Search Engines in Indian Languages

Dr. Sudeshna Sarkar
Dept of Computer Science & Engineering,
Indian Institute of Technology Kharagpur

Abstract

In this talk we will discuss the issues related to search by Indian language users. Issues pertaining to both monolingual search where the query language and the target language are the same, as well as the case of cross language search where the query and target languages may be different will be discussed. I will first give a brief overview of a search engine, and its various components. I will then discuss issues related to monolingual search in Indian languages. For example, we will like to discuss what components of the search engine must be modified for a searcher to query in Hindi, and to be able to retrieve documents in Hindi that are present in the Web. The role of a stemmer, spelling variation rules, dictionary, thesaurus and other modules will be discussed in this connection. The next part of the talk will focus on cross-language retrieval and access. This is the case when the query is given in one language but the information may be retrieved from other languages. Various techniques for effective cross language retrieval will be discussed. This will include dictionary based approaches as well as approaches based on comparable corpora. We will also discuss the status of these systems and components in various Indian languages.

About the Speaker

Dr. Sudeshna Sarkar completed her B.Tech. in Computer Science & Engineering in 1989 from IIT Kharagpur, MS in Computer Science from University of California, Berkeley in 1991, and PhD from IIT Kharagpur in 1995. She has served in the faculty of IIT Guwahati and at IIT Kanpur before joining IIT Kharagpur. Her broad research interests are in Artificial Intelligence and Machine Learning. She is currently working in the fields of natural language processing, multilingual question answering, personalized retrieval, and recommender systems. Some related projects on which she is currently working on are Cross language information access, Machine Translation between Indian languages, NER and POS tagging, and Multilingual processing of languages. She is the principal scientist of Minekey, a company incubated at IIT Kharagpur that provides personalization services for content discovery.