Chapter 16
Qualitative Pre-Processing for Semantic Search of Unstructured Knowledge

Shyamanta M. Hazarika
Tezpur University, India

ABSTRACT

This chapter describes semantic search of unstructured data through a qualitative pre-processor. Using the spatial representation language Region Connection Calculus, qualitative relationships inherent in the background knowledge is made explicit. The pre-processor obtained by exploiting such qualitative information can overcome some fundamental problems associated with information retrieval and is an efficient approach to retrieve relevant results.

INTRODUCTION

There has been a growing interest in treating *knowledge* as a significant organizational resource. Consistent with the interest in organizational knowledge, *knowledge management* has been constantly evolving over the last few years (Nonaka, 2005). Identification, acquisition, storage, dissemination, development and utilization of knowledge have been identified as the main tasks of knowledge management (Abecker and Decker, 1999). The advent of high throughput data acquisition technologies as well as digital storage technologies have made it possible to gather and store large amount of data in electronic form resulting in vast amount of information. Accurate and speedy access is required to such repositories, as knowledge is acquired by retrieving information (facts, text, document, images, audio, video etc.).

Knowledge Management Systems have to struggle to overcome the information overload. Finding particular piece of knowledge within large knowledge repositories can be very difficult. Two related problems to do with knowledge retrieval are - a. the issue of finding knowledge again, once it has been stored and b. the problem of retrieving the subset of content from the repository that is relevant to a particular problem. A large section of researchers within knowledge management is

DOI: 10.4018/978-1-60566-954-0.ch016

working on developing efficient tools for retrieving knowledge from such repositories.

Knowledge is seen as a mixture of various elements, which are sometimes codified and sometimes tacit. Knowledge is communicated through database entities, documents, workflows and e-mails etc (Dustdar, 2005). Such an item that allows knowledge to be communicated independently of its holder is a knowledge artifact. Knowledge artifacts may either be of structured, semi-structured, or unstructured nature (Hahn and Subramani, 2000). For subsequent discussion in this chapter, knowledge is interpreted as pieces of information (serving a specific need). The focus here is to retrieve these pieces of information.

Search and retrieval from unstructured information resources is usually by *best match*, i.e. to find documents which are *most relevant* to a given query (e.g. web searching, bibliographic database searches, digital libraries). Majority of unstructured information is either personal or developed through interaction between two or more individuals (Hicks et. al., 2002). Organizing and leveraging content from unstructured information has become a key concern today. There have been several attempts to bring 'structure' to this unstructured information. *Ontology* is used to capture implicit knowledge of the knowledge workers and to associate it with knowledge artifacts for classification, search, and browsing purposes (Decker et. al., 1999; Wang et. al., 2003). On the basis of ease and representational completeness, (Yang and Oh, 1993; Ounis and Pasca, 1998) argue in favor of general and intuitive knowledge representation formalisms such as *conceptual graphs*. Flexible and precise knowledge representation and retrieval can be achieved through *knowledge representation languages* that support logical inference. Many metadata languages (to support such inferences) are being developed to let people index Web information resources with knowledge representations (logical statements) and store them in Web documents (Martin and Eklund, 2000). However, these metadata languages are insufficient to satisfy several requirements necessary to allow precise, flexible, and scalable retrieval.

There has been work on design of knowledge management framework based on improved *information retrieval* (IR) systems (Yang et. al., 2002). An IR system helps to give answers to questions or find solutions by making relevant documents available via search queries. Each document contains search words (or descriptions) which can be used to describe the contents of the document. However, a user may want to retrieve on the basis of the conceptual content. The individual words provide unreliable evidence about the conceptual topic or meaning of a document. The two major problems of information retrieval systems are that of a. *synonymy* - literal terms in a user's query may not match those of a relevant document and b. *polysemy* - term in a user's query will literally match terms in documents that are of no interest to the user. To overcome such problems and provide relevant results for queries, a technique known as *latent semantic indexing* (LSI) (Deerwester et. al., 1990) have been developed. Improvements on LSI (Cegloski et. al., 2003; Dumais, 1990) have been reported in the literature.

Relevance has an important bearing and is crucial to information retrieval. Within an IR system, the retrieval strategy is to retrieve all the relevant documents, at the same time retrieving as few of the non-relevant as possible. Therefore, when characterisation of a document is worked out for subsequent use within an IR system, it should be such that it enables the document to be retrieved (when the document is relevant) in response to a query. One needs to construct a technique within which relevance decisions can be quantified. Research in IR for efficient and relevant retrieval frameworks seems to have been concerned with different aspects of such a technology (Deerwester et. al., 1990; Cegloski et. al., 2003; Dumais, 1990).

This chapter presents a method of searching unstructured knowledge artifacts. The technique presented is based on the *qualitative spatial*