.1))

Ambiguous queries mentioned above in the table are tested for results against three search engines Google, Bing and Guruji. Results are shown below in tables 6.4.1, 6.4.2 and 6.4.3 as.

(Space for Tables 6.4.1, 6.4.2 and 6.4.3)

From the above results obtained in tables it is observed that all three search engines return documents without differentiating between the contexts of keyword in the query. In the above table the last column labeled as "other Context" holds the number of results which are not relevant to the query supplied or those documents which contains the keywords in other non required context. From the results it is clear that all search engines return documents in different contexts. Therefore it can be concluded that search engines underperform when supplied with ambiguous queries. Numbers in column labeled as "other Context" signifies the deviation from relevance. For example for query युद्ध में कुल विनाश (aggregate destruction in wars) the column "Other Context" for Google contains 5 documents for Bing contains 8 documents and for Guruji contains all 10 documents.

In another query सपेरों का फन (art of snake charmers) another context (Snake charmer's snake head) retrieved documents are expected to be in context (art) but from the above results obtained it can be seen that google returns all 10 documents in non required context (snake head) and Bing returns 9 documents where as Guruji fails to retrieve even a single document.

In the above scenario it becomes important for the search engines to address to the issue of ambiguity in keywords to obtain better results.

7 Discussion

We tested the performance of three search engines Google. Bing and Guruji for the challenges mentioned in the section (5). After the comparative analysis it is concluded that for *morphology*, query when supplied with root word yields maximum results. It is also evident that Google indexes the keyword in its root form and also lists the documents consisting of the morphological variants of the keywords. The other two engines Bing and Guruji do not list any morphological variant and hence they entail to have stemmer.

It is apparent from the table (6.2.1) that *phonetic* variation in Hindi keyword has a great impact on the performance of search engines. For each phonetically different word different set of results are obtained. Precision graph 6.2.2 also shows that the relevancy factor for all phonetic equivalent keywords contained in the queries is nearly equal for average case.

Word *synonyms* also play a major role in Hindi information retrieval process as it has been shown in table 6.3.1 above that a word with its synonyms when supplied to the search engines results in the retrieval of hand full of documents and none of the search engines is capable of listing a synonym of a keyword in the documents retrieved. However, the precision values of Google are better than other two search engines.

Ambiguous words in a query bring down the performance of search engines. None of the search engine is capable enough to handle the problem of ambiguity in query. It was observed that search results were far away from relevance and results obtained are out of context in almost all the cases.

8. Conclusions

In this paper we discussed the issues and problems which a user may face while finding Hindi information on web. We tested the parameters that affect the Hindi search on web. Search engines may have the performance and throughput problems if these parameters are implemented at root level. However this problem can be solved at interface level. Therefore in this direction we have developed software with a large scale Hindi database which is an interface between Hindi user and search engine. The software takes care of the Hindi phonetic variants, word synonyms and regional/foreign words that influence the Hindi Language. Complete description and implementation details will be reported shortly.

9 References

[1]. Praveen Kumar, Shrikant Kashyap, Ankush Mittal Indian Institute of Technology, Roorkee, India. "A Query Answering System for E-Learning Hindi Documents" SOUTH ASIAN LANGUAGE REVIEW VOL.XIII, Nos 1&2, January-June,2003

[2] S.K. Dwivedi and Parul Rastogi Babasaheb Bhimrao Ambedkar University (A Central University), Lucknow, Uttar Pradesh, India "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

[3] 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.

[4] Prasad Pingali Jagadeesh Jagarlamudi Vasudeva Varma Language Technologies Research Centre IIIT, Hyderabad India. WebKhoj: Indian language IR from Multiple Character Encodings. URL: - www2006.org/programme/files/pdf/5503.pdf

[5] Dimple Patel 1 and Devika P. Madalli 2 1 Mahatma Gandhi National Institute of Research and Social Action, Hyderabad, India 2 Documentation Research & Training Centre, Indian Statistical Institute, Bangalore, India "Information Retrieval in Indian Languages: A Case Study of Plural Resolution in Telugu Language". URL: - drtc.isibang.ac.in:8080/bitstream/handle/1849/396/054_p45_dimple.pdf

[6] Government of India Ministry of Home Affairs Department of Official Language Technical Cell "Office Memorandum No 12015/7/2008-OL (TC.)

[7] Priyendra S. Deshwal Kalyanmoy Deb Department of Computer Science and Engineering Indian Institute of Technology, Kanpur. Kanpur PIN-208016, India."Design of an Optimal Hindi Keyboard for Convenient and Efficient Use" URL: - www.iitk.ac.in/kangal/papers/k2003004.pdf

[8] Ranjan Srivastava, Chief of Bureau, Prabhat Khabar, (Friday, April 28, 2006) "The future of Hindi on the Internet" http://www.raftaar.in/thehoot.htm.

[9] 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

[10] GANAPATHIRAJU Madhavi, BALAKRISHNAN Mini, BALAKRISHNAN N. REDDY Raj "Om: One tool for many (Indian) languages" Journal of Zhejiang University SCIENCE ISSN 1009-3095.

[11] Vishal Goyal, Ph.D. Development of a Hindi to Punjabi Machine Translation System
A Doctoral Dissertation Volume 10: 10 October 2010 ISSN 1930-2940 Language in India www.languageinindia.com

S. No	Query in Hindi	Meaning In English	S.NO	Query in Hindi	Meaning In English
1	भारत वर्षावन	Indian rain forest	6.2	पक्षियों की प्रजातियां	Bird's species
1.1	भारतीय वर्षा वनों	Indian rain forests	7	कृषि समस्या	Agriculture problem
2	हवाई दुर्घटना का कारण	Reason for air crash	7.1	कृषि समस्याओं	Agriculture problems
2.1	हवाई दुर्घटनाओं का कारण	Reason for air crashes	8	कीटनाशक के इस्तेमाल	Use of pesticide
3	भारत में बोली जाने वाली भाषा	Language spoken in India	8.1	कीटनाशकों के इस्तेमाल	Use of pesticides
3.1	भारत में बोली जाने वाली भाषाएँ	Languages spoken in India	9	मानसिक रोग	Mental illness
3.2	भारत में बोली जाने वाली भाषाओँ	Languages spoken in India	9.1	मानसिक रोगियों	Mental patients
4	विलुप्त होने पर झील	Lake on the verge of extinction	9.2	मानसिक रोगी	Mental Patient
4.1	विलुप्त होने पर झीलें	Lakes on the verge of extinction	10	ग्रामीण विकास योजना	Policy for village
5	प्राकृतिक आपदा	Natural calamity	10.1	ग्रामीण विकास योजनाओं	Policies for village
5.1	प्राकृतिक आपदाएँ	Natural calamities	10.2	ग्रामीण विकास योजनाएं	Policies for village
6	पक्षी की प्रजाति	Bird species	11	प्रमुख कृषि केंद्र	Major agricultural office
6.1	पक्षियों की प्रजाति	Bird's species	11.1	प्रमुख कृषि केन्द्रों	Major agricultural offices

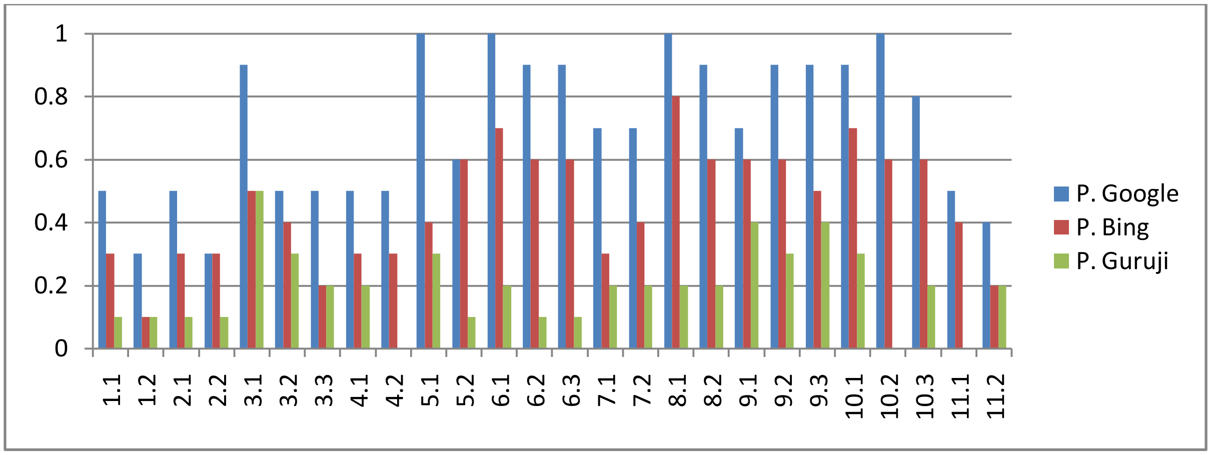
Table 6.1 List of Hindi queries

S. No	Root word/s	Listing of Keywords			Morphological variants		Documents Returned		
		Google	Bing	Guruji			Google	Bing	Guruji
1	भारत , वर्षा वन	भारत, वर्षावन, वर्षा वन, वन	वर्षा वन, वन	भारत ,वर्षा वन, वन	1.1	भारत , वर्षा वन	50,500	4,680	485
					1.2	भारतीय वर्षा वनों	40,400	680	61
2	दुर्घटना	दुर्घटना,दुर्घटनाओं,	दुर्घटना	दुर्घटना	2.1	दुर्घटना	133,000	2,410	284
					2.2	दुर्घटनाओं	117,000	420	23
3	भाषा	भाषा, भाषाओं,	भाषा	भाषा	3.1	भाषा	161,000	8,330	961
					3.2	भाषाएँ	6,200	935	188
					3.3	भाषाओँ	6,090	441	356
4	झील	झील, झीलों,	झील	झील	4.1	झील	4,740	278	25
					4.2	झीलें	1,270	28	1
5	आपदा	आपदा, आपदाओं,	आपदा	आपदा	5.1	आपदा	102,000	4,030	410
					5.2	आपदाएँ	1,160	64	20
6	पक्षी,प्रजाति	पक्षी, पक्षियों, प्रजाति, प्रजातियों, प्रजातियां	पक्षी, प्रजाति	पक्षी, प्रजाति	6.1	पक्षी ,प्रजाति	48,200	1,670	98
					6.2	पक्षियों, प्रजाति	47,600	1,150	84
					6.3	पक्षियों ,प्रजातियां	33,800	747	25
7	समस्या	समस्याएं,समस्याओं, समस्या	समस्या	समस्या	7.1	समस्या	584,000	30,200	1,889
					7.2	समस्याओं	584,000	7,150	1,356
8	कीटनाशक	कीटनाशक, कीटनाशकों	कीटनाशक	कीटनाशक	8.1	कीटनाशक	36,300	1,360	333
					8.2	कीटनाशकों	35,800	800	270
9	रोग	रोगों, रोग	रोग	रोग	9.1	रोग	205,000	21,600	1,423
					9.2	रोगियों	128,000	3,280	239
					9.3	रोगी	112,000	6,280	647
10	योजना	योजनाओं,योजना	योजना	योजना	10.1	योजना	673,000	18,500	3,343
					10.2	योजनाओं	669,000	6,020	990
					10.3	योजनाएं	673,000	2,860	416
11	केंद्र	केंद्रीय, केंद्र,	केंद्र,	केंद्र,	11.1	केंद्र	261,000	11,300	655
					11.2	केन्द्रों	29,500	1,850	105

Table 6.1.1 Effect of morphological factors on Hindi queries

S. No	**Query**	**Precision@ 10**			**S.NO**	**Query**	**Precision@ 10**		
		Google	**Bing**	**Guruj i**			**Google**	**Bing**	**Guruji**
1.1	भारत वर्षा वन	0.5	0.3	0.1	**6.3**	पक्षियों की प्रजातियां	0.9	0.6	0.1
1.2	भारतीय वर्षा वनों	0.3	0.1	0.1	**7.1**	कृषि समस्या	0.7	0.3	0.2
2.2	हवाई दुर्घटना का कारण	0.5	0.3	0.1	**7.2**	कृषि समस्याओं	0.7	0.4	0.2
2.2	हवाई दुर्घटनाओं का कारण	0.3	0.3	0.1	**8.1**	कीटनाशक के इस्तेमाल	1	0.8	0.2
3.1	भारत में बोली जाने वाली भाषा	0.9	0.5	0.5	**8.2**	कीटनाशकों के इस्तेमाल	0.9	0.6	0.2
3.2	भारत में बोली जाने वाली भाषाएँ	0.5	0.4	0.3	**9.1**	मानसिक रोग	0.7	0.6	0.4
3.3	भारत में बोली जाने वाली भाषाओँ	0.5	0.2	0.2	**9.2**	मानसिक रोगियों	0.9	0.6	0.3
4.1	विलुप्त होने पर झील	0.5	0.3	0.2	**9.3**	मानसिक रोगी	0.9	0.5	0.4
4.2	विलुप्त होने पर झीलें	0.5	0.3	0	**10.1**	ग्रामीण विकास योजना	0.9	0.7	0.3
5.1	प्राकृतिक आपदा	1	0.4	0.3	**10.2**	ग्रामीण विकास योजनाओं	1	0.6	0
5.2	प्राकृतिक आपदाएँ	0.6	0.6	0.1	**10.3**	ग्रामीण विकास योजनाएं	0.8	0.6	0.2
6.1	पक्षी की प्रजाति	1	0.7	0.2	**11.1**	प्रमुख कृषि केंद्र	0.5	0.4	0
6.2	पक्षियों की प्रजाति	0.9	0.6	0.1	**11.2**	प्रमुख कृषि केन्द्रों	0.4	0.2	0.2

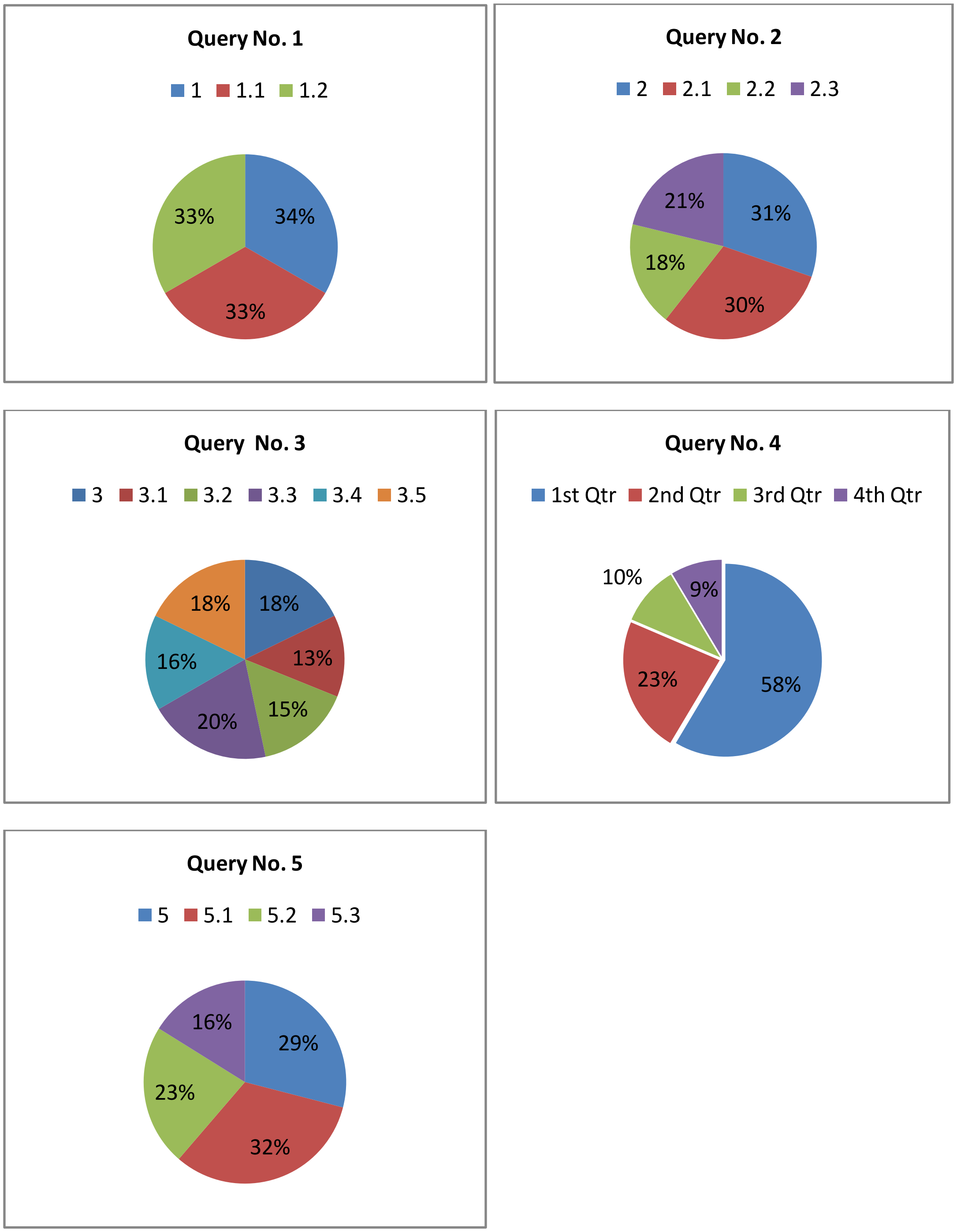
Table 6.1.2 precision values of the three search engines



Graph 6.1.3 precision values of the three search engines

Hindi Query With Bold Standard Keywords	Phonetic variations of the Keywords			Google Results for query having keywords	No. of Results	Precision @10
सब्ज़ियों में ज़हरीले पदार्थ	सब्जियों सब्जीयों,	जहरीले, जहरिले		**सब्ज़ियों, ज़हरीले**	97	0.9
				सब्जियों, ज़हरीले	632	0.9
				सब्ज़ियों, **जहरीले**	194	0.9
आसमान छूती महंगाई	आसमां,	महँगाई , मेहंगाई		आसमान **महंगाई**	35300	1.0
				आसमान **महँगाई**	1040	1.0
				आसमां महँगाई	14	0.6
				आसमां महंगाई	563	0.7
भ्रष्टाचार से आज़ादी	भ्रश्टाचार भरष्टाचार	आजादी		**भ्रष्टाचार** आज़ादी	211,000	0.8
				भ्रश्टाचार आज़ादी	214	0.6
				भरष्टाचार आज़ादी	447	0.7
				भ्रष्टाचार **आजादी**	1,090,000	0.9
				भ्रश्टाचार **आजादी**	1,040	0.7
				भरष्टाचार **आजादी**	1,190	0.8
अन्ना हज़ारे का आन्दोलन	अनना	हजारे	आंदोलन आँदोलन	अन्ना हज़ारे **आन्दोलन**	84,700	0.3
				अन्ना हज़ारे **आंदोलन**	85,100	0.8
				अन्ना हज़ारे का **आँदोलन**	78	0.6
				अनना हजारे का आन्दोलन	399	0.5
				अन्ना **हजारे** का आन्दोलन	3,260,000	1.0
बेरोज़गारी समस्या समाधान	बेरोजगारी बरोजगारी बेरोज़ग़ारी			**बेरोज़गारी**	9,650	0.9
				बेरोजगारी	80,600	1.0
				बरोजगारी	170	0.7
				बेरोज़ग़ारी	30	0.5

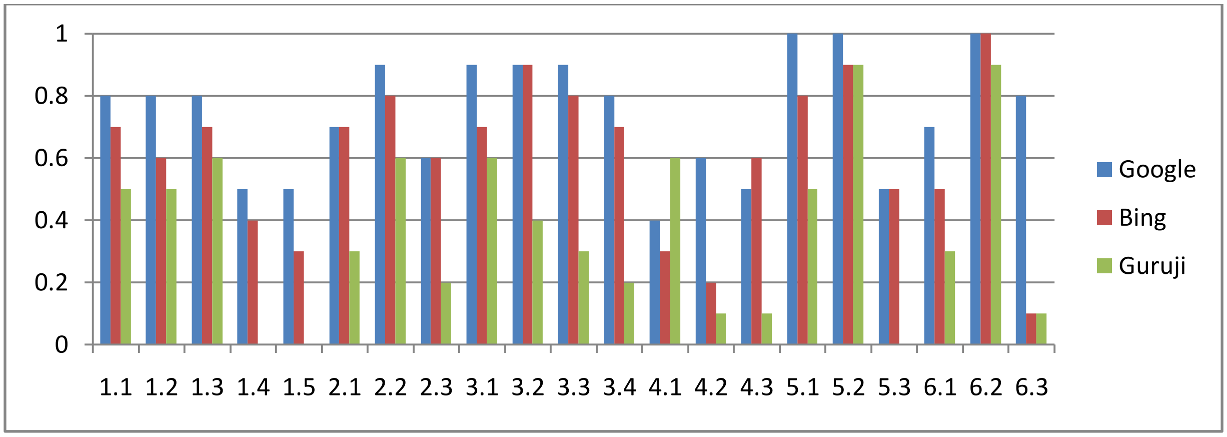
Table 6.2.1 results of three search engines on Phonetic nature of Hindi language



Graphs 6.2.2 Precision Charts for Phonetic nature of Hindi language

S. NO	Query	Standard Hindi Word/s	Synonyms		Documents Returned					
					Google	Per. @10	Bing	Per. @10	Guruji	Per. @10
1	सोने के आभूषण	आभूषण/ गहने	1.1	सोने के आभूषण	217,000	0.8	3,250	0.7	381	0.5
			1.2	सोने के गहने	188,000	0.8	2,590	0.6	389	0.5
			1.3	सोने के जेवर	78,900	0.8	1,670	0.7	311	0.6
			1.4	सोने के अलंकार	9,490	0.5	633	0.4	70	0
			1.5	सोने के आभरण	493	0.5	38	0.3	1	0
2	काले बादल	बादल	2.1	काले बादल	233,000	0.7	7,510	0.7	733	0.3
			2.2	काले मेघ	40,700	0.9	1,500	0.8	99	0.6
			2.3	काले जलधर	1,570	0.6	54	0.6	2	0.2
3	स्त्री सशक्तिकरण	स्त्री,नारी	3.1	स्त्री सशक्तिकरण	9,950	0.9	1,570	0.7	760	0.6
			3.2	नारी सशक्तिकरण	29,300	0.9	1,910	0.9	736	0.4
			3.3	महिला सशक्तिकरण	96,300	0.9	5,160	0.8	1,091	0.3
			3.4	औरत सशक्तिकरण	7,670	0.8	680	0.7	510	0.2
4	सिकंदर का अंहकार	अंहकार	4.1	सिकंदर का अंहकार	1,990	0.4	18	0.3	60	0.6
			4.2	सिकंदर का अभिमान	2,400	0.6	304	0.2	16	0.1
			4.3	सिकंदर का घमंड	495	0.5	54	0.6	9	0.1
5	वृक्ष लगाओ	वृक्ष	5.1	वृक्ष लगाओ	6,960	1	698	0.8	29	0.5
			5.2	पेड़ लगाओ	13,400	1	1,080	0.9	143	0.9
			5.3	दरख्त लगाओ	481	0.5	19	0.5	0	0
6	आँख दान	आँख	6.1	आँख दान	34,000	0.7	3,690	0.5	312	0.3
			6.2	नेत्र दान	77,500	1	3,240	1	159	0.9
			6.3	चक्षु दान	2,450	0.8	427	0.1	36	0.1

Table 6.3.1 effect of word synonyms on Hindi IR



Graph 6.3.2 comparison of precision values against three search engines

S.No	Query	For keyword as	In English	For keyword as	In English
1	नारी को सोना पसंद है	सोना (Gold)	Women like Gold	सोना (To sleep)	Women like to sleep
2	आम लोगों की पसंद	आम (common)	Common man's choice	आम (Mango)	Mango is people's choice
3	बाल विकास और पोषण	बाल (Children)	Child Development and Nutrition	बाल(Hair)	Hair Development and Nutrition
4	सपेरों का फन	फन (Art)	Art of snake charmers	फन(Snake head)	Snake charmer's snake head
5	युद्ध में कुल विनाश	कुल (Aggregate)	Aggregate destruction in wars	कुल(family)	Destruction of families in war

Table 6.4.1 List of randomly selected ambiguous queries

Query	Ambiguous keyword	Documents returned	Google				Other Context
			Results Found				
			Context		Context		
1	सोना	50,800	Gold	5	To sleep	2	3
2	आम	488,000	Common	3	Mango	3	4
3	बाल	2,900,000	Children	7	Hair	3	0
4	फन	184	Art	0	Snake head	10	0
5	कुल	17,800	Aggregate	2	Family	3	5

Table 6.4.2 Ambiguity test for Google

Query	Ambiguous keyword	Documents returned	Bing				Other Context
			Results Found				
			Context		Context		
1	सोना	2,680	Gold	2	To sleep	2	6
2	आम	17,800	Common	3	Mango	3	4
3	बाल	4,030	Children	6	Hair	2	2
4	फन	25	Art	0	Snake head	9	1
5	कुल	1,900	Aggregate	0	Family	2	8

Table 6.4.3 Ambiguity test for Bing

Query	Ambiguous keyword	Documents Returned	Guruji				Other Context
			Results Found				
			Context		Context		
1	सोना	109	Gold	0	To sleep	0	10
2	आम	6,756	Common	3	Mango	0	7
3	बाल	635	Children	5	Hair	2	3
4	फन	No Results Found	Art	n/a	Snake head	n/a	n/a
5	कुल	84	Aggregate	0	Family	0	10

Table 6.4.4 Ambiguity test for Guruji