Chapter 7
Conclusion and Future Scope

7.1 Discussion

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 Hindi language queries. The various factors that affect Hindi search on web are:

- Morphological Factors
- Phonetic nature of Hindi Language
- Words Synonyms
- Ambiguous Words
- Influence of English on Hindi Information retrieval

Today web search engines provide the easiest way to reach information on the web. In this scenario, more than 90% of Indian language content on the web is not searchable due to multiple encodings of web pages. Most of these encodings are proprietary and hence need some kind of standardization for making the content accessible via a search engine. Therefore we developed an interface to search engine called Hindi Query Optimizer which helps improve low recall in Hindi IR. The interface helps to solve the recall problem for Hindi Language. 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 wider Scale.

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. 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.

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 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 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, medical including Ayurveda, tourism etc is now available on web. Therefore it becomes very important 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 and other domain (Tourism, Science and Technology etc) based keywords in the database has been started which includes phonetic variations of the keywords. As Hindi language is concerned the Medical terms in English Language are oftenly miss spelled 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. We also tested the affect of phonetics on proper nouns, names of popular individuals e.g. Authors, Politicians, Scientists and names of the places and found that phonetics do affect those. Therefore we attempt to enrich our database with proper nouns and their phonetic variations proper nouns include names of the famous persons and places. Wrongly transliterated keywords when supplied to search engines fetch handful of results. Therefore we also take into consideration the wrongly transliterated
Chapter 7: Conclusion and Future Scope

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 designed by us addresses all these issues and provide 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 by us 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.

7.2 Conclusion Basic Issues

It has been concluded that all the factors that affect the performance of Hindi search on web are not addressed by the any of the search engines selected for experiments. It has been shown in the experiments that effective mining can be carried- out by including these parameters for searching in Hindi language.

7.2.1 Phonetic variations

It has been concluded that each phonetically equivalent Hindi keyword is considered as a different keyword and search engines provide different set of results for same keyword. None of the three search engines used for experiments address this problem. Making variations based on this factor recall improves.

7.2.2 Synonyms

It has been concluded that by including the synonyms of the Hindi keywords improve the scope of search by providing relevant results. Including synonyms by replacing the keyword with closely related synonym recall improves.
7.2.3 Morphology

It has been concluded that by using a keyword in its root form improves the performance of search engines. Only Google handles morphology up a certain level whereas Yahoo and Guruji do not.

7.2.4 Influence of English on Hindi

It has been concluded that inclusion the English equivalents of Hindi keywords in search helps to fetch the handful of relevant results. Including such keywords by replacing the keyword with English keyword in Hindi recall improves.

7.2.5 Wrongly transliterated keywords

It has been concluded that Hindi keywords with wrong transliteration are able to fetch large quantity of web results wrong transliteration can be one more amongst the factors contributing the growth of non standard Hindi data on web.

7.2.6 Proper nouns

It has been concluded that due to phonetic variations of proper nouns, the Hindi search is affected. As names of the persons or places are treated as different entities and hence results scatter. Search engines do not address this problem.

7.3 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 address 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 get increased.

The interface and the query optimization have solved the problem of recall up to a great extent. Users using Guruji search engine are benefited most by using the interface. Also the database can be further used as component for various applications like Hindi spell checking, translation, transliteration and thesaurus.

7.4 Future Scope of the Work

As we discussed that by query optimization problem of recall has been solved up to a great extent. The query optimization works well for all parameters except one but only in certain cases. The parameter is inclusion of word synonyms. In certain cases query optimization suffers due to ambiguity involved in the Hindi language because of the multiple synonyms (not closely related) 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. 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 thesis this drawback of the interface has been addressed but not resolved and hence becomes one important issue to be resolved in the future research work.

As we discussed above that 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.

As discussed a feature of relevance feedback method in chapter 5 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.