A Search Engine for Indian Languages

Ashwani Mujoo, Manoj Kumar Malviya, Rajat Moona, TV Prabhakar

(mujoo, manojkm, moona, tvp}@cse.iitk.ac.in
Department of Computer Science and Engineering,
Indian Institute of Technology, Kanpur, India

Abstract. There is a great need for a search engine for web documents written in languages other than English. In this paper, we describe the design issues of a Search Engine for Indian Languages. We also describe the implementation of two Search Engines for Indian Languages, one for documents in ISCII and the other for documents in Unicode. The software allows full-text indexing and searching of a database of documents written in any Brahmi-based Indian Language. The Search engine gathers the HTML documents from the web, indexes and compresses the documents and then searches for the given keywords. The main features of the search engines are phonetic tolerance, morphological analysis, compression and indexing, leading and trailing substring matches for keywords, search through compressed documents. The implementation includes a search server architecture, which can be accessed from a WYSIWYG front end, which is a Java swing applet. Performance results show that the search engine achieves a compression of almost 80 percent and has an appreciable precision and recall.

1 Introduction

The Internet in India is growing by more than 300% a year, and poses some unique problems not found elsewhere. There are 15 officially recognized languages and various sites with localized content in multiple languages are coming up rapidly. Though there are search engines that search documents in different languages of the world, none of them are enabled to search documents written in Indian Languages. Another problem is that many Indian Language documents take the route of using the FONT FACE tag in HTML along with a glyph set for Indian Languages. This is sufficient for displaying documents but is a major obstacle for searching. Encoding schemes like ISCII and Unicode not only facilitate display but also give a handle on indexing and searching.

In this paper we describe the implementation of two Search Engines that can search documents written in Devanagari (Deavanagari is a script used by several Indian languages like Hindi, Marathi, Sanskrit etc.). First we describe the Search Engine for documents in ISCII encoding where we handle the major linguistic issues and then for Unicode documents where we take care of the compression of the web documents.

2 Design Issues

Language specific features can be expected to have a significant influence on the architecture of an Internet search engine. We identify some of the issues that need to be considered when building an Indian language search engine.

- **Different Forms of the Words**: Almost in every language grammatical markers (like *s*, *es*, *ing* etc in English and ध, बाला, याँ in Hindi) are used with a word stem (or root word) and new words are made. These new words called morphological variants of the stem, present the same concept but differ in tense, plurality etc. An example, लड़के, लड़कों are morphological variant of the root word लड़का. It is possible that a document contains a root word, say लड़का and the user gives a morphological variant of the root word, say लड़के for searching. The search engine should be able to handle such conditions and should give matches for all the morphological variants of the word.

- **Phonetic Tolerance**: Indian languages alphabet contains many characters (ृ, ॠ, ५), which sound similar, i.e. they are phonetically equivalent. These characters are often used interchangeably. For example, झांडा, झण्डा, झन्डा, झंडा are phonetically equivalent words. Users may use any phonetically equivalent word of a keyword for searching. So the search engine should be able to support some form of phonetic tolerance. If the user gives the word झंडा for searching, the search engine should be able to get matches for all its phoneme equivalent words झण्डा, झन्डा, झंडा.

- **Font Independence**: Due to lack of standardization, Indian language documents on the Web are being written in different fonts. Each web author uses her own font encoding in the HTML documents using the FONT FACE tag. There is no universally accepted font or display encoding. A user can see the Hindi content only when he has the same fonts as were used by the web page author. So the character sequence in a web page is a function of the fonts used for viewing the web page. To search a word in such documents is not possible, since the font encoding is not a standard.

- **Language Independence**: Since there are a number of spoken and written languages in world, a user may want a search engine with no language barriers. A truly internationalized search engine will be able to search for different sites having content in any language of the world.

- **Indian Language Front End**: Since the content that is being searched and gathered by the search engine is in an Indian Language, it is expected that the user would like to give the query in an Indian Language. So the front end of the search engine should be able to support typing keywords in an Indian script.