A Study of Web Mining Tools for Query Optimization

UNIVERSITY OF JAMMU
India's first ISO 9001: 2000 University

THESIS SUBMITTED
To
THE FACULTY OF Mathematics
UNIVERSITY OF JAMMU
IN
PARTIAL FULFILMENT
FOR THE AWARD
OF
DEGREE OF
Doctor of Philosophy
In
Computer Science & Information Technology

Supervised By: Prof. Vibhakar Mansotra. Investigated By: Kumar Sourabh

DEPARTMENT OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY
UNIVERSITY OF JAMMU
JAMMU-180006
A B S T R A C T

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 mean 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 support 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 specification. 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. 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. 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. They are sending and receiving 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 changed the scenario.

Realizing the potential of Indian languages, Microsoft and Google has launched various products in the past two years. With the 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.

India centric localized search engines market is saturating 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 current scenario. Here are some of the search engines who fall in 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 has been launched a couple of months back. Also, we do have localized versions of 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 such content.

Information Retrieval in Hindi language is getting popularity and IR systems face low recall if existing systems are used as-is. Upon analyzing the failure cases, we found that certain characteristics of Indian languages cause the existing algorithms not being able 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 feedback 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 get increased.

The entire study is divided into seven chapters. A brief description of each chapter is given below as:

**Chapter-1: “Web Mining and Information Retrieval”** This chapter provides the introduction of Web Mining and its characteristics with detailed description of Information retrieval and historical background. Chapter also covers basic processes and models of IR.

**Chapter-2: Second Chapter “Multilingual-Monolingual Information Retrieval”** provides a theoretical framework for the study. Chapter covers the multilingual and monolingual IR with respect to Asia and provides an insight into challenges in Asian language processing including national language of India i.e. Hindi.
Chapter-3: The third chapter “Literature Review“. In this chapter various IR and NLP tools and their working principles are explained. Various developments in multilingual and monolingual IR have also been discussed. A detailed literature review for developments in IR in India has been done.

Chapter-4: The fourth chapter “Issues in Information Retrieval for Hindi Language”. This chapter provides detailed discussion on IR in Hindi Language. Chapter addresses various problems encountered in Hindi search. The factors which are responsible for low recall in Hindi IR are well addressed and highlighted with detailed examples. Following are the Highlighted factors that affect the Hindi Search on web:

- Hindi Morphological Factors
- Phonetic nature of Hindi Language and Spelling variations
- Multiple words Synonyms and Ambiguity
- Influence of English on Hindi Information retrieval
- Wrong transliteration and named entity recognition

Chapter-5: The fifth chapter “Research Methodology and Software Design” deals with the research methodologies and an approach toward the solution of the problem. This chapter covers the basic details of the design and development of the Interface and database with the working principles of the same. This chapter mainly focuses on the development of the tool for rephrasing and optimizing the Hindi query which ultimately leads as a solution for low recall problem in Hindi language.

Chapter-6: The sixth chapter “Experimental Analysis”. In this chapter we particularly want to draw the attention towards our objectives through experiments and analysis.

- That how web query optimization helps to solve the problem of recall in Hindi information retrieval
- That how the Hindi data can be efficiently mined out by using the search parameters
- That how various search engines perform for data retrieval in Hindi Language
- That how the users can be benefited by using the interface for Hindi search

To perform the experiments the software was distributed to various people (with particular domain expertise e.g., research scholars of Hindi and other departments, Medical, MTM 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 varied 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. In this chapter we presented the detailed experimental analysis of the queries and results (quantity and quality both) with present feedbacks.
The chapter also covers additional experiments and analysis on name entity recognition, wrong transliteration of Hindi keywords. The potential drawback of the interface has also been discussed in this chapter.

**Chapter-7:** The seventh chapter “Conclusion and Future scope” summarizes the findings, recommendations, conclusions and scope for further research, of the study. Interface helps users to mine the Hindi information from web and hence chances of retrieving relevant information get increased. The interface and the query optimization have solved the problem of recall up to a great extent. The interface addresses all monolingual issues in Hindi Language 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. The tool can be used as component feature for many text oriented systems e.g. Hindi Spell correction, Hindi Transliteration, Hindi Translation, Mailing Component, and thesaurus in word processing applications.

The drawback of the interface has been addressed but not resolved and hence becomes one important issues to be resolved in the future research work. Keywords from domains like medical, tourism etc have been added the addition is yet not appropriate more additions are needed to cover subareas in the domains. Also various domains are left untouched for which database have to be generated and organized in future.

Since development in the area of Hindi monolingual IR is less as compared to multilingual information retrieval we shall attempt to develop database for users who want web information in other languages than Hindi in the northern region of India especially in the state of J&K and Himachal Pradesh. To make the information available to such users we shall develop database for languages like Dogri, Urdu and Kashmiri.
We have provided a feedback method so that user behavior of search can be analyzed. With the help of relevance feedback method rank for queries can be generated so that perfect query optimization can be done. In our work we have implemented only explicit feedback. There are implicit feedback methods which are yet to be implemented. We take into consideration the implementation of implicit feedback methods and techniques in future.