LANGUAGE IN INDIA
Strength for Today and Bright Hope for Tomorrow
Volume 12 : 11 November 2012
ISSN 1930-2940

Query Optimization: A Solution for Low Recall Problem in Hindi Language Information Retrieval

Kumar Sourabh
Vibhakar Mansotra

==

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. Whatever they may have been, over the years, many of the motivations for an almost exclusive focus on English as the language of choice in IR have lost their validity. 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 (after English and Mandarin): 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 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, and lack of standards for spelling words. Taking into consideration the aforesaid issues we introduce Hindi Query Optimization technique (design and development) which solved the problem of recall up to a great extent.

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

1. Introduction

The World Wide Web, or simply the *web* may be seen as a huge collection of documents freely produced and published by a very large number of people, without any solid editorial control. This is probably the most democratic – and anarchic –widespread means for anyone to express feelings, comments, convictions and ideas, independently of ethnics, sex, religion or any other characteristic of human societies. The web constitutes a comprehensive, dynamic, up-to-date repository of information regarding most of the areas of human knowledge; and, it supports an increasingly important part of commercial, artistic, scientific and personal transactions, which gives rise to a very strong interest from individuals, as well as from institutions, at a universal scale. However, the web also exhibits some characteristics that are adverse to the process of collecting information from it in order to satisfy specific needs; some of the characteristics are: the large volume of data it contains, its dynamic nature, constituted by unstructured or semi-structured data, content and format heterogeneity and irregular data quality. End-users also introduce some additional difficulties in the retrieval process. Information needs are often imprecisely defined, generating a semantic gap between user needs and their specifications. The satisfaction of a specific information need on the web is supported by search engines and other tools, aimed at helping users to gather information from the web.

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. Over the years other languages have made some inroads in IR. The Internet shows more inclination toward the use of plurality of languages, as the non- English content is growing rapidly. Asia is the largest and the most culturally and linguistically diverse continent. It covers 39 million square kilometers, about 60% of land area of the world, and has an estimated 3.8 billion population, which is approximately 60% of the world's population. There are more than 50 countries and roughly 2200 languages spoken in Asia.

Hindi is the third most widely-spoken language in the world (after English and Mandarin): an estimated 500-600 million people speak this language. A direct descendant of Sanskrit through Prakrit and Apabhramsha, Hindi belongs to the Indo-Aryan group of languages, a subset of the Indo-European family. Rise of Hindi, Urdu and other Indian languages on the Web, has lead millions of non-English speaking Indians to discover uses of the Internet in their daily lives. 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. Two American IT companies, Microsoft and Google, have played a big role in making this possible.

A decade ago, there were many problems involved in using Indian languages on the Internet. "There was mismatch of fonts and keyboard layouts, which made it impossible to read any Hindi document if the user did not have the same fonts .There was chaos, more than 50 fonts and 20 types of keyboards were being used and if two users were following different styles, there was no way to read the other person's documents. But the advent of Unicode support for Hindi and Urdu has changed the scenario.

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. This space has become so crowded right now that it is difficult to know who is really winning. However, we attempt to put forth a brief overview of the current scenario.

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*.

Taking into consideration the aforesaid issues we introduce Hindi Query Optimization technique as a database-oriented approach by bringing morphological variants, spelling variations, synonyms and English equivalent Hindi words under one platform. The data base can be accessed via an Interface which serves as an input platform for user queries. The query entered by user is then fed to database to fetch the Morphological variants, spelling variations, Synonyms and English- equivalent Hindi words. The rephrased variations of the query generated by the interface are then fed to search engine/s via interface to obtain search results. The interface uses database as backend for matching and retrieval of Hindi keywords. Search engines like Google, Yahoo, Bing and Guruji can be used as selections for information retrieval which makes the interface as Meta search platform. The queries supplied by the user are saved in query log which is a separate database used for processing the keywords for their further optimization. To accomplish this purpose we used the keyword ranking and explicit relevance feed-back method. A Hindi keyboard and transliterator has also been provided for query input.

The interface addresses all the monolingual search issues and provides a better platform for Hindi users to search Hindi information on web. Query optimization and Interface 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. The interface provides wide range of options to the users to choose correct keyword against the keyword supplied by him/her which saves time and effort and also gives the ability to search variety of information without changing the basic nature/meaning of their query. Interface helps users to mine the Hindi information from web and hence chances of retrieving relevant information can be increased.

2. A brief literature review

As far as development in IR with respect to Indian languages is concerned, a lot work is going on particularly in the field of information retrieval. Research is also going on in other related areas as well such as NLP machine translation etc. Various regional languages have been taken into consideration by researchers for IR. Even government organization like TDIL (Technology Development for Indian Languages) has made significant contributions for standardization of Indian Languages on the web. In the following section we present the various developments in Indian IR and NLP system.

2.1 Developments in Indian Language IR system

2.1.1 Bengali and Hindi to English CLIR

Debasis Mandal, Mayank Gupta, Sandipan Dandapat, Pratyush Banerjee, and Sudeshna Sarkar Department of Computer Science and Engineering IIT Kharagpur, India presented a cross-language retrieval system for the retrieval of English documents in response to queries in Bengali and Hindi, as part of their participation in CLEF1 2007 Ad-hoc bilingual track. They followed the dictionary-based Machine Translation approach to generate the equivalent English query out of Indian language topics. [2]

2.1.2 Hindi and Marathi to English Cross Language Information Retrieval

Manoj Kumar Chinnakotla, Sagar Ranadive, Pushpak Bhattacharyya and Om P. Damani Department of CSE IIT Bombay presented Hindi -> English and Marathi->English CLIR systems developed as part of their participation in the CLEF 2007 Ad-Hoc Bilingual task. They took a query translation-based approach using bi-lingual dictionaries. Query words not found in the dictionary are transliterated using a simple rule-based approach which utilizes the corpus to return the 'k' closest English transliterations of the given Hindi/Marathi word. The resulting multiple translation/transliteration choices for each query word are disambiguated using an iterative page-rank style algorithm, which based on term-term co-occurrence statistics, produces the final translated query. [3].

2.1.3 Hindi and Telugu to English Cross Language Information Retrieval

Prasad Pingali and Vasudeva Varma Language Technologies Research Centre IIIT, Hyderabad presented the experiments of Language Technologies Research Centre (LTRC) as part of their participation in CLEF 2006 ad-hoc document retrieval task. They focused on Afaan Oromo, Hindi and Telugu as query languages for retrieval from English document collection and contributed to Hindi and Telugu to English CLIR system with the experiments at CLEF. [4]

2.1.4 English-Hindi CLIR

Tan Xu and Douglas W.Oard College of Information Studies and CLIP Lab, Institute for Advanced Computer Studies, University of Maryland. Forum for Information Retrieval Evaluation (FIRE), the University of Maryland participated in the Ad-hoc task cross-language document retrieval task, with English queries and Hindi documents. Their experiments focused on evaluating the effectiveness of a “meaning matching” approach based on translation probabilities [5].

2.1.5 English to Kannada / Telugu Name Transliteration in CLIR

Mallamma v reddy, Hanumanthappa Department of Computer Science and Applications, Bangalore University, They present a method for automatically learning a transliteration model from a sample of name pairs in two languages. However, they faced the problem of translating Names and Technical Terms from English to Kannada/Telugu. [6].

2.1.6 Kannada and Telugu Native Languages to English Cross Language Information Retrieval

Mallamma V. Reddy, Hanumanthappa Department of Computer Science and Applications, Bangalore University conducted experiments on translated queries. One of the crucial challenges in cross lingual information retrieval is the retrieval of relevant information for a query expressed in a native language. While retrieval of relevant documents is slightly easier, analyzing the relevance of the retrieved documents and the presentation of the results to the users are not trivial tasks. To accomplish the above task, they present their Kannada English and Telugu English CLIR systems as part of Ad-Hoc Bilingual task by translation based approach using bi-lingual dictionaries. [7]

2.1.7 Bilingual Information Retrieval System for English and Tamil

Dr.S.Saraswathi, Asma Siddhiqaa.M, Kalaimagal.K, Kalaiyarasi.M address the design and implementation of BiLingual Information Retrieval system on the domain, Festivals. A generic platform is built for BiLingual Information retrieval which can be extended to any foreign or Indian language working with the same efficiency. Search for the solution of the query is not done in a specific predefined set of standard languages, but is chosen dynamically on processing the user's query. Their research deals with Indian language Tamil apart from English. [8].

2.1.8 Recall Oriented Approaches for improved Indian Language Information Access

Pingali V.V. Prasad Rao Language Technologies Research Centre International Institute of Information Technology Hyderabad:

Their research is an investigation into Indian language information access. The investigation shows that Indian language information access technologies face severe recall problem when using conventional IR techniques (used for English-like languages). During this investigation they crawled the web extensively for Indian languages, characterized the Indian language web and in the process came up with some solutions for the low recall problem. [9]

2.1.9 English Bengali Ad-hoc Monolingual Information Retrieval Task Result at FIRE 2008

Sivaji Bandhyopadhyay, Amitava Das, Pinaki Bhaskar Department of Computer Science and Engineering Jadavpur University, Kolkata.

Their experiments suggest that simple TFIDF based ranking algorithms with positional information may not result in effective ad-hoc mono-lingual IR systems for Indian language queries. [10]

2.1.10 Using Morphology to Improve Marathi Monolingual Information Retrieval

Ashish Almeida, Pushpak Bhattacharyya IIT Bombay. They study the effects of lexical analysis on Marathi monolingual search over the news domain corpus (obtained through FIRE-2008) and observe the effect of processes such as lemmatization, inclusion of suffixes in indexing and stop-words elimination on the retrieval performance. Their results show that lemmatization significantly improves the retrieval performance of languages like Marathi which is agglutinative in nature. [11].

2.1.11 Om: One tool for many (Indian) languages

Ganpathiraju, Madhavi, Balakraishnan, Mini Balakrishnan, N., Reddy Raj (Language Technologies Institute, Carnegie Mellon University, Pittsburgh) (Supercomputer Education and Research Centre, Indian Institute of Science, Bangalore 560 012, India)

They describe the development of a transliteration scheme Om which exploits this phonetic nature of the alphabet. Om uses ASCII characters to represent Indian language alphabets, and thus can be read directly in English, by a large number of users who cannot read script in other Indian languages than their mother tongue. It is also useful in computer applications where local language tools such as email and chat are not yet available. Another significant contribution presented in their research is the development of a text editor for Indian languages that integrates the Om input for many Indian languages into a word processor such as Microsoft WinWord. The text editor is also developed on Java platform that can run on UNIX machines as well. They propose this transliteration scheme as a possible standard for Indian language transliteration and keyboard entry [12].

2.1.12 Post Translation Query Expansion using Hindi Word-Net for English-Hindi CLIR System

Sujoy Das, Anurag Seetha, M. Kumar, and J.L. Rana have investigated impact of query expansion using Hindi WordNet in the context of English-Hindi CLIR system. The WordNet is a lexical database, machine readable thesaurus for Hindi language. They have translated English query using Shabdanjali dictionary. The translated queries have been expanded using Hindi WordNet and nine query expansion strategies have been formulated. In these runs title field of topic was used for query formulation and expansion and in one run title + description field was used for query formulation and expansion. The queries are translated, then expanded and are submitted to the retrieval system to retrieve documents from the Fire Hindi Test collection. Their observations suggest that simple query expansion using Hindi WordNet is not effective for English- Hindi CLIR system [13].

2.2 Machine Translation in India

Although Translation in India is old, Machine Translation is comparatively young. Earlier efforts in this field have been noticed since 1980, involving different prominent Institutions such as **IIT** Kanpur, **University of Hyderabad, NCST** Mumbai and **CDAC** Pune. During late 1990 many new projects initiated by **IIT** Mumbai, **IIIT** Hyderabad, **AU-KBC** Centre, Chennai and **Jadavpur University,** Kolkata were undertaken. **TDIL** has started a consortium mode project since April 2008, for building computational tools and Sanskrit-Hindi MT under the leadership of Amba Kulkarni (University of Hyderabad). The goal of this Project is to build children's stories using multimedia and e-learning content.

2.2.1 Anglabharati

IIT Kanpur has developed the Anglabharti Machine Translator technology from English to Indian languages under the leadership of Prof. R.M.K Sinha. It is a rule-based system and has approximately 1750 rules, 54000 lexical words divided into 46 to 58 paradigms. It uses pseudo Interlingua named as PLIL (Pseudo Lingua for Indian Language) as an intermediate language.

The architecture of **Anglabharti** has six modules: Morphological analyzer, Parser, Pseudo code generator, Sense disambiguator, Target text generators, and Post-editor. The Hindi version of Anglabharti is **AnglaHindi** which is web- based application which is also available for use at http://anglahindi.iitk.ac.in. To develop automated translator system for regional languages, Anglabharti architecture has been adopted by various Indian institutes for example, IIT Guwahati.

2.2.2 Anubharti

Prof. R.M.K. Sinha developed **Anubharti** during 1995 at IIT Kanpur. **Anubharti is** based on hybridized example-based approach. The Second phase of both the projects (Anglabharti II and Anubharti II) has started from 2004 with new approaches and some structural changes.

2.2.3 Anusaaraka

Anusaaraka is a Natural Language Processing (NLP) Research and Development project for Indian languages and English undertaken by CIF (Chinmaya International Foundation). It is fully-automatic general-purpose high-quality machine translation systems (FGH-MT). It has software which can translate the text of any Indian language(s) into another Indian Language(s), based on Panini Ashtadhyayi (Grammar rules). It is developed at the International Institute of Information Technology, Hyderabad (IIIT-H) and Department of Sanskrit Studies, University of Hyderabad.

2.2.4 Mantra

Machine Assisted Translation Tool (**Mantra**) is a brain child of Indian Government during 1996 for translation of Government orders, notifications, circulars and legal documents from English to Hindi. The main goal was to provide the translation tools to government agencies. Mantra software is available in all forms such as desktop, network and web based. It is based on **Lexicalized Tree Adjoining Grammar (LTAG)** formalism to represent the English as well as the Hindi grammar. Initially, it was domain specific such as Personal Administration, specifically Gazette Notifications, Office Orders, Office Memorandums and Circulars, gradually the domains were expanded. At present, it also covers domains like Banking, Transportation and Agriculture etc. Earlier Mantra technology was only for English to Hindi translation, but currently it is also available for English to other Indian Languages such as Gujarati, Bengali and Telugu. **MANTRA-Rajyasabha** is a system for translating the parliament proceedings such as papers to be laid on the Table [PLOT], Bulletin Part-I, Bulletin Part- II, List of Business [LOB] and Synopsis. Rajya Sabha Secretariat of Rajya Sabha (the upper house of the Parliament of India) provides funds for updating the **MANTRARajyasabha** system.

2.2.5 UNL-based MT System between English Hindi and Marathi

IIT Bombay has developed the Universal Networking Language (UNL) based machine translation system for English to Hindi Language**. UNL** is United Nations project for developing the Interlingua for world's languages. **UNL**-based machine translation is being developed under the leadership of Prof. Pushpak Bhattacharya IIT Bombay.

2.2.6 English-Kannada MT System

Department of Computer and Information Sciences of Hyderabad University has developed an **English-Kannada MT system**. It is based on the transfer approach and Universal Clause Structure Grammar (UCSG).This project is funded by the Karnataka Government and it is applicable in the domain of government circulars

2.2.7 SHIVA and SHAKTI MT

Shiva is an Example-based system. It provides the feed-back facility to the user. Therefore, if the user is not satisfied with the system generated, translated sentence, then the user can provide the feedback of new words, phrases and sentences to the system and can obtain the newly interpretive translated sentence. Shiva MT system is available at (http://ebmt.serc.iisc.ernet.in/mt/login.html).

Shakti is a statistical approach based rule-based system. It is used for the translation of English to Indian languages (Hindi, Marathi and Telugu). Users can access the **Shakti** MT system at (http://shakti.iiit.net).

2.2.8 Tamil-Hindi MAT System

K B Chandrasekhar Research Centre of Anna University, Chennai has developed the machine-aided Tamil to Hindi translation system. The translation system is based on ***Anusaaraka*** Machine Translation System and follows lexicon translation approach. It also has small sets of transfer rules. Users can access the system at http://www.aukbc.org/research_areas/nlp/demo/mat/

2.2.9 Anubadok

Anubadok is a software system for machine translation from English to Bengali. It is developed in Perl programming language which supports processing of Unicode encoded and text for text manipulations. The system uses the Penn Treebank annotation system for part-of-speech tagging. It translates the English sentence into Unicode based Bengali text. Users can access the system at http://bengalinux.sourceforge.net/cgibin/anubadok/index.pl

2.2.10 Punjabi to Hindi Machine Translation System

During 2007, Josan and Lehal at the Punjab University, Patiala, designed Punjabi to Hindi machine translation system. The system is built on the paradigm of foreign machine translation systems such as RUSLAN and CESILKO. The system architecture consists of three processing modules: Pre Processing, Translation Engine and Post Processing.

2.3 Contribution of Private Companies in Evolving the ILT – Indian language Search Engine

2.3.1 Guruji

Guruji.com is the first Indian language search engine founded by the two IIT Delhi graduates Anurag Dod and Gaurav Mishra, assisted by the Sequoia Capital. guruji.com uses crawl technology, based on propriety algorithms. For any query, it goes into Indian languages contents deep and tries to return the appropriate output. Guruji search engine covers a range of specific content news, entertainment, travel, astrology, literature, business, education and more.

2.3.2 Google

Internet searching giant Google also supports major Indian Languages such as Hindi, Bengali, Telugu, Marathi, Tamil, Gujarati, Kannada, Malayalam, and Punjabi and also provides the automated translation facility from English to Indian Languages. Google Transliteration Input Method Editor is currently available for different languages such as Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Nepali, Punjabi, Tamil, Telugu and Urdu.

2.3.3 Microsoft Indic Input Tool

Microsoft has developed the Indic Input Tool for Indianization of computer applications. The tool supports major Indian languages such as Bengali, Hindi, Kannada, Malayalam, Tamil and Telugu. It is based on a syllable-based conversion model. WikiBhasa is Microsoft multilingual content creation tool for translating Wikipedia pages into multilingual pages. So, source language in WikiBhasa will be English and Target language can be any Indian local language(s).

2.3.4 Webdunia

Webdunia is an important private player which assists the development of Indian language technology in different areas such as text translation, software Localization, and Website localizations. It is also involved in research and development of Corpus creation/collection, and Content Syndication. Moreover, it provides the facility of language consultancy. It has developed various applications in Indian Languages such as My Webdunia, Searching, Language Portals, 24 Dunia, Games, Dosti, Mail, Greetings, Classifieds, Quiz, Quest, Calendar etc.

2.3.5 Modular InfoTech

Modular InfoTech Pvt. Ltd. is a pioneer private company for development of Indian Languages software. It provides the Indian language enablement technology to many state governments and central government in e-governance programs. It has developed the software for multilingual content creation for publishing newspapers and also has developed the qualitative Unicode based Fonts for major Indian languages. It has specifically developed the Shree-Lipi Gurjrati pacakage for the Gujarati language which is useful in DTP sector, corporate offices and e-Governance program of the Government of Gujarat.

2.4 Government Effort for Evolving Language Technology

Indian government was aware about this fact. Since 1970, the Department of Electronics and the Department of Official Language were involved in developing the Indian language Technology. Consequently ISCII (Indian Script Code for Information Interchange) is developed for Indian languages on the pattern of ASCII (American Standard Code for Information Interchange). Also "**I**ndian languages **Trans**literation" (**ITRANS**) developed by Avinash Chopde and ITRANS represents Indian language alphabets in terms of ASCII (Madhavi et al, 2005). The Department of Information Technology under Ministry of Communication and Information Technology is also putting the efforts for proliferation of Language Technology in India, And other Indian government ministries, departments and agencies such as the Ministry of

Human Resource, DRDO (Defense Research and Development Organization), Department of Atomic Energy, All India Council of Technical Education, UGC (Union Grants Commission) are also involved directly and indirectly in research and development of Language Technology. All these agencies help develop important areas of research and provide funds for research to development agencies. As an end-result IndoWordNet was developed for the Indian languages on the pattern of English WordNet.

2.4.1 TDIL Program

Government of India launched TDIL (Technology Development for Indian Language) program. TDIL decides the major and minor goal for Indian Language Technology and provide the standard for language technology TDIL journal ***Vishvabharata*** **(Jan 2010)** outlined short-term, intermediate, and long-term goals for developing Language Technology in India.

From the above literature study it can be clearly analyzed that a good amount of research work has been done and is still going on in the field of CLIR, NLP etc but at the same time very much less work has been done in the field of monolingual IR for Hindi language in particular. It seems that Hindi language which is the national language of India and widely used worldwide, has not been given much importance.

Indian Search engines like Guruji, Raftaar etc. are now present for Hindi IR but the monolingual issues are not well addressed by any of them. The objective of our research work is to highlight various issues involved in monolingual IR and suggest ways and means to solve those issues through the design and development of a specialized tool which will take care of such issues. [14]

3. Issues in Information Retrieval for Hindi Language

The preliminary investigation into typical information access technologies by applying present day popular Techniques show a severe problem of low recall while accessing information using Indian language queries. For instance, many times popular web search engines such as Google, Yahoo and Guruji result in `0' search results for Indian language queries giving an impression that no documents containing this information exist. In reality these search engines face a low recall problem while dealing with Indian languages. Table 1 illustrates a few such cases. For example, a Hindi query for “world trade center aatank-waadi hamlaa” “वर्ल्ड ट्रेड सेन्टर आतंकवादी हलमा “ “Terrorist attack on World Trade Center” is shown to result in `0' documents in table 1, however a small rephrasing of the query in table 2 shows that these keywords exist in second search result. But just saying we have a recall problem may not be sufficient. The next obvious question that follows would be `how much is it a problem?’

Table 1: Problems faced while search in Hindi / Low recall

वर्ल्ड ट्रेड **सेंटर आतंकी** हमला	8,820	92	12
वर्ल्ड ट्रेड सेंटर **आतंकीअटैक**	331	10	1
इंडियनइंस्टिट्यूटस्वास्थ्य शिक्षा और रिसर्च	708	50	1
भारतीय संस्थान स्वास्थ्य शिक्षा और शोध	37,100	7400	93

Table 2: Improved Recall

Keyword	Spelling Variant	English Equivalent	Synonym

		expected	Mostly Used
आरक्षण	आरक्षन	रिजर्वेशन/ रिज़र्वेशन	
लाभ	N/A	N/A	फायदा/फ़ायदा

4. Factors responsible for low recall in Hindi information retrieval

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 such content. Information Retrieval in Hindi language is getting popularity and IR systems face low recall if existing systems are used as-is. Some characteristics which 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. We conducted many experiments to show importance of these parameters in Hindi information searching on web [15] [16].

Relevant information can be mined out by transforming the Hindi queries. Search engines neither make transformations of the query nor find keyword equivalents. We present an interface to the search engine called Hindi Query Optimizer which helps improve low recall in Hindi IR. In this paper we focus on the Design and development of the Hindi Query Optimizer and show how the recall problem for Hindi Language is solved up to a certain level for Monolingual IR.

5. The Hindi Query Optimization

We introduce Hindi Query Optimization technique as a database oriented approach by bringing morphological variants, spelling variations, synonyms and English equivalent Hindi words under one large scale database. The data base can be accessed via an Interface which serves as an input platform for user queries. The query entered by user is then fed to database to fetch the Morphological variants, spelling variations, Synonyms and English equivalent Hindi words. The variations of the query generated by the interface are then fed to search engine/s via interface to obtain search results. Users provide input to search systems at their own convenience. No particular standard is followed for writing Hindi on web. Hindi is India's official language which is further under the influence of regional and foreign languages particularly English [16] and this result in synonym and spelling variation of Hindi keywords. Below we present an example table 3 that shows how different results of the same nature can be obtained by making variations of Hindi query based on the above factors. For Query आरक्षण से लाभ.

Table 3: Query organization

S.No	Hindi Query	Google Results
1	आरक्षण से लाभ	891,000
2	आरक्षन (Spelling Variation) से लाभ	31
3	रिजर्वेशन (Reservation)से लाभ	24,900
4	रिज़र्वेशन (Reservation) से लाभ	1,090

Table 4: Spelling Variation of at least one keyword of the query

S.No	Hindi Query	Google Results
1	आरक्षण से फायदा	365,000
2	आरक्षन (Spelling Variation) से फायदा	25
3	रिजर्वेशन (Reservation)से फायदा	15,100
4	रिज़र्वेशन (Reservation) से फायदा	1,150

Table 5: Synonym Variation of at least one keyword of the query

S.No	Hindi Query	Google Results
1	आरक्षण से फ़ायदा	13,800
2	आरक्षन (Spelling Variation) से फ़ायदा	1
3	रिजर्वेशन (Reservation)से फ़ायदा	371
4	रिज़र्वेशन (Reservation) से फ़ायदा	426

Table 6: Phonetic difference in Synonym Variation of at least one keyword of the query

S.NO	New Data Base Entries of Popular Keywords and their phonetic equivalents
1	:अजन्मा:अजात:अनुत्पन्न:अनुद्भूत:अप्रादुर्भूत:अज:अजन:अजन्म:अनन्यभव:अनागत:अयोनि:**उन्बोर्ण:*अन-बोर्न*:**
2	:घोटाला:गोलमाल:घपला:धाँधली:हेराफेरी:हेरफेर:**स्कैम**: ***स्केम***:**करप्शन**: ***करपशन:कोरप्शन:कोर्रुप्तिओन***:भ्रष्टाचार: ***भ्रश्टाचार***:
3	:शेरनी:मादा_बाघ:मादा_व्याघ्र:बाघिन:व्याघ्री:**टाइग्रेस:** ***टाइग्रस***:
4	:प्रवीण:निपुण:पारंगत:दक्ष:माहिर:अभ्यस्त:क़ाबिल:कुशल:होशियार:पक्का:सिद्धहस्त:शातिर:पटु:पका:निष्णात:परिपक्व:कार्यकुशल:विचक्षण:अभिज्ञ:अभ्यासी:आप्त:धौंताल:अवसित:संसिद्ध:आकर:आगर:प्रवण:आढ़:**केपेबल**: ***कैपेबल***:
5	:योजना:आयोजना:अभिकल्पन:**स्कीम:प्लान:पलैन:** ***प्लैन***:पॉलिसी: ***पालिसी*. *पोलिसी:पोलिस्य***
6	:स्वास्थ्य:तंदरुस्ती:सेहत:तबीयत:तबियत:आरोगिता:अरोग्यता:अरोगिता:स्वास्थ: ***हेअलथ*:हैलथ: हैल्थ *:हेल्थ:***

7	:बीमा:**इंश्योरेंस**:इन्श्योरेन्स:***इन्सुरांस***:
9	:विश्वविद्यालय:यूनिवर्सिटी:**युनिवर्सिटी**:***उनिवेर्सित्य***:
10	:मित्र:दोस्त:साथी:संगी:सखा:यार:सहचर:संगाती:संगतिया:बंधु:मीत:बांधव:बान्धव:बाँधव:मितवा: दोस्तदार:हितैशी:अभिसर:अविरोधी:असामी:इयारा:इष्ट:ईठ:**फ्रेंड**: ***फरेंड***:***फ्रैंड***:

By observing the above example tables a Hindi query can find its variations in different forms and in each case different set of search results can be obtained. The major problem for Hindi data retrieval is due to spelling variations. Not only basic Hindi keywords have spelling variations but synonyms and English equivalent Hindi keywords also suffer from spelling variations. In the light of above example the keyword आरक्षण (AARAKSHAN) is a basic keyword having आरक्षन (AARAKSHAN) as a spelling / Phonetic variation, the English equivalent Hindi keyword for (AARAKSHAN) (आरक्षण/आरक्षन) is (रिजर्वेशन) (RESERVATION) and it further has a spelling variation (रिज़र्वेशन) (RESERVATION). Similarly the basic keyword लाभ has a synonym (फायदा) (FAAYDA) which finds its frequent use in the context of the query and has one more phonetic variation (फ़ायदा) (FAAYDA).

In the above table we present the google results against different forms of the query and observe that for each semantically equivalent query we get a different quantity of results which is due to spelling / phonetics, synonyms and English Hindi equivalent keywords. The phonetic difference between the keywords (रिज़र्वेशन) / (रिजर्वेशन) and (फायदा) / (फ़ायदा) is because of pronunciation. In both the keywords the dot under the letter ज (J) and फ (PH) brings a huge change in web results. Keyword (रिजर्वेशन) containing letter ज is pronounced as RE***J***ESRVATION and letter ज़ as RE***Z***ERVATION which is more appropriate similarly

Keyword (फायदा) containing letter फ is pronounced as ***PH***AAYDA and letter फ़ as ***F***AAYDA which is more appropriate.

Without changing the semantic nature of query search results can be mined out by including the spelling variations, synonyms and English equivalent Hindi Keywords in the query. We also used keyword/query ranking system to suggest the users so that one can pick the highly ranked query to pursue the search. Detailed description of the keyword/query ranking system is explained later in this paper. Search engines do not include these factors in searching. To facilitate users we attempt to develop an interface which acts as a query optimizer, supported with large scale Hindi database different from Query expansion. Query expansion uses different techniques and methods which have least role in query optimization.

6. Database

The Hindi Query Optimizer interface has been developed by using a Database Approach. The study of the structure of Hindi language and its importance in Hindi IR suggests a need of Database which could help in handling of Morphology, Spelling Variations, Word Synonyms and Foreign Language words that directly influence Hindi Language on a wider Scale. The portion data for development of database has been obtained from Hindi Wordnet and subsequent modifications and additions have been made to the database as per the interface and language *platform* requirements.

6.1 Hindi Wordnet a brief introduction

The Hindi WordNet is a system for bringing together different lexical and semantic relations between the Hindi words. It organizes the lexical information in terms of word meanings and can be termed as a lexicon based on psycholinguistic principles. The design of the Hindi WordNet is inspired by the famous English WordNet.

`In the Hindi WordNet the words are grouped together according to the similarity of their meaning. Two words that can be interchanged in a context are synonymous in that context. For each word there is a synonym set, or synset, in the Hindi WordNet, representing one lexical concept. This is done to remove ambiguity in cases where a single word has multiple meanings. Synsets are the basic building blocks of WordNet. The Hindi WordNet deals with the content words, or open class category of words. Thus, the Hindi WordNet contains the following category of words- Noun, Verb, Adjective and Adverb.

Each entry in the Hindi WordNet consists of the following elements

Synset: It is a set of synonymous words. For example, "विद्यालय, पाठशाला, स्कूल" (vidyaalay, paaThshaalaa, skuul) represents the concept of school as an educational institution. The words in the synset are arranged according to the frequency of usage.

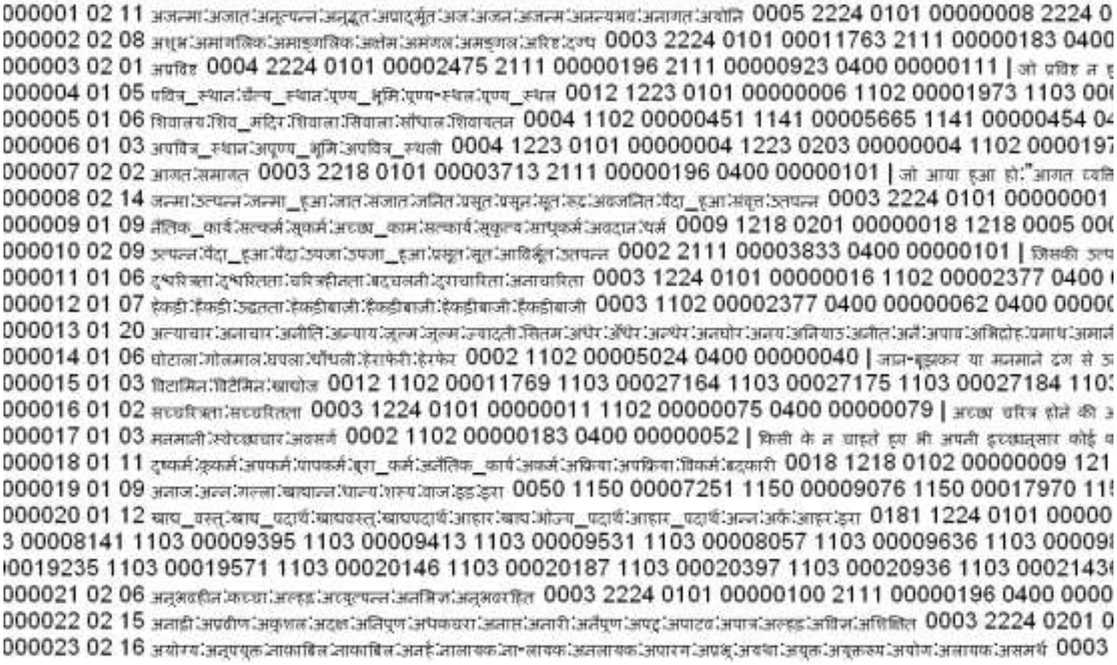
Gloss: It describes the concept. It consists of two parts:

Text definition: It explains the concept denoted by the synset. For example,"वह स्थान जहाँ प्राथमिक या माध्यमिक स्तर की औपचारिक शिक्षा दी जाती है" (vah sthaan jahaan praathamik yaa maadhyamik star kii aupachaarik sikshaa dii jaatii hai) explains the concept of school as an educational institution.

Example sentence: It gives the usage of the words in the sentence. Generally, the words in a synset are replaceable in the sentence. For example, "इस विद्यालय में पहली से पाँचवी तक की शिक्षा दी जाती है" (is vidyaalay men pahalii se paanchaviin tak kii shikshaa dii jaatii hai) gives the usage for the words in the synset representing school as an educational institution [17]

The Hindi Wordnet API is available online at http://www.cfilt.iitb.ac.in/wordnet/webhwn/ and has no direct application to information retrieval. We also focus on exploring the usage of Hindi Wordnet for its application to Hindi IR. Below we present a snapshot of Hindi database offered by WordNet.

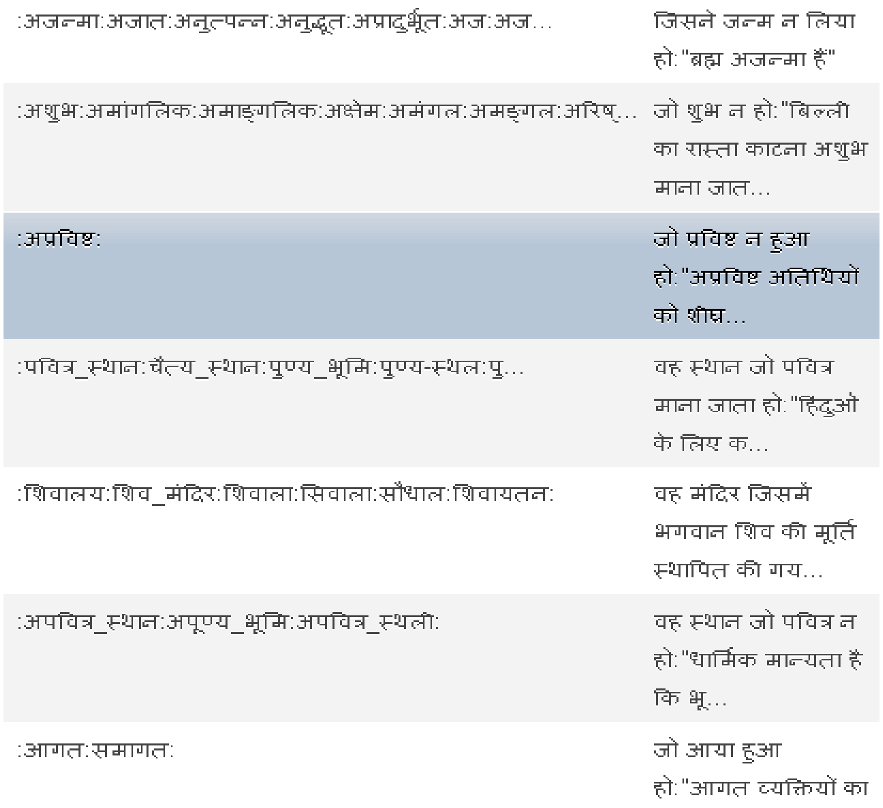
Figure 1: Snapshot of Hindi database obtained from Wordnet



```
000001 02 11 अजन्मा:अजात:अनुत्पन्न:अनुद्भूत:अप्रादुर्भूत:अज:अजन:अजन्म:अनन्यभव:अनागत:अयोनि 0005 2224 0101 00000008 2224 0
000002 02 08 अशुभ:अमांगलिक:अमाङ्गलिक:अक्षेम:अमंगल:अमङ्गल:अरिष्ट:दुष्ट 0003 2224 0101 00011763 2111 00000183 0400
000003 02 01 अपवित्र 0004 2224 0101 00002475 2111 00000196 2111 00000923 0400 00000111 | जो पवित्र न
000004 01 05 पवित्र_स्थान:चैत्य_स्थान:पुण्य_भूमि:पुण्य-स्थल:पुण्य_स्थल 0012 1223 0101 00000006 1102 00001973 1103 00
000005 01 06 शिवालय:शिव_मंदिर:शिवाला:सिवाला:सौपाल:शिवायतन 0004 1102 00000451 1141 00005665 1141 00000454 0
000006 01 03 अपवित्र_स्थान:अपुण्य_भूमि:अपवित्र_स्थली 0004 1223 0101 00000004 1223 0203 00000004 1102 000019
000007 02 02 आगत:समागत 0003 2218 0101 00003713 2111 00000196 0400 00000101 | जो आया हुआ हो:"आगत व्यक्ति
000008 02 14 जन्मा:उत्पन्न:जन्मा_हुआ:जात:संजात:जनित:प्रसूत:प्रसून:सूत:रूढ़:अवजनित:पैदा_हुआ:संभूत:उत्पन्न 0003 2224 0101 00000001
000009 01 09 नैतिक_कार्य:सत्कर्म:सुकर्म:अच्छा_काम:सत्कार्य:सुकृत्य:साधुकर्म:अवदान:धर्म 0009 1218 0201 00000018 1218 0005 00
000010 02 09 उत्पन्न:पैदा_हुआ:पैदा:उपजा:उपजा_हुआ:प्रसूत:सूत:आविर्भूत:उत्पन्न 0002 2111 00003833 0400 00000101 | जिसकी उत्प
000011 01 06 दुश्चरित्रता:दुश्चरितता:चरित्रहीनता:बदचलनी:दुराचारिता:अनाचारिता 0003 1224 0101 00000016 1102 00002377 0400
000012 01 07 हेकड़ी:हैकड़ी:उद्धतता:हेकड़ीबाजी:हैकड़ीबाजी:हेकड़ीबाज़ी:हैकड़ीबाज़ी 0003 1102 00002377 0400 00000062 0400 0000
000013 01 20 अत्याचार:अनाचार:अनीति:अन्याय:ज़ुल्म:जुल्म:ज्यादती:सितम:अंधेर:अँधेर:अन्धेर:अनघोर:अनय:अनियाउ:अनीत:अनै:अपाव:अभिद्रोह:प्रमाथ:अमानु
000014 01 06 घोटाला:गोलमाल:घपला:धाँधली:हेराफेरी:हेरफेर 0002 1102 00005024 0400 00000040 | जान-बूझकर या मनमाने ढंग से
000015 01 03 विटामिन:विटैमिन:खाद्योज 0012 1102 00011769 1103 00027164 1103 00027175 1103 00027184 110
000016 01 02 सच्चरित्रता:सच्चरितता 0003 1224 0101 00000011 1102 00000075 0400 00000079 | अच्छा चरित्र होने की
000017 01 03 मनमानी:स्वेच्छाचार:अवसर्ग 0002 1102 00000183 0400 00000052 | किसी के न चाहते हुए भी अपनी इच्छानुसार कोई
000018 01 11 दुष्कर्म:कुकर्म:अपकर्म:पापकर्म:बुरा_कर्म:अनैतिक_कार्य:अकर्म:अक्रिया:अपक्रिया:विकर्म:बदकारी 0018 1218 0102 00000009 121
000019 01 09 अनाज:अन्न:गल्ला:खाद्यान्न:धान्य:शस्य:नाज:इड:इरा 0050 1150 00007251 1150 00009076 1150 00017970 11
000020 01 12 खाद्य_वस्तु:खाद्य_पदार्थ:खाद्यवस्तु:खाद्यपदार्थ:आहार:खाद्य:भोज्य_पदार्थ:आहार_पदार्थ:अन्न:अर्क:आहर:इरा 0181 1224 0101 00000
3 00008141 1103 00009395 1103 00009413 1103 00009531 1103 00008057 1103 00009636 1103 00009
00019235 1103 00019571 1103 00020146 1103 00020187 1103 00020397 1103 00020936 1103 0002143
000021 02 06 अनुभवहीन:कच्चा:अल्हड़:अव्युत्पन्न:अनभिज्ञ:अनुभवरहित 0003 2224 0101 00000100 2111 00000196 0400 0000
000022 02 15 अनाड़ी:अप्रवीण:अकुशल:अदक्ष:अनिपुण:अधकचरा:अजान:अनारी:अनैपुण:अपटु:अपाटव:अपात्र:अल्हड़:अविज्ञ:अशिक्षित 0003 2224 0201 0
000023 02 16 अयोग्य:अनुपयुक्त:नाकाबिल:नाक़ाबिल:अनर्ह:नालायक:ना-लायक:अनलायक:अपारग:अप्रवृ:अयथा:अयुक्त:अयुक्तरूप:अयोग:अलायक:असमर्थ 0003
```

The database obtained from Hindi Wordnet has been parsed and modified as per the application requirements. New additions have also been made. A snapshot of the parsed/modified database that has been used for our research purpose is presented below in figure 2.

Figure 2: Snapshot of Hindi database Modified as per application requirements



:अजन्मा:अजात:अनुत्पन्न:अनुद्भूत:अप्रादुर्भूत:अज:अज...	जिसने जन्म न लिया हो:"ब्रह्म अजन्मा हैं"
:अशुभ:अमांगलिक:अमाङ्गलिक:अक्षेम:अमंगल:अमङ्गल:अरिष्...	जो शुभ न हो:"बिल्ली का रास्ता काटना अशुभ माना जात...
:अप्रविष्ट:	जो प्रविष्ट न हुआ हो:"अप्रविष्ट अतिथियों को शीघ्र...
:पवित्र_स्थान:चैत्य_स्थान:पुण्य_भूमि:पुण्य-स्थल:पु...	वह स्थान जो पवित्र माना जाता हो:"हिंदुओं के लिए क...
:शिवालय:शिव_मंदिर:शिवाला:सिवाला:सौंधाल:शिवायतन:	वह मंदिर जिसमें भगवान शिव की मूर्ति स्थापित की गय...
:अपवित्र_स्थान:अपूण्य_भूमि:अपवित्र_स्थली:	वह स्थान जो पवित्र न हो:"धार्मिक मान्यता हैं कि भू...
:आगत:समागत:	जो आया हुआ हो:"आगत व्यक्तियों का

The Modified Database is a two column database first column holds the keywords with all their morphological, phonetic, synonyms and English equivalent Hindi variants and second column holds the text definition.

6.2 Additions in the database

The Wordnet database includes the variants of the keywords up to a certain level. On closely observing almost all entries in the database we found that there is a need for inclusion of more English equivalent Hindi keywords along with their phonetic variants (which are least in the Wordnet version of the database under use) and more phonetic variations of existing Hindi keywords. Since the synonyms and English equivalent keywords have more phonetic variations, we have added those keyword variations up to a certain level. Below we present the examples of appropriate additions (highlighted) to the Hindi Wordnet database table 7. Keywords highlighted in bold are English Equivalent Hindi keywords, and those highlighted in bold and italicized are phonetic variations. All highlighted entries are new additions to the existing Wordnet and the process of adding new entries and new keywords will be a continuous process.

Table7: Appropriate additions (highlighted)

English Word	Transliteration	Actual Hindi	Google results: Number of documents returned for Transliterated words
Policy	पोलिस्य	पॉलिसी	625
Corruption	कोर्रुप्तिओन	करप्शन	2,280
Insurance	इन्सुरांस	इंश्योरेंस	460
Health	हेअलथ	हैल्थ	871
University	उनिवेर्सित्य	यूनिवर्सिटी	870

Hindi web developers make use of the transliteration software. The most popular and freely available transliterator is Google Transliteration software. As it is quite clear that now-a-days Hindi data available on the web is influenced with English Language and many commonly used English keywords written in Hindi are used in the web pages. Some of the common words are Policy, Corruption, Insurance, Health, University etc. The Transliteration of these words yields non-standard and wrong Hindi output and the web pages are affected due to this and hence the need for retrieval. To make our point clear we present an example table 8 below.

Table 8: Results obtained for wrongly transliterated Hindi words

Medical Terms	Hindi	Spelling Variant/s
Antiretroviral	एंटीरेट्रोवाइरल	एन्टीरेट्रोवाइरल

campylobacter	कैम्पाइलोबैक्टर	काम्प्य्लोबक्टेर
Dyspepsia	डिसपेप्सिया	डिस्पेप्सिया
Filariasis	फाइलेरिएसिस	फिलारिअसिस
Hepatitis	हिपेटाइटिस	हैपेटाइटिस / हिपैटाइटिस
Impetigo	इम्पेटिगो	इम्पेटीगो
Tuberculosis	ट्यूबर्क्युलोसिस	टुबेर्कुलोसिस

The data retrieved against these keywords is relevant in the context of the keywords. But the question is how a native Hindi user will be able to fetch this kind of available data from the web. To ensure the retrieval of such data we include keywords like these in our database as additional entries as shown in table above.

6.3 Domain Specific Database approach

As the Hindi literature is growing on web it becomes very important to focus on the various domains. Hindi information on domains like agriculture, medicine including Ayurveda, tourism, etc., is now available on the web. Therefore it becomes veryimportant to make this information available to the users. Hindi Wordnet does not include domain specific keywords on larger scale. We take into consideration the inclusion of domain specific keywords into our database. The development of all domain specific databases is a time consuming process. However the process of inclusion of medical domain based keywords in the database has been started which includes names of the diseases and names of the medicines along with the phonetic variations of the keywords. As Hindi language is concerned the Medical terms in English Language are often misspelled in Hindi. We attempt to include in our database the correct senses of such words in Hindi by including a separate domain with dictionary look up style. Some of the (medical terms) keywords are listed below in the table.

Table 9: Addition of medical domain specific keywords in Hindi

Users	**Keyword Frequency**
User 1	बीमा
User 2	**इंश्योरेंस**
User 3	इन्श्योरेन्स
User 4	**इंश्योरेंस**
User 5	इन्श्योरेन्स
User 6	**इंश्योरेंस**
User 7	**इंश्योरेंस**

7. The Interface Design

The interface has been designed to carry out the Hindi Search activity where Hindi query can be issued by the user either by typing from keyboard or using transliteration API. We have also provided a standard soft Hindi keyboard for the query input. The interface uses database as backend for matching and retrieval of Hindi keywords. Search engines like Google, Yahoo, Bing and Guruji can be used as selections for information retrieval which makes the interface a Meta search platform. The queries supplied by the user are saved in query-log which is a separate database used for processing the keywords for their further optimization. To accomplish this purpose we used the keyword ranking approach. The process of keyword ranking is simple to use and implement. The generalized working model of the system is shown below as a graphical layout.

Figure 3: The General Model

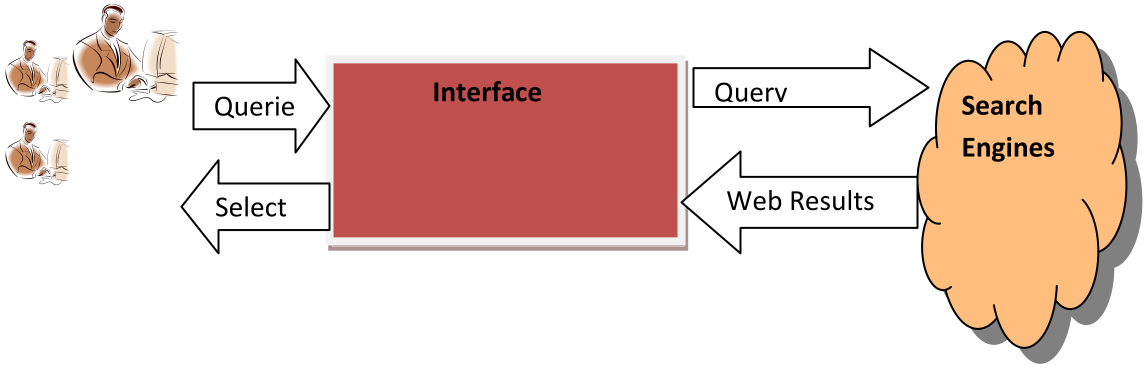
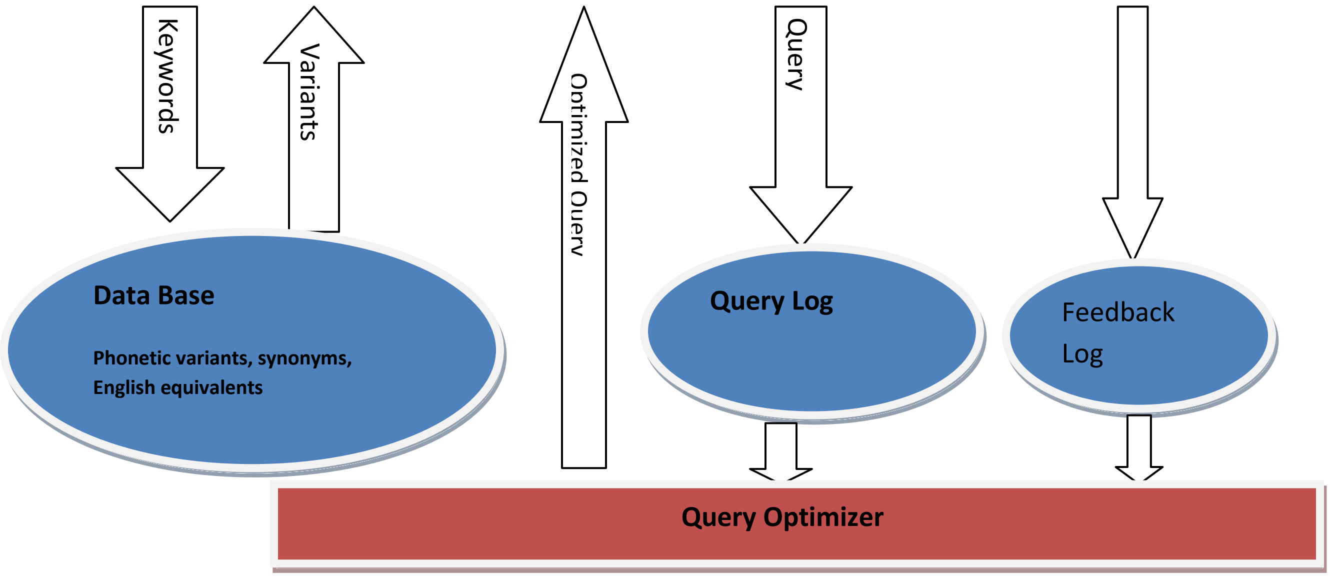


Figure 4: The working model of the query interface/optimization system



7.1 Keyword Ranking

The Hindi keywords are present in the database with their variants as a group. The maximum usage of a particular keyword in a group gives it a high score. Example: The keyword insurance has following variants: बीमा:इंश्योरेंस:इन्श्योरेन्स:इन्सुरांस. The maximum usage of a particular keyword will be suggested for its use.

Table 10: Frequency of word

Users	Keyword Frequency
User 1	बीमा
User 2	**इंश्योरेंस**
User 3	इन्श्योरेन्स
User 4	**इंश्योरेंस**
User 5	इन्श्योरेन्स
User 6	**इंश्योरेंस**
User 7	**इंश्योरेंस**

Since the keyword इंश्योरेंस has been used more frequently in its group; it is recommended for use in search.

7.2 Query Optimization

The query optimization can be done on the basis of keyword ranking. When multiple keywords are supplied as a query each winning keyword from its group is arranged in an order and is suggested as an optimized query. The example below throws light on this procedure.

Let a Hindi query be भारत में विदेशी निवेश (Foreign investment in India). The keywords have following variants associated with them. Variants are phonetic, synonyms and English equivalent keywords in Hindi.

Table 11: Multiple keywords in a group

Keywords	Variants in Database
भारत	हिंदुस्तान:हिन्दुस्तान:भारत_वर्ष:हिंदोस्ताँ:हिंद:हिंदोस्तान:भारतवर्ष:हिन्दोस्तान: हिन्द:नरभू:इण्डिया:इंडिया:
में	:के_अंदर:के_अन्दर:के_भीतर:
विदेशी	:परदेशी:परदेसी:बिदेसी:ग़ैरमुल्की:गैरमुल्की:विलायती:अजनबी:अजन:अन्यदेशी य:अन्यराष्ट्रीय:देसावरी:फारन:फारेन:फारेनर:
निवेश	:पूँजी_निवेश:पूँजी-निवेश:निवेश:इन्वेस्टमेंट:इन्वेस्टमेन्ट:इन्वेस्टमैन्ट:इन्वेस्टमैंट:

Depending upon the number of hits on the particular selected keyword,` the ranking procedure will generate the combination of the most frequently used keywords into different queries in ascending order. The following queries can be generated by using the keyword ranking method.

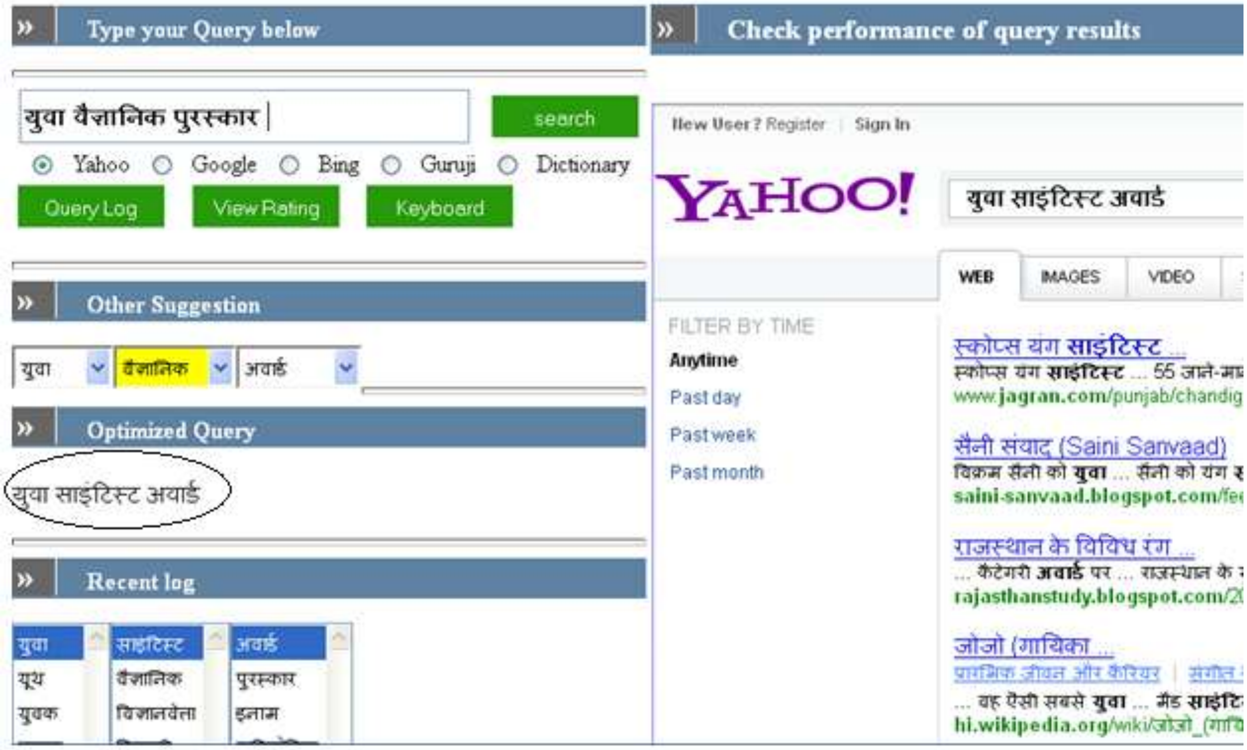
Table 12: Number of queries that can be generated

S.NO	Order of suggested queries
1	इंडिया में विदेशी पूँजी-निवेश
2	भारत में फारेन इन्वेस्टमेंट
3	हिन्दुस्तान के_अंदर ग़ैरमुल्की निवेश
4	And so on

By using the aforesaid approach the problem of recall and precision in Hindi language has improved up to a significant level**.** The original query can have its variants without changing the sense of the query thus reducing the efforts of users to pursue search**.** Not only the recall and precision have been impacted but the scope of search for user has become easy and simple. The database oriented approach saves time and efforts of user for making searches as all possible variations are provided to user for his reference. User is benefited in two possible ways first he/she can use his own selections for query generation second he/she can use the suggested queries which are generated by the interface.

To facilitate the Hindi input the transliteration service along with soft Hindi keyboard is also provided to the novice Hindi users to submit their queries. The keywords supplied by the users are sent to the database to fetch their phonetic variants, synonyms and English equivalent keywords. Users can select the keywords from the list and proceed with further search. Below we present a brief demo example of the working of the Interface. A Hindi query is supplied to the interface, Hindi Query: युवा वैज्ञानिक पुरस्कार which means *Young scientist award*. The result provided by interface is the select list of the keywords obtained from the database. The select list contains the possible variations of the keywords in the query which can be selected for further search. The example is explained below in the figure. Snapshot of the interface

Figure 5: Snapshot of the Interface

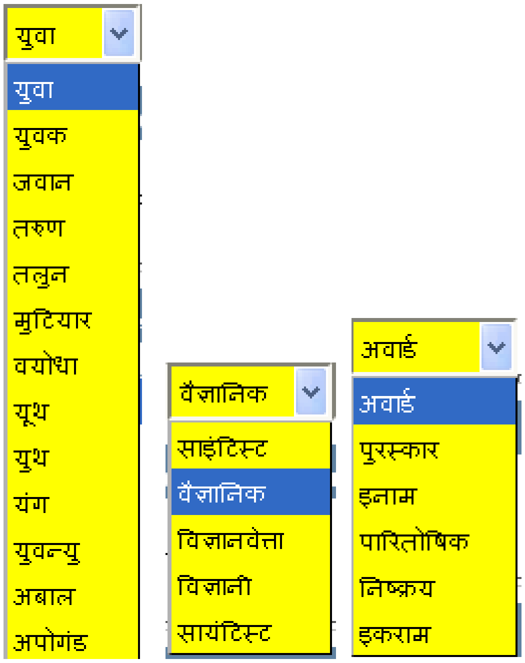


स्कोप्स यंग साइंटिस्ट ...
स्कोप्स यंग साइंटिस्ट ... 55 जाने-मा
www.jagran.com/punjab/chandig

सैनी संवाद (Saini Sanvaad)
विक्रम सैनी को युवा ... सैनी को यंग
saini-sanvaad.blogspot.com/fe

राजस्थान के विविध रंग ...
... कैटेगरी अवार्ड पर ... राजस्थान के
rajasthanstudy.blogspot.com/2

जोजो (गायिका ...
प्रारंभिक जीवन और कैरियर | संगीत
... वह ऐसी सबसे युवा ... मैड साइंटि
hi.wikipedia.org/wiki/जोजो_(गायि



7.3 The query log and optimization

As we discussed above, using keyword ranking method, the query can be optimized to the highest level. To accomplish this, the interface maintains a query log so that the queries submitted by the users can be recorded and optimized. The number of hits on a particular keyword decides its score. The major factor that influences the Hindi language is its phonetic nature. A keyword can have various spelling variations. Hindi users make use of different spellings for the same keyword. Our approach to keyword ranking depends on the probability of the selection of the phonetically right keyword, right synonym and right English equivalent keyword. An example below shows how the query optimization can be implemented and be helpful to the Hindi web users to search the web.

For query युवा वैज्ञानिक पुरस्कार most frequent searches have been made for साइंटिस्ट (scientist) as compared to वैज्ञानिक and अवार्ड (award) as compared to पुरस्कार which is recorded in the log. Therefore the optimized query against युवा वैज्ञानिक पुरस्कार becomes युवा साइंटिस्ट अवार्ड and meaning of the query remains the same. The optimized query can be seen encircled in the figure above.

7.4 Relevance Feedback

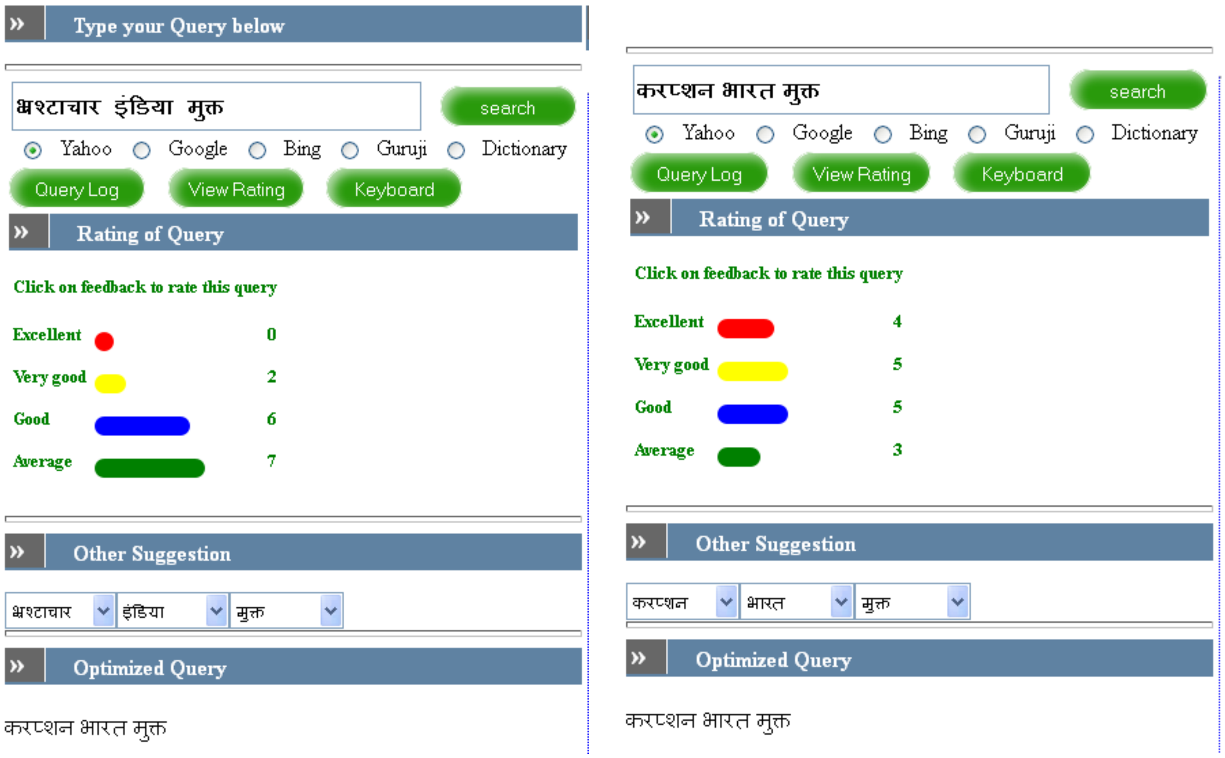
Measuring the information retrieval effectiveness of Web search engines can be expensive if human relevance judgments are required to evaluate search results. Using implicit and explicit user feedback for search engine evaluation provides a cost and time effective manner of addressing this problem. Web search engines can use human evaluation of search results without the expense of human evaluators. An additional advantage of this approach is the availability of real time data regarding system performance. We use the explicit feedback to calculate performance metrics, such as precision. We show that the presentation of relevance feedback to the user is important in the success of relevance feedback.

To observe the relevance of the search results the feedback feature has been provided with the following parameters:

1. Average
2. Good
3. Very Good
4. Excellent

The feedback provided by the users is saved in a separate log to analyze information on how users are searching. Below we present a snapshot of the interface where the working of the feature of the feedback module is shown. Also it can be seen that a query भ्रश्टाचार इंडिया मुक्त is optimized as करप्शन भारत मुक्त. In figure 6 it can be clearly seen that feedback results for the query करप्शन भारत मुक्त are better than the former; therefore, the feedback feature plays an important role for optimizing the query.

Figure 6: Feedback for query करप्शन भारत मुक्त



भ्रश्टाचार इंडिया मुक्त

● Yahoo ○ Google ○ Bing ○ Guruji ○ Dictionary

Click on feedback to rate this query

Excellent	🟥	0
Very good	🟨	2
Good	🟦	6
Average	🟩	7

भ्रश्टाचार | इंडिया | मुक्त

करप्शन भारत मुक्त

करप्शन भारत मुक्त

● Yahoo ○ Google ○ Bing ○ Guruji ○ Dictionary

Click on feedback to rate this query

Excellent	🟥	4
Very good	🟨	5
Good	🟦	5
Average	🟩	3

करप्शन | भारत | मुक्त

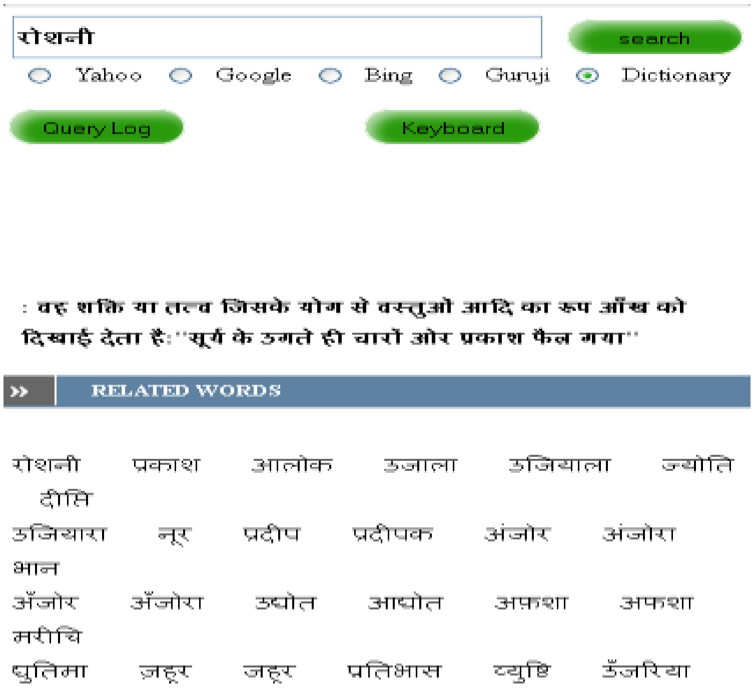
करप्शन भारत मुक्त

7.5 Interface as a source of word look up dictionary

The database has been organized as a two column data source first column being the keyword/s source and second being the text description. It can be used as an online dictionary at the interface level as shown below in the snapshot.

As discussed above we also have included a dictionary feature so that meanings of complex Hindi words can be understood. In the Database section of the paper we discussed, the two column arrangement of the database where the first column holds the keyword/s variant/s etc and second holds the explained meaning of that keyword. A sample snapshot of the interface with dictionary feature is shown below.

Figure 7: Dictionary feature with the meaning of the keyword along with the related words.



रोशनी

: वह शक्ति या तत्व जिसके योग से वस्तुओं आदि का रूप आँख को दिखाई देता है:''सूर्य के उगते ही चारों ओर प्रकाश फैल गया''

RELATED WORDS

रोशनी प्रकाश आलोक उजाला उजियाला ज्योति
दीप्ति
उजियारा नूर प्रदीप प्रदीपक अंजोर अंजोरा
भान
अँजोर अँजोरा उद्योत आद्योत अफ़शा अफशा
मरीचि
द्युतिमा ज़हूर जहूर प्रतिभास व्युष्टि उँजरिया

We conducted experiments related to monolingual IR and web IR in the context of Hindi language. 1245 Queries received by users were organized into various domains namely “Agriculture”, “Science and Technology”, “Medical”, “General” and “Tourism”. Some additional experiments on the 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 by using our interface. In our experiments we concluded that the query optimization helped to solve the low recall problem up to a great extent. However, keeping in consideration the length of the present paper the experiments will be reported shortly.

8. Conclusion: Query Optimization as solution

All the issues discussed above lead to low recall in Hindi search process and are needed to be addressed. In our work we addressed all these problems and found that the problem of recall can be solved by including these parameters in Hindi search.

The Interface supported with large scale database designed by us handles all these issues and thus solves the problem of 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 queries supplied by the user are saved in query log which is a separate database used for processing the keywords for their further optimization. To accomplish this purpose we used the keyword ranking approach. The Hindi keywords are present in the database with their variants as a group. The maximum usage of a particular keyword in a group gives it a high score.

When multiple keywords are supplied as a query each winning keyword from its group is arranged in an order and is suggested as an optimized query. The optimized query is further suggested to the user to use as it contains optimized keywords which have been searched most of the times. Interface helps users to mine the Hindi information from web and hence chances of retrieving relevant information are increased. The query optimization has solved the problem of low recall for Hindi IR up to a great extent**.**

==

References

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

[2] DebasisMandal, Mayank Gupta, SandipanDandapat, Pratyush Banerjee, and SudeshnaSarkar "*Bengali and Hindi to English CLIR Evaluation*" Department of Computer Science and Engineering IIT Kharagpur, India – 721302Springer Berlin Heidelberg Series Volume 5152 Series ISSN 0302-9743 Pages pp 95-102.

[3] Manoj Kumar Chinnakotla, SagarRanadive, Pushpak Bhattacharyya and Om P. Damani"*Hindi and Marathi to English Cross Language Information Retrieval*" at CLEF 2007 Department of CSE IIT Bombay Mumbai, India Advances in Multilingual and Multimodal Information Retrieval Pages 111 - 118 Springer-Verlag Berlin, Heidelberg ©2008 ISBN: 978-3-540-85759-4

[4] Prasad Pingali and VasudevaVarma"*Hindi and Telugu to English Cross Language Information Retrieval*" at CLEF 2006 Language Technologies Research Centre IIIT, Hyderabad, India.Evaluation of Multilingual and Multi-modal Information Retrieval Lecture Notes in Computer Science, 2007, Volume 4730/2007, 35-42, DOI: 10.1007/978-3-540-74999-8_4

[5] Tan Xu1 and Douglas W. Oard1 "*FIRE-2008 at Maryland: English-Hindi CLIR*" College of Information Studies and 2CLIP Lab, Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742, USA.

[6] Mallammavreddy, Hanumanthapa."*ENGLISH TO KANNADA/TELUGU NAME TRANSLITERATION IN CLIR: A STATISTICAL APPROACH*" Department of Computer Science and Applications, Bangalore University, Bangalore-560 056, INDIA IJMI International Journal

of Machine Intelligence ISSN: 0975–2927 & E-ISSN: 0975–9166, Volume 3, Issue 4, 2011, pp-340-345

[7] Mallamma V Reddy, Dr. M. Hanumanthappa "*Kannada and Telugu Native Languages to English Cross Language Information Retrieva*l" Department of Computer Science and Applications, Bangalore University, Bangalore, INDIA. (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 2 (5) , 2011, 1876-1880

[8] Dr.S.Saraswathi, AsmaSiddhiqaa.M, Kalaimagal.K, Kalaiyarasi.M"*BiLingual Information Retrieval System for English and Tamil*" JOURNAL OF COMPUTING, VOLUME 2, ISSUE 4, APRIL 2010, ISSN 2151-9617

[9] Pingali V.V. Prasad Rao "*Recall Oriented Approaches for improved Indian Language Information Access*" Language Technologies Research Centre International Institute of Information Technology Hyderabad - 500 032, INDIA August 2009 Source iiit.ac.in

[10] SivajiBandhyopadhyayAmitava Das PinakiBhaskar"*English Bengali Ad-hoc Monolingual Information Retrieval Task Result at FIRE 2008*" Department of Computer Science and EngineeringJadavpur University, Kolkata-700032, India Source www.amitavadas.com/Pub/Fire_2010.pdf

[11] Ashish Almeida, Pushpak Bhattacharyya "*Using Morphology to Improve Marathi Monolingual Information Retrieval*" IIT Bombay.Source http://www.isical.ac.in/~fire/paper/Ashish_almeida-IITB-fire2008.pdf

[12] GANAPATHIRAJU Madhavi, BALAKRISHNAN Mini, BALAKRISHNAN N., REDDY Raj "*Om: One tool for many (Indian) languages*" Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA Journal of Zhejiang University SCIENCE ISSN 1009-3095

[13] Sujoy Das AnuragSeetha M. Kumar J.L. Rana“*Post Translation Query Expansion using Hindi Word-Net for English-Hindi CLIR System*” source www.isical.ac.in/~fire/paper_2010/sujaydas-manit-fire2010.pdf

[14] KeshavNiranjan“*Language Technology in India*” Ph.D. Candidate LANGUAGE IN INDIA Strength for Today and Bright Hope for Tomorrow Volume 12: 4 April 2012 ISSN 1930-2940

[15] Kumar Sourabh and Vibhakar Mansotra“*FactorsAffecting 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

[16] 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

[17] Hindi WordNet documentation Pushpak Bhattacharya http://www.cfilt.iitb.ac.in/wordnet/webhwn/

==

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

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