CHAPTER 1

INTRODUCTION

This chapter presents an overview of existing systems and motivation for the presented work. It is followed by the organization of contributions in the form of chapters in the rest of the thesis.

1.1 Introduction

The decade of 1990’s witnessed the inception of WWW/ Web 1.0 by Tim Berner-Lee (Berner-Lee, 1998). The concept enabled connection of all kinds of documents available on the web. However, the documents which were linked were more information centric / static HTML web pages. The web was thus popularly known as surface web/static web or information centric web. Although a great invention at that stage, Web 1.0 provided very little support for user interaction. The need to boost user interaction over web led to the evolution of Web 2.0 popularly known as people-centric web/read-write web also known as social web. However increase in user interaction over the WWW led to an exponential proliferation of information leading to the problem of information overload. Information overload refers to a situation where users are overwhelmed with abundance of information. The problem of information availability thus transformed into the problem of obtaining relevant information.

Search engine technology came into being as an initial attempt (Berghel, 1997) to deal with this challenge. Web search engines, solved the problem of information overload to some extent by indexing the resources available on the WWW and making it available to the users. However, search engines have two major problems associated with them; incompleteness and relevance. Incompleteness refers to the inability of the prevalent search engines to provide complete information to the user. Bergman (2001) identified two dimensions of WWW - Surface web and Deep web. Surface web refers to all the static web pages whose content is directly visible to the user whereas deep web refers to the dynamic web applications whose
content is hidden inside the databases and can be accessed only via dedicated search engines provided by these dynamic web applications. Prevalent search engines such as MSN, Yahoo and Google etc employ generic web crawlers which have limited capabilities to submit the queries to the dynamic web applications and index their content pages. Thus these search engines are unable to provide complete information to the users.

Lawrence (2000) highlighted the “one size fits all” approach of existing search engines which limits the diversity of results retrieved by these systems. For short and ambiguous queries, these systems fail to provide desired search results. For example, for a query “mouse” conventional search engines return a list of most popular urls that contain the specified keyword (i.e., mouse). However, the semantics of word “mouse” varies according to its context of use e.g. mouse pads (Business), mouse mammals (Biological), Mickey Mouse (Cartoon) and mouse device (Computer) etc. Therefore extra efforts need to be directed towards refinement of such search queries by appending additional terms and filtering out irrelevant results (Aridor et al, 2000). The concept of Semantic web/ Web 3.0 (Berner-Lee, 2001) focuses on development of technologies which can be used to understand the precise information needs of the users so as to provide relevant content. However, user’s needs are contextual in nature for example for query ‘apple’, semantics will vary according to the user’s current need context. Thus, a need is felt for web search systems that can identify the user’s contextual information to understand the underlying intent of the users query and provide relevant results by indexing the complete web (both surface and deep web).

This thesis presents a distinct framework of Contextual Web Search using Nature Inspired Algorithms. It focuses on the user dimension of the context especially the user’s personal context and social context in order to serve context aware content from both the surface and deep web.

1.2 Motivation

In the current scenario search engines are considered as primary tools for finding answers to questions, sometimes positioning themselves above books. However, existing search engines flood the user with varied set of results drawn from different domains, irrespective of the users’ desires. The user is thus left grappling for relevant information in
the sea of provided results. Besides, conventional web search systems index only surface web urls where a huge part of the web is hidden behind the search interfaces. Extensive studies have been performed to improve the performance of web search systems such as understanding and incorporating context (Schilit, 1994), context aware web search (Allan et al, 2002), deep web search (Bergman (2001), and nature inspired algorithms (Menczer and Belew, 2000) for web search etc. However, in these studies researchers focussed on individual issues.

A need is felt for a unified system which identifies user’s immediate context and adapts this context to retrieve relevant content from both the surface and deep web. This work presents unified model of contextual web search from both the surface and the deep web. The model employs Shuffled Frog Leaping Algorithm (SFLA), a nature inspired algorithm to enhance the contextual retrieval. Subsequently, presented improvisations of existing nature inspired algorithm namely Shuffled Frog Leaping Algorithm are deployed to further enhance the performance of context aware web search systems.

1.3 Outline of Thesis

The organization of the thesis is as follows:

Chapter 2 reviews the work done by various practitioners in the field of contextual web search. The extensive literature survey discusses the various perspectives of context, nature inspired algorithms and the role of agents in surface and deep web search systems. A comprehensive review of the work done in this area indicates the existing problems which form the motivation of our work.

Chapter 3 presents the architecture of contextual web search based on Belief-Desire-Intention MultiAgent architecture. The model adapts user’s personal context to perform contextual retrieval from both the surface and deep web. To identify the best algorithm for contextual retrieval, existing nature inspired algorithms are analyzed. Due to the limited capability of prevalent tools to simulate these algorithms over benchmark test problems, a Swarm and Evolutionary Optimization (SEVO) toolbox is developed and validated.
Subsequently, comparative analysis of nature inspired algorithms is performed using the SEVO tool.

**Chapter 4** presents the techniques used to perform contextual retrieval from both the surface and deep web in BDI model of contextual web search discussed in Chapter 3. Detailed methodology of Shuffled Frog Leaping Algorithm based contextual meta-crawler for both the surface and deep web, semantic query formulation, deep web classifier are subsequently discussed and validated by experimental evaluation. A comparison of presented techniques with Genetic algorithm over the varied contextual domains such as Health, Computer, Travel and Consumer Electronics substantiate the efficacy of presented approach.

**Chapter 5** discusses the users’ personal and social context in deep web search systems. The chapter includes the proposed approaches to solve the inherent problems (such as cold-start, scalability and sparsity) of both the content and social context based deep web search systems. To further augment the performance, ‘Web of Trust’ and demographic context of users are studied. The efficacy of proposed approaches is evaluated over MovieLens and Epinion datasets with respect to Case Based Reasoning (CBR), K-means, genetic algorithm and memetic algorithm.

**Chapter 6** presents the two extensions of conventional shuffled frog leaping algorithm (SFLA) namely ISFLA (Improved SFLA) and SRSFLA (Self-Reformed SFLA), to further enhance the performance of contextual web search systems. ISFLA and SRSFLA have been evaluated with respect to established and contemporary nature inspired algorithms over continuous uni-modal and multimodal benchmark test functions. The chapter also deliberates in the enhancement of contextual web search systems using SRSFLA.

**Chapter 7** concludes the work contributions presented in the various chapters and highlights the future directions of the presented work.