LANGUAGE IN INDIA

Strength for Today and Bright Hope for Tomorrow

Volume 13 : 1 January 2013

ISSN 1930-2940

Query Optimization: Solution for low recall problem in Hindi Language IR - Revisited with Experimental Results and Analysis

Dr. Kumar Sourabh
Prof. Vibhakar Mansotra
Rakesh Goswami, Research Scholar

==========

Abstract

While information retrieval (IR) has been an active field of research for decades, for much of its history it has had a very strong bias towards English as the language of choice for research and evaluation purposes. The Internet is no longer monolingual, as the non-English content is growing rapidly. Hindi is the third most widely spoken language in the world. An estimated 500-600 million people speak this language. Information Retrieval in Hindi language is getting popularity and IR systems face low recall if existing systems are used as-is. Certain characteristics of Indian languages do not enable the existing algorithms to match relevant keywords in the documents for retrieval.

Some of the major characteristics that affect Indian language IR are due to language morphology, compound word formations, word spelling variations, ambiguity, word synonym, foreign language influence, and lack of standards for spelling words.

Taking into consideration the aforesaid issues we introduced Hindi Query Optimization technique in our previous work [4]. In this paper we extend our work by presenting various experiments carried out by using query optimization technique to solve low recall problem in Hindi Language IR.

Keywords: Information retrieval, Hindi, Monolingual, Query optimization, Interface, Hindi WordNet.

1. Introduction

While information retrieval (IR) has been an active field of research for decades, for much of its history it has had a very strong bias towards English as the language of choice for research and evaluation purposes. Internet shows more inclination toward the use of plurality of languages, as the non-English content is growing rapidly. More people have begun to send and receive e-mails, searching for information, reading e-papers, blogging and launching web sites in their own languages. Hindi is the third most widely-spoken language in the world (after English and Mandarin): an estimated 500-600 million people speak this language. Two American IT companies, Microsoft and Google, have played a big role in making this possible.

Realizing the potential of Indian languages, Microsoft and Google have launched various products in the past two years. With Google Hindi and Urdu search engines, one can search all the Hindi and Urdu Web pages available on the Internet, including those that are not in Unicode font. Google also provides transliteration in Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Nepali, Punjabi, Tamil, Telugu and Urdu and offers searching in 13 languages, Hindi, Tamil, Kannada, Malayalam and Telugu, to name a few. [1].

India-centric localized search engines market is growing fast. In last year alone there have been more than 10-15 Indian local search engines launched. Here are some of the search

engines who fall into the localized Indian search engine category. Guruji, Raftaar Hinkhoj, Hindi Search Engine, Yanthram, Justdial, Tolmolbol, burrp, Dwaar, onyomo, khoj, nirantar, bhramara, gladoo, lemmefind.in along with Ask Laila which have been launched a couple of months back. Also, we do have localized versions of those big giants Google, Yahoo and MSN. Each of these Indian search engines have come forward with some or the other USP (Unique Selling Proposition). However, it is too early to pass a judgment on any of them as these are in testing stages and every start-up is adding new features and making their services better.

Many 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 for such content.

Information Retrieval in Hindi language is getting popularity and IR systems face low recall if existing systems are used as is. Certain characteristics of Indian languages cause the existing algorithms to become unable to match relevant keywords in the documents for retrieval. Some of the major characteristics that affect Indian language IR are due to language *morphology, compound word formations, word spelling variations, Ambiguity, Word Synonym, foreign language influence, lack of standards for spelling words*. [2][3]

Taking into consideration the aforesaid issues we introduced Hindi Query Optimization technique [4]. Query optimization is one in its own kind. It is the first initiative taken in the field of monolingual Hindi IR. Almost all phonetic, synonym English equivalent Hindi keywords, phonetic variations of proper nouns and wrongly transliterated keywords converted to correct form are at their disposal and the optimized version of the query is suggested to the user so that effective process of Hindi IR can be carried out.

In this paper we discuss our experiments related to monolingual IR and web IR in the context of Hindi language. Queries received by users were organized into various domains namely “Agriculture”, “Science and Technology”, “Medical”, “General” and “Tourism”. Some additional experiments on effect of phonetics and transliteration on proper nouns (names of

individuals and places) were also conducted .The primary objective of the experiments was to study the impact of rephrasing and optimization of query in improving the problem of recall for Hindi language using our interface. The process of data retrieval and precision statistics are discussed in detail in the subsequent sections through various tables. The limitations of the present working system have also been discussed later in this paper.

2. Research Methodology and Data Collection

The software was distributed to various people (with particular domain expertise e.g. Tourism, research scholars of Hindi and other departments, Medical students etc.) and to the novice Hindi users for general purpose Hindi search. The users were also encouraged to make use proper nouns like names of the famous people and places in their searches. Also the users were asked to make use of Hindi queries containing English keywords (written in Hindi). To observe precision of the results the interface has been provided with feedback feature for which existing users have been guided to check the relevance of first ten results and report the relevance as average, Good, Very good and Excellent for a query supplied.

All the queries supplied by the users have been collected in the query log as different groups. Queries from tourism domain are group one queries. Queries from Hindi experts are group two queries. Similarly the pattern follows. Fifteen minutes of training session was organized for each group which indicates the ease of the use/handling of the search interface.

A total of 1245 queries of different nature were collected along with the feedback for the results obtained. The log has been examined for the variations of the queries performed by the uses and it was found that on an average a small query has been variated for four to five times. The maximum variation of the query has been observed as 8-10 times for large queries containing 6-7 keywords words.

From the large number of queries we present randomly picked queries from each group, particularly the queries which contains for which feedback has been provided. Measuring the information retrieval effectiveness of Web search engines can be expensive if human relevance judgments are required to evaluate search results. Using implicit and explicit user feedback for search engine evaluation provides a cost and time effective manner of addressing this problem.

Web search engines can use human evaluation of search results without the expense of human evaluators. An additional advantage of feedback approach is the availability of real time data regarding system performance. We use the explicit feedback to calculate performance metrics, such as precision. This information can lead to more successful relevance feedback techniques.

3. Organization of tables

Before explaining the process of retrieval and precision statistics through the use of tables it is obligatory to explain the role of various tables in handling different types of data. A brief explanation about the kind of data handled by a particular table is given as:

Table 1st: Contains queries related to Agriculture.

Table 2nd: Contains queries related to Science and Technology

Table 3rd: Contains queries related to Medical Domain

Table 4th: Contains queries related to General searches

Table 5th: Contains queries related to Tourism

Column wise representation/organization of the data in these tables is described as:

Column: 1st: Index

Column 2nd: Original query supplied by the user.

Column 3rd: Variants of the query generated through our interface.

Column 4th, 6th, 8th: Search engine results: Quantity of the documents

Column 5th, 7th, 9th: Precision for first 10 results.

Columns 4th, 6th, 8thof the tables given below contain quantity of results returned by the search engines. The quantity of results varies with time and hence does not remain constant as queries are tested live. The arrangement of the results does not follow any order ascending or descending. The first result reflects quantity of the documents returned by the original query and rest reflects the quantity of the documents returned by the variants of the query as received through the interface.

To check the relevance of the results only first ten results are considered as it is generally believed that the most relevant data is present / available in the first few results. The search

engines used for live searches are Google, Yahoo and Guruji. Selection of these search engines is made because of their usage in India as per "Juxt consult online research survey" [5].

The details of the results are shown in the tables given below.

S. NO	Original Query	Query Variations Generated		Documents Returned and precision @10					
				Google	Per. @10	Yahoo	Per. @10	Guruji	Per. @10
Table 1 Queries related to Agriculture Domain									
1	भारतीय कृषि संस्थान	1	भारतीय कृषि संस्थान	354,000	8	10,600	7	779	4
		1.1	भारतीय **किसानी** संस्थान	385,000	8	6,040	5	13	2
		1.2	**इंडियन कृषि संस्था**	80,800	7	13,700	6	75	3
		1.3	भारतीय **खेती संस्था**	265,000	7	7,920	4	272	4
		1.4	**इंडियन एग्रीकल्चर इंस्टिट्यूट**	754	9	20	8	1	1
2	चावल की पैदावार बढ़ाने की तरकीब	2	चावल की पैदावार बढ़ाने की तरकीब	394	6	14	4	6	3
		2.1	चावल की **फसल** बढ़ाने के **उपाय**	10,300	6	84	4	177	4
		2.2	चावल की **फसल** बढ़ाने का **तरीका**	16,500	5	66	3	7	2
3	हिंदुस्तान खेती क्षेत्र में बैंक	3	हिंदुस्तान खेती क्षेत्र में बैंक	37,200	5	9,550	4	9	3
		3.1	**भारत कृषि** क्षेत्र में बैंक	464,000	7	14,400	9	737	6
		3.2	भारत **खेतीबाड़ी** क्षेत्र में बैंक	1,840	7	41	1	20	1
		3.3	इंडिया एग्रीकल्चर फील्ड में बैंक	610	6	36	4	2	1
4	क्रॉप	4	क्रॉप इंश्योरेंस	104	4	3	1R	0	n/a

	इंश्योरेंस पॉलिसी		पॉलिसी						
		4.1	**फसल बीमा योजना**	84,900	9	5,260	6	283	5
		4.2	**फसल** इंश्योरेंस **स्कीम**	2,990	7	33	5	3	1
		4.3	**फसल बीमा** पॉलिसी	5,440	7	83	5	6	2
5	मवेशियों का चुनाव	5	मवेशियों का चुनाव	88,200	1	4,920	1	54	1
		5.1	**जानवरों** का चुनाव	907,000	2	7,840	2	363	2
		5.2	मवेशियों का **चयन**	21,900	2	4,740	2	14	1
		5.3	**जानवरों** का **चयन**	160,000	1	10,600	1	96	1
6	किसान क्रेडिट-कार्ड योज़ना	6	किसान क्रेडिट-कार्ड योज़ना	97	5	10	4	1	1
		6.1	किसान क्रेडिट-कार्ड **योजना**	209,000	6	4,230	8	188	6
		6.2	किसान **क्रेडीट**-कार्ड योजना	522	6	16	7	3	1
		6.3	**कृषक** क्रेडिट कार्ड **योजना**	9,080	7	138	7	6	2
7	कपास फ़सल बचाव	7	कपास फ़सल बचाव	216	3	11	4	0	n/a
		7.1	कपास **फसल** बचाव	8,540	6	7,910	8	17	5
		7.2	कपास **फसल रक्षा**	8,060	4	138	7	186	6
		7.3	कपास **उपज रक्षा**	6,930	4	7,660	4	177	4
8	किसान के	8	किसान के	468,000	8	9,160	6	446	5

	हित में सरकार		हित में सरकार						
		8.1	किसान के हित में **प्रशासन**	186,000	8	9,550	4	196	5
		8.2	किसान के **कल्याण** में सरकार	973,000	5	8,260	4	520	4
		8.3	किसान के हित में **गवर्नमेंट**	4,040	6	86	1	1	1
Table 2 Queries related to Science and Technology Domain									
1	विज्ञानं साहित्य का प्रकाशन	1	विज्ञानं साहित्य का प्रकाशन	3,510	5	345	3	9	3
		1.1	**विज्ञान** साहित्य का प्रकाशन	762,000	7	29,600	5	10,77	5
		1.2	**साइंस लिटरेचर** का प्रकाशन	463	5	124	5	2	1
		1.3	**साइंस लिटरेचर** का **पब्लिकेशन**	69	5	6	4	0	n/a
2	प्लूटो ग्रह का पाँचवाँ चाँद	2	प्लूटो ग्रह का पाँचवाँ चाँद	151	9	178	1	0	n/a
		2.1	प्लूटो ग्रह का **पंचम चंद्र**	187	4	10,300	2	1	0
		2.2	प्लूटो ग्रह का **पंचम चंद्रमा**	204	2	10,700	1	2	1
3	सैटेलाइट मेजरमेंट तकनीक	3	सैटेलाइट मेजरमेंट तकनीक	6	5	3	1	0	n/a
		3.1	**सेटेलाइट** मेजरमेंट तकनीक	21	6	0	n/a	1	0

		3.2	**कृत्रिम उपग्रह मापन तकनीक**	225	8	23	4	4	1
4	पृथ्वी पर सौर सूनामी	4	पृथ्वी पर सौर सूनामी	776	8	77	4	0	n/a
		4.1	**धरती** पर सौर **सुनामी**	2,640	8	214	4	7	4
		4.2	पृथ्वी पर सौर **सुनामी**	4,070	7	339	7	2	1
		4.3	**धरती** पर सौर **सूनामी**	533	7	59	6	0	n/a
5	युवा वैज्ञानिक पुरस्कार	5	युवा वैज्ञानिक पुरस्कार	111,000	8	18,900	8	311	6
		5.1	युवा **विज्ञानी** पुरस्कार	220,000	8	442	4	5	1
		5.2	**यंग साइंटिस्ट अवार्ड**	915	10	71	6	0	n/a
		5.3	**यंग सायंटिस्ट अवार्ड**	102	8	0	n/a	0	n/a
6	देखें विज्ञान विडियो	6	देखें विज्ञान विडियो	46,900	4	11,400	6	162	6
		6.1	देखें विज्ञान **फिल्म**	1,220,000	5	775,000	2	779	2
		6.2	देखें साइंस **मूवी**	40,600	3	482	4	260	2
		6.3	देखें विज्ञान **फ़िल्म**	35,800	8	11,600	5	568	6
7	ब्रह्मोस हाइपरसोनिक मिसाइल परीक्षण	7	ब्रह्मोस हाइपरसोनिक मिसाइल परीक्षण	1,630	9	18	9	0	n/a
		7.1	ब्रह्मोस हाइपरसोनिक मिसाइल **टेस्ट**	536	6	10	9	0	n/a
		7.2	ब्रह्मोस **सुपर**	11,200	9	154	8	1	0

			सोनिक मिसाइल परीक्षण						
8	नासा अंतर्राष्ट्रीय अंतरिक्ष स्टेशन	8	नासा अंतर्राष्ट्रीय अंतरिक्ष स्टेशन	20,400	10	357	10	9	5
		8.1	नासा अंतर्राष्ट्रीय **स्पेस** स्टेशन	17,500	10	190	10	3	1
		8.2	नासा अंतर्राष्ट्रीय **अन्तरिक्ष** स्टेशन	439	9	89	8	0	n/a
		8.3	नासा **इंटरनेशनल** अंतरिक्ष स्टेशन	3,910	8	249	7	1	0
		8.4	नासा **बहुराष्ट्रीय** अंतरिक्ष स्टेशन	336	7	19	6	2	1
Table 3 Queries related to Medical Domain									
1	इन्डियन इंस्टिच्युट हेल्थ एजुकेशन ऐन्ड रिसर्च	1	इन्डियन इंस्टिच्युट हेल्थ एजुकेशन ऐन्ड रिसर्च	1	1	0	n/a	0	n/a
		1.1	**भारतीय संस्थान स्वास्थ्य शिक्षा और शोध**	37,100	5	32,000	1	93	2
		1.2	**इंडियन इंस्टिट्यूट स्वास्थ्य शिक्षा और** रिसर्च	708	3	30,500	3	1	1
		1.3	**इन्डियन संस्थान**	14	2	289	1	0	n/a

			स्वास्थशिक्षा और अनुसंधान						
2	हेल्थ एक्शन प्लान वर्ष 2010	2	हेल्थ एक्शन प्लान वर्ष 2010	937	5	35	7	0	n/a
		2.1	**स्वास्थ्य कार्रवाई योजना साल**2010	364,000	2	7570	3	13	2
		2.2	**स्वास्थ** कार्रवाई योजना **वर्ष**2010	3,430	1	86	2	0	n/a
3	डायबीटीज़ में डिप्रेशन	3	डायबीटीज़ में डिप्रेशन	500	3	13	3	0	n/a
		3.1	**डायबीटीज** में डिप्रेशन	1,090	6	26	4	3	1
		3.2	**डायबिटीज़** में डिप्रेशन	18,900	6	38	5	1	0
		3.3	**मधुमेह** में डिप्रेशन	25,000	7	3,580	8	16	4
4	स्वास्थ संबंधी कार्यक्रम	4	स्वास्थ संबंधी कार्यक्रम	6,210	6	3,750	5	2,443	5
		4.1	**स्वास्थ्य** संबंधी कार्यक्रम	344,000	6	8,700	4	740	4
		4.2	**सेहत** संबंधी कार्यक्रम	56,700	5	1,400	4	133	4
		4.3	**सेहत** संबंधी **प्रोग्राम**	8,180	7	3,510	3	17	1
5	ज्वाहर लाल नेहरू कैंसर अस्पताल	5	ज्वाहर लाल नेहरू कैंसर अस्पताल	7	3	0	n/a	0	n/a
		5.1	नेहरू **कैंसर हॉस्पिटल**	80	10	3	8	0	n/a
		5.2	**पंडित**_जवाहरलाल_नेहरू कैंसर अस्पताल	1,050	1	19	0	0	n/a

		5.3	**नेहरू कैंसर अस्पताल**	37,700	8	15,800	8	31	2
		5.4	नेहरू कैंसर **चिकित्सालय**	3,690	7	64	6	2	1
6	रक्त कैंसर के कारण	6	रक्त कैंसर के कारण	2,100	1	105	2	78	3
		6.1	रक्त **कैंसर** के कारण	213,000	9	8,140	10	99	4
		6.2	**ब्लड** कैंसर के कारण	1,030	1	53	3	6	3
		6.3	**ब्लड कैंसर** के कारण	102,000	9	4,910	10	67	5
7	स्वास्थ्य जानकारी नेटवर्क	7	स्वास्थ्य जानकारी नेटवर्क	470,000	3	6,860	2	293	2
		7.1	स्वास्थ्य **सूचना** नेटवर्क	1,950,000	5	5,860	3	272	1
		7.2	**हेल्थ** सूचना नेटवर्क	27,000	4	11,400	2	18	1
		7.3	स्वास्थ्य जानकारी **संजाल**	8,630	1	4,810	1	28	1
8	हृदय बीमारी इलाज	8	हृदय बीमारी इलाज	153,000	8	5,220	8	349	2
		8.1	दिल बीमारी इलाज	607,000	7	8,490	6	692	3
		8.2	हृदय रोग चिकित्सा	195,000	10	7,450	5	221	4
		8.3	हार्ट डिजीज ट्रीटमेंट	681	9	21	8	2	1

Table 4 Queries related to Tourism Domain

1	टूरिज्म के	1	टूरिज्म के लिए	14,000	9	1,660	7	62	3

	लिए हिमाचल		हिमाचल						
		1.1	**पर्यटन** के लिए हिमाचल	636,000	10	227,000	9	476	8
		1.2	**टुरिज्म** के लिए हिमाचल	126	5	6	4	2	1
		1.3	**टूरिज़म** के लिए हिमाचल	164	8	5	2	0	n/a
2	सेंट्रल टूरिजम डिपार्टमेंट	2	सेंट्रल टूरिजम डिपार्टमेंट	57	7	6	8	0	n/a
		2.1	**केंद्रीय पर्यटन विभाग**	702,000	10	13,700	7	205	5
		2.2	**केंद्रीय टूरिज्म विभाग**	28,700	9	375	7	9	2
3	पर्यटक सूचना केंद्र	3	पर्यटक सूचना केंद्र	2,020,000	9	9,120	8	27	3
		3.1	पर्यटक **जानकारी** केंद्र	919,000	8	10,110	6	97	4
		3.2	**सैलानी** सूचनाकेंद्र	162,000	8	381	4	6	4
		3.4	**सैलानी** सूचना **सेंटर**	14,000	7	143	7	1	0
4	घरेलू टूरिजम इन्फ्रास्ट्रक्चर	4	घरेलू टूरिजम इन्फ्रास्ट्रक्चर	21	6	3	1	0	n/a
		4.1	घरेलू **टूरिज्म इंफ्रास्ट्रक्चर**	1,080	5	23	7	2	1
		4.2	घरेलू **पर्यटन आधारभूत**	4,750	7	405	4	4	3
		4.3	घरेलू **पर्यटन अवसंरचना**	709	7	137	5	3	1
5	टूरिजम के अध्यक्ष	5	टूरिजम के अध्यक्ष	1,410	6	3	5	1	1
		5.1	**पर्यटन** के **अध्यक्ष**	1,150,000	5	66,600	4	2,314	4
		5.2	**टूरिज्म** के	5,170	8	447	6	8	1

			चेयरमैन						
		5.3	**पर्यटन** के **चेयरमैन**	109,000	4	44,700	5	106	4
6	डायरेक्टर टूरिजम दिल्ली	6	डायरेक्टर टूरिजम दिल्ली	287	6	6	4	3	1
		6.1	डायरेक्टर **टूरिज्म नई_दिल्ली**	1,660	7	155	5	9	3
		6.2	**निर्देशक पर्यटन दिल्ली**	287,000	6	10,400	8	78	4
		6.3	**निर्देशक परिभ्रमण** दिल्ली	9,000	2	113	6	11	5
7	नेशनल इंस्टीट्युट टूरिजम	7	नेशनल इंस्टीट्युट टूरिजम	2	2	0	n/a	0	n/a
		7.1	नेशनल इंस्टिट्यूट टूरिजम	7	4	1	0	0	n/a
		7.2	राष्ट्रीय संस्था पर्यटन	362,000	7	28,000	6	289	6
			नेशनल **इंस्टिट्यूट टूरिज्म**	187	5	18	8	0	n/a
8	भारत टूरिजम सेक्टर में रोजगार	8	भारत टूरिजम सेक्टर में रोजगार	187	9	6,440	1	3	1
		8.1	भारत **टूरिज्म** सेक्टर में **रोज़गार**	1,050	8	602	4	0	n/a
		8.2	भारत **पर्यटन क्षेत्र** में **रोज़गार**	59,700	7	79,600	6	26	5
		8.3	भारत **पर्यटन**	592,000	7	98,500	5	436	5

			क्षेत्र में रोजगार						
Table 5 General Search Queries									
1	मुफ़्त शैक्षिक संसाधन	1	मुफ़्त शैक्षिक संसाधन	4,230	4	71	4	1	0
		1.1	**मुफ्त** शैक्षिक संसाधन	31,700	5	565	8	9	5
		1.2	**निशुल्क** शैक्षिक संसाधन	3,450	4	73	7	1	0
		1.3	**निशुल्क** शैक्षिक **साधन**	4500	7	102	6	2	1
		1.4	**निःशुल्क** शैक्षिक संसाधन	2,580	7	84	6	3	2
2	केंद्रीय हिन्दी बोर्ड	2	केंद्रीय हिन्दी बोर्ड	403,000	4	7,430	3	414	3
		2.1	**केन्द्रीय हिंदी परिषद्**	18,000	4	8,630	5	24	3
		2.2	**सेंट्रल हिंदी** बोर्ड	94,900	3	10,200	6	33	4
3	अंताराष्ट्रीय नारी दिवस	3	अंताराष्ट्रीय नारी दिवस	102	2	4	1	0	n/a
		3.1	**अंतर्राष्ट्रीय महिला** दिवस	806,000	9	11,500	6	756	5
		3.2	**इंटरनेशनल वूमन डे**	1,480	2	21	1	1	0
		3.3	**इंटरनैशनल स्त्री** दिवस	916	2	33	2	1	0
4	कन्या गर्भ हत्या	4	कन्या गर्भ हत्या	37,200	9	9,830	6	73	4
		4.1	कन्या **भ्रूण** हत्या	220,000	10	15,400	10	355	4

		4.2	**लड़की पेट** हत्या	519,000	5	7,770	9	181	3
5	वर्ल्ड ट्रेड सेन्टर आतंकवादी हलमा	5	वर्ल्ड ट्रेड सेन्टर आतंकवादी हलमा	0	n/a	0	n/a	0	n/a
		5.1	वर्ल्ड ट्रेड **सेंटर आतंकी** हमला	8,820	10	92	10	12	3
		5.2	वर्ल्ड ट्रेड सेंटर **आतंकी अटैक**	331	8	10	10	1	0
6	फैशन ऐंड स्टाइल	6	फैशन ऐंड स्टाइल	3850	8	51	6	36	6
		6.1	**फ़ैशन और** स्टाइल	176,000	5	47	4	17	6
		6.2	फैशन **एंड** स्टाइल	191,000	10	245	7	591	3
		6.3	**प्रचलन और शैली**	36,600	6	22,330	5	138	2
7	कला और मनोरन्जन	7	कला और मनोरन्जन	3,060	3	37	2	3	1
		7.1	कला और **मनोरंजन**	7,550,000	9	72,110	7	2,364	7
		7.2	**आर्ट्स** और **एंटरटेनमेंट**	8,850	4	42	3	24	3
		7.3	**आर्ट एंड एंटरटेनमेंट**	6,130	6	53	5	11	2

4. Experimental Analysis

As Hindi literature on web is growing on an exponential rate the availability of the same to the end users becomes a prime concern. As already mentioned in our previous work [4] that growth and demand of Hindi users is increasing day by day and various private and government organizations are making their continuous efforts to provide Hindi information to users in India

and abroad as well. Inspite of all such efforts Information retrieval in Hindi language is still suffering due to various factors [2] [3]. All such factors ultimately boil down to a single major problem known as problem of low recall.

In the above tables one can easily observe that problem of low recall can be solved in an effective manner by making variation/s in the query. The variations in the query are dependent on many factors that influence the Hindi search on web. The factors like morphology, phonetics, synonyms and influence of English language on Hindi are major factors which are required to be addressed. An attempt has been made in this investigation to address these factors by including them in Hindi search. A detailed analysis, based on the importance of these factors, has been carried out in this paper. The design and development of the interface for query optimization already discussed in [4] addresses these factors very efficiently and improves the recall for Hindi data on web.

4.1 Observations

In the above table it can be clearly seen that variations of query are generated by making variations in the keywords without changing the meaning of the query and for each variated query a different set of results have been mined out. These variations have less to do with English language but have more impact on Hindi language due to the complexity of Language itself. The tables given above provide lot of information about basic query submitted by the user and the variated query generated by the interface. Because of large number of queries only a few of interest have been picked from tables from each domain to discuss and analyze. The queries picked from various domains are analyzed briefly in the proceeding section.

4.1.1 Agricultural Domain

The original query supplied by the user is भारतीय कृषि संस्थान (*Bhaartiya Krishi Sansthaan*) which means (Indian Agricultural Institutes) in English. The following variations are made to the query through the interface (Variated synonym) भारतीय **किसानी** संस्थान (Variated synonym and English equivalent) **इंडियन** कृषि **संस्था** (Variated two synonyms) भारतीय **खेती संस्था.**

(Variated English equivalents) **इंडियन एग्रीकल्चर इंस्टिट्यूट.** For each variated query a different set of results has been obtained where the meaning of the query remains same. In case of Google for the original query 354,000 results for 1st variation 385,000 results for 2nd 3rd and 4th 80,800, 265,000 and 764 results are obtained which are distinct. The pattern with different figures in quantity of results can be seen for other queries. As far as relevance is concerned it can be clearly seen in the table that for original query 8 out of 10 results are relevant for 1stvariated query 8 out of 10 results are relevant similarly for 2nd 3rd 4thvariated query 7/10, 7/10 and 9/10 results are relevant. Therefore it can be analyzed that by generating the variations of the query not only the quantity but quality of results is also affected.

In another queryकिसान क्रेडिट-कार्ड योज़ना phonetic variations are made to keywords क्रेडिट-कार्ड and योज़ना and its impact on recall can be seen. For original query only 97 results are obtained and the variated query किसान क्रेडिट-कार्ड **योजना** gives 209,000 results similarly the other variations किसान **क्रेडीट-कार्ड** योजना and **कृषक** क्रेडिट कार्ड **योजना** gives 522 and 9080 results which are more than the results obtained against original query.

The figures become more interesting when queries on tested on Guruji Search engine for the above query. Original queryकिसान क्रेडिट-कार्ड योज़ना only **one** result variated queryकिसान क्रेडिट-कार्ड **योजना** 188 results किसान **क्रेडीट-कार्ड** योजना 3 results and **कृषक** क्रेडिट कार्ड **योजना** 6 results.

The optimized query generated by the Interface for original query isकिसान क्रेडिट-कार्ड **योजना**

4.1.2 Science and Technology Domain

The original query supplied by the user isसैटेलाइट मेजरमेंट तकनीक (*saatelaaiet mejermaint takneek*) which means (satellite measurement technique) in English. The following variations are made to the query through the interface (Variated phonetic) **सेटेलाइट** मेजरमेंट तकनीक (Variated synonym and Hindi equivalent) **कृत्रिम उपग्रह मापन तकनीक.** For each variated query a different set of results has been obtained and the meaning of the query remains same. In case of Google for

the original query 6 results for 1st variation 21 results and for 2nd variation 255 results are obtained which are distinct. The pattern with different figures in quantity of results can be seen for other queries. The optimized query generated by the Interface for original query is**कृत्रिम उपग्रह** मेजरमेंट तकनीक

As far as relevance is concerned it can be clearly seen in the table that for original query 5 out of 6 results are relevant for 1st variated query 6 out of 10 results are relevant similarly for 2nd variated query 8/10 results are relevant.

The figures become more interesting when queries on tested on Guruji Search engine for the same query. Original querySैटेलाइट मेजरमेंट तकनीक NO result and variated queries **सेटेलाइट** मेजरमेंट तकनीक and **कृत्रिम उपग्रह मापन तकनीक**1 and 4 results are obtained. This shows how recall improves by inclusion of query variations in Hindi search.

4.1.3 Medical Domain

The original query supplied by the user is डायबीटीज़ में डिप्रेशन (*daayabteez mein dipreshn*) which means (depression during diabetes) in English. The following variations are made to the query through the interface (Variated phonetic) **डायबीटीज** में डिप्रेशन (Variated phonetic) **डायबिटीज़** में डिप्रेशन (Variated synonym) **मधुमेह** में डिप्रेशन. For each variated query a different set of results has been obtained and the meaning of the query remains same. In case of Google for the original query 500 results for 1st variation 1090 results and for 2nd and 3rd variation 18,900 and 25,000 results are obtained which are distinct. The same pattern with different figures in quantity of results can be seen for other queries. The optimized query generated by the Interface for original query is **मधुमेह** में डिप्रेशन.

As far as relevance is concerned it can be clearly seen in the table that for original query 3 out of 10 results are relevant for 1st variated query 6 out of 10 results are relevant similarly for 2nd and 3rd variated queries 6/10 and 7/10 results are relevant. Therefore it can be analyzed that by generating the variations of the query not only the quantity but quality of results is also affected.

Queries when tested on Guruji Search engine. Original query डायबीटीज़ में डिप्रेशन NO result and variated queries **डायबीटीज** में डिप्रेशन, **डायबिटीज़** में डिप्रेशन and **मधुमेह** में डिप्रेशन 3, 1 and 16 results are obtained.

4.1.4 Tourism Domain

The original query supplied by the user isडायरेक्टर टूरिजम दिल्ली (*daayrectr toorizm dilli*) which means (Director of tourism Delhi) in English. The following variations are made to the query through the interface (Variated phonetic) डायरेक्टर **टूरिज्म नई_दिल्ली** (Variated Hindi equivalent and synonym) **निर्देशक पर्यटन** दिल्ली (Variated Hindi equivalent and synonym) **निर्देशक परिभ्रमण** दिल्ली. For each variated query a different set of results has been obtained and the meaning of the query remains same. In case of Google for the original query 287 results for 1st variation 1,660 results and for 2nd and 3rd variation 287,000 and 9,000 results are obtained which are distinct. The pattern with different figures in quantity of results can be seen for other queries. The optimized query generated by the Interface for original query is **निर्देशक पर्यटन** दिल्ली.

As far as relevance is concerned it can be clearly seen in the table that for original query 6 out of 10 results are relevant for 1stvariated query 7 out of 10 results are relevant similarly for 2nd and 3rdvariated queries 6/10 and 2/10 results are relevant. Therefore it can be analyzed that by generating the variations of the query not only the quantity but quality of results is also affected.

By making variations of the queries depending upon various factors as discussed the problem of recall in Hindi IR can be solved up to a great extent. ***It should be noticed that data of similar nature can be mined out regardless the quantity and relevant results can be obtained effectively***. ***Our focus is not only on the quantity of data retrieved but the*** *NATURE* ***of the data retrieved***. In the above analysis of various domains it has been observed that not only recall has improved but the relevant data can also been mined out. Also we discussed in [3] that English language has its impact on Hindi IR and in the above tables we show that inclusion of these English equivalent Hindi keywords improve the recall and relevance up to a certain level.

4.1.6 Proper Nouns (Names of Individuals and Places)

In addition to conducting tests on various queries from different domains a set of experiments have also been conducted to test the affect of phonetics on proper nouns, names of popular individuals e.g. Authors, Politicians, Scientists and names of the places. Through these tests it is observed that phonetics also affect the search results. We show this with examples tested on "Google". The table below shows the results of the tests conducted on search queries.

S.No	Name of Person	Variation Name	Google Results	Name of the Place	Variation Name	Google Results
1	सुभाष चंद्र बोस	सुभाषचन्द्र बोस	212,000 and 7,130	राजस्थान	राजसथान	5,250,000 and 3,880
2	मुंशी प्रेमचंद	मुन्शी प्रेमचन्द	51,400 and 2,600	उत्तर प्रदेश	ऊ.प्र. (Short form) and उत्तर प्रदेष	7,620,000 , 162,000 and 9,480
3	जवाहर नेहरु	जवाहर नेहरू	30,500 and 217,000	अमेरिका	अम्रीका and अमरीका	11,800,000 , 5,600 and 1,090,000
4	उमर अब्दुल्ला	ओमर अब्दुल्ला	213,000 and 38,800	देहली	दिल्ली	56,200 and 15,000,000
5	मदर टेरेसा	मदर टरेसा	37,600 and 621	इंग्लैंड	इंग्लैड इंग्लॅण्ड and इंगलैड	2,180,000 , 15,800, 4,040 and 1,950

Table 6 Effect of phonetics on proper nouns

It can be clearly seen that even search results for proper nouns in Hindi are affected by phonetics. To overcome this problem we have made an attempt to flood our database with proper nouns

which includes names of famous people from different areas and names of the places with their phonetic variants. Through the interface these variants can be easily accessed and scope of search can be further improved.

5. Behavior and comparison of search engines for Hindi Search

In this section we focus on the behavior (Not Working) of search engines for Hindi search. In most cases, morphological variants of words have similar semantic interpretations and can be considered as equivalent for the purpose of IR applications. For this reason, a number of stemming algorithms, or stemmers, have been developed, which attempt to reduce a word to its stem or root form. Thus, the key terms of a query or document are represented by stems rather than by the original words. For IR purposes, it doesn't usually matter whether the stems generated are genuine words or not – thus, "computation" might be stemmed to "comput" – provided that (a) different words with the same 'base meaning' are conflated to the same form, and (b) words with distinct meanings are kept separate. An algorithm which attempts to convert a word to its linguistically correct root ("compute" in this case) is sometimes called a lemmatiser. Examples of products using stemming algorithms would be search engines such as Lycos and Google, and also thesauruses and other products using NLP for the purpose of IR [6].

In the Hindi search it is apparent that Google makes use of word stemming which serves as one of the reasons for retrieval is of documents in very large quantity. A table below shows the listing of keywords by Google for a Hindi keyword.

S.No	Hindi Keyword	Google Listing of keywords
1	योजना	योजना, योजनाओं and योजनाएं.
2	फसल	फसल, फसलों and फसलें
3	पक्षी	पक्षी and पक्षियों
4	समस्या	समस्या and समस्याएं
5	रोग	रोग, and रोगों

Table 7 Listing of keywords in results by Google for a Hindi keyword

However in case of Yahoo and Guruji it is observed that no such use of stemming has been done. Therefore the quantity of documents returned by both search engines is less as compared to Google.

All three search engines do not handle Hindi phonetics. It has been observed that the Hindi phonetic equivalent keywords are treated as different keywords. For example (सेंटर and सेन्टर) (रोज़गार, रोजग़ार and रोजगार) are treated as different keywords. Also the synonyms and other parameters are not handled by the search engines. It has been observed that proper nouns can have their phonetic variants and for each phonetic equivalent proper noun the distinct results can be obtained but all search engines fail to handle this. Guruji underperforms as compared to Google and Yahoo as far as retrieval of quantity of results is concerned. But an improvement has been seen in its performance when query variations are used to obtain results. In all above tables it can be clearly seen that problem of recall has been well addressed and becomes more meaningful in case of Guruji search engine.

It should be noticed that 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. Also no particular standard exists for writing the keyword to fetch Hindi web data. The native Hindi user may not be aware of the Phonetic and other issues in Hindi IR and may miss the relevant information of his/her use. As we mentioned in [4] that ***wrongly transliterated*** keywords when supplied to search engines fetch handful of results. Therefore we also take into consideration the wrongly transliterated keywords. For a wrongly transliterated keyword correct keywords can be fetched from the database and are provided to the end user for their use in search.

The interface addresses all these issues and provides a better platform for Hindi users to search Hindi information on web. ***Almost all phonetic, synonym English equivalent Hindi keywords, phonetic variations of proper nouns and wrongly transliterated keywords converted to correct***

form are at their disposal and the optimized version of the query is suggested to the user so that effective process of Hindi IR can be carried out.

In the above tables for a given query an optimized version of the queries has been generated by the interface. The optimized queries for the basic queries submitted by the user are presented in the table given below. Here in this table we can easily see how queries submitted by the users have been optimized by the interface for generating better results.

S .No	Original User Queries	Optimized Queries Generated by Interface
1	हिंदुस्तान खेती क्षेत्र में बैंक	भारत **कृषि** क्षेत्र में बैंक
2	चावल की पैदावार बढ़ाने की तरकीब	चावल की **फसल** बढ़ाने की उपाय
3	भारतीय कृषि संस्थान	भारतीय कृषि संस्थान
4	क्रॉप इंश्योरेंस पॉलिसी	फसल इंश्योरेंस **योजना**
5	मवेशियों का चुनाव	मवेशियों का **चयन**
6	किसान क्रेडिट-कार्ड योज़ना	किसान क्रेडिट-कार्ड **योजना**
7	कपास फ़सल बचाव	कपास **फसल रक्षा**
8	किसान के हित में सरकार	किसान के हित में **गवर्नमेंट**
9	विज्ञानं साहित्य का प्रकाशन	**विज्ञान** साहित्य का प्रकाशन
10	प्लूटो ग्रह का पाँचवाँ चाँद	प्लूटो ग्रह का पाँचवाँ **चंद्रमा**
11	सैटेलाइट मेजरमेंट तकनीक	**कृत्रिम उपग्रह** मेजरमेंट तकनीक
12	पृथ्वी पर सौर सूनामी	पृथ्वी पर सौर **सुनामी**
13	युवा वैज्ञानिक पुरस्कार	युवा **साइंटिस्ट अवार्ड**
14	देखें विज्ञान विडियो	देखें विज्ञान विडियो
15	ब्रह्मोस हाइपरसोनिक मिसाइल परीक्षण	ब्रह्मोस हाइपर सोनिक मिसाइल परीक्षण

16	नासा अंतर्राष्ट्रीय अंतरिक्ष स्टेशन	नासा अंतर्राष्ट्रीय **स्पेस** स्टेशन
17	इन्डियन इंस्टिच्युट हेल्थ एजुकेशन ऐन्ड रिसर्च	भारतीय **संस्थान** हेल्थ एजुकेशन **और** रिसर्च
18	हेल्थ एक्शन प्लान वर्ष 2010	हेल्थ एक्शन **योजना** वर्ष 2010
19	डायबीटीज़ में डिप्रेशन	**मधुमेह** में डिप्रेशन
20	स्वास्थ संबंधी कार्यक्रम	**हेल्थ** संबंधी कार्यक्रम
21	ज्वाहर लाल नेहरू कैंसर अस्पताल	**नेहरूकेंसर** अस्पताल
22	रक्त कैंसर के कारण	रक्त कैंसर के कारण
23	स्वास्थ्य जानकारी नेटवर्क	**हेल्थ सूचना** नेटवर्क
24	हृदय बीमारी इलाज	हृदय बीमारी **उपाय**
25	मुफ़्त शैक्षिक संसाधन	**मुफ्त** शैक्षिक संसाधन
26	केंद्रीय हिन्दी बोर्ड	**सेंट्रल भारतीय** बोर्ड
27	अंताराष्ट्रीय नारी दिवस	अंतर्राष्ट्रीय **महिला** दिवस
28	कन्या गर्भ हत्या	कन्या **भ्रूण** हत्या
29	वर्ल्ड ट्रेड सेन्टर आतंकवादी हलमा	वर्ल्ड ट्रेड **सेंटर आतंकी** हलमा
30	फैशन ऐंड स्टाइल	फैशन **और** स्टाइल
31	कला और मनोरन्जन	कला और **मनोरंजन**
32	टूरिज्म के लिए हिमाचल	**पर्यटन** के लिए हिमाचल
33	सेंट्रल टूरिजम डिपार्टमेंट	सेंट्रल **पर्यटन विभाग**
34	पर्यटक सूचना केंद्र	पर्यटक सूचना **सेंटर**
35	घरेलू टूरिजम इन्फ्रास्ट्रक्चर	घरेलू टूरिजम **आधारभूत**
36	टूरिजम के अध्यक्ष	टूरिजम के **चेयरमैन**

37	डायरेक्टर टूरिजम दिल्ली	**निर्देशकपर्यटन** दिल्ली
38	नेशनल इंस्टीट्युट टूरिजम	नेशनल इंस्टीट्युट टूरिजम
39	भारत टूरिजम सेक्टर में रोजगार	भारत **पर्यटन क्षेत्र** में रोजगार

Table 8 Optimized queries generated by the interface

6. Limitations/ Drawbacks of the Interface

The interface supported by large scale database (including morphological, phonetic variants, synonyms, English equivalents of Hindi keywords, phonetic variants proper nouns and correct variants of wrongly transliterated keywords) designed for Hindi search has benefited the users to pursue Hindi search. It has been observed that problem of recall has been solved up to a great extent and relevant results can also be mined out by making query variations. Interface also generates optimized queries for search suggestions so that users can choose correct phonetics, synonyms and English equivalents of Hindi keywords etc.

Hindi language is a rich language with multiple synonyms of one word which leads to ambiguity. Interface works well for all parameters including synonyms but in certain cases due to ambiguous nature of Hindi language wrong optimized query is generated. The problems arise for those keywords whose synonyms are not closely related. In this section we show how this problem affects the process of generating optimized query. In the table below we present some queries for which wrong optimized queries can be generated by the interface.

S.No		Synonyms Variations	Optimized Queries
1	मुम्बई मधुशाला पाबंदी	:मधुशाला:मदिरालय:मधुशाला:मद्यशाला: शराब_घर:शराबघर:मयख़ाना:शराबख़ाना:पा नागार:**शराबखाना**:मयखाना:सुरागार:बार: आपान:	मुम्बई शराबखाना पाबंदी

1.1	हरिवंशराय बच्चन कविता मधुशाला	शराबखाना	हरिवंशराय बच्चन कविता शराबखाना
2	विरासतपर अनैतिकअधिकार	अधिकार::अधिकार:वश:काबू:क़ाबू:हक़:हक:**क़ब्ज़ा**:**कब्जा**:आधिपत्य:अख़्तियार:अख्तियार:इख़्तियार:इख्तियार:ज़ोर:दावा:संरक्षण:इमकान:	विरासत पर अनैतिक कब्जा
2.1	समानता का अधिकार	कब्जा	समानता काकब्जा
3	लोक मंगल एवं लोक कल्याण	मंगल:हित:कल्याण:फ़ायदा:फायदा:भला:मंगल:भलाई:सलामती:भला:शुभ:हित:स्वस्ति:भद्र:मंगलवार:मंगल:भौमवार	लोक कल्याणएवं लोक हित
3.1	नायक मंगल पांडे	कल्याण	नायक कल्याण पांडे

Table 9 Drawback of Interface: wrong optimized queries

From the above table it is clear that in certain cases query optimization suffers due to ambiguity involved in the Hindi language because of the multiple synonyms for a particular word. The problem has a very adverse impact on query optimization as the meaning of optimized query gets changed completely and becomes completely irrelevant. This problem occurs for certain keywords and not all keywords. There are various other keywords listed below for which this problem in query optimization does not occur because of close relation between the synonyms.

S .No	Keywords	Closely related synonyms
1	योग्य	उपयुक्त:काबिल:समर्थ:हुनरमंद:उदात्त: सलीकामंद:

2	अनाथ	:अनाथ:यतीम:लावारिस:बेकस:छेमंड:बैतला:मुरहा:निगोड़ा_नाथा:ओर्फन:ओर्फन_चिल्ड्रन:ओर्फन_चाइल्ड:ओरफन:
3	अपमान	:अपमान:अनादर:बेइज्जती:निरादर:तिरस्कार:असत्कार:असन्मान:हेठी:तौहीन:तोहीनी:जिल्लत:ज़िल्लत:फ़ज़ीहत:अवमान:अवमानना:अवमानन:मानध्वंस::गञ्जन:इन्सुल्ट:इंसुल्ट:and more
4	रक्तदान	:रक्तदान:रुधिरदान:रक्त_दान:रुधिर_दान:ब्लड_डोनेशन:
5	अनुपस्थित	:अनुपस्थित:ग़ैरहाज़िर:ग़ैरमौज़ूद:नदारद:नदारत:अविद्यमान:अप्रस्तुत:अप्राप्त:अवर्त्तमान:अवर्तमान:अविद्य:मिस्सिंग:अब्सेंट:एब्सेंट: and more

Table 10: Closely related synonyms

Closely related synonyms do not affect the query optimization but there are many such keywords with synonyms not related closely and create ambiguity in a sentence. The drawback is serious and it needs to be addressed. In this paper drawback of the interface has been addressed but not resolved and hence becomes one important issue to be resolved in the future research work.

7. Conclusion

In our work we addressed certain factors/parameters that are responsible for low recall in Hindi Language and found that the problem of recall can be solved by optimizing the Hindi query at interface level. The Query optimizing interface handles all these issues and thus solves the problem of low recall in Hindi search.

In our database Keywords are provided with their morphological, phonetic, synonym, English equivalent Hindi variants. We also include wrongly transliterated keywords and their correct forms. Database also includes keywords related to various domains and proper nouns (names of famous persons and places) with their phonetic equivalents.

The interface has been developed to provide wide range options to the users to choose correct keyword against the keyword supplied by him/her which saves time and effort and also gives them ability to search variety of information without changing the basic nature/meaning of their query. The optimized query is further suggested to the user to use as it contains optimized keywords. Query optimizing Interface helps users to mine the Hindi information from web and

hence chances of retrieving relevant information are increased. From our experiments presented in this paper we show that query optimization solves the problem of low recall for Hindi IR up to a great extent**.** The present system suffers from serious drawbacks. The limitations of the system shall be taken care of in our future work.

==========

References

[1] GirirajAgarwal"*Indian Languages on the Internet*" Article Source http://span.state.gov/wwwfspseptoct0948.pdf

[2] Kumar Sourabh and Vibhakar Mansotra "*Factors Affecting the Performance of Hindi Language searching on web: An Experimental Study*". Department of Computer Science and IT, University of Jammu J&K 180001. INDIA International Journal of Scientific & Engineering Research Volume 3, Issue 4, April-2012 ISSN 2229-5518

[3] Kumar Sourabh and Vibhakar Mansotra "*An Experimental Analysis on the Influence of English on Hindi Language Information Retrieval*". Department of Computer Science and IT, University of Jammu J&K 180001. INDIA International Journal of Computational Linguistics Research Volume 2 Number 3/4 Sep/Dec 2011

[4] Kumar Sourabh and Vibhakar Mansotra "*Query Optimization: A Solution for Low Recall Problem in Hindi Language Information Retrieval*". Department of Computer Science and IT, University of Jammu J&K 180001. INDIA LANGUAGE IN INDIA Strength for Today and Bright Hope for Tomorrow Volume 12: 11 November 2012 ISSN 1930-2940

[5] Juxt Consult India Online Survey Source: www.digitaltribe.in/digi-data/eMarketer_India_Online.pdf

[6] Article Information retrieval and stemming Source: http://www.comp.lancs.ac.uk/computing/research/stemming/general/

==========

Kumar Sourabh, Ph.D.
Department of Computer Science and IT
University of Jammu
J&K 180001 INDIA
Kumar9211.sourabh@gmail.com

Prof. Vibhakar Mansotra
Department of Computer Science and IT
University of Jammu
J&K 180001 INDIA
Vibhakar20@yahoo.co.in

Rakesh Goswami, Research Scholar
Department of Computer Science and IT
University of Jammu
J&K 180001 INDIA
rahulgoswami95@gmail.com