CHAPTER 3

RESEARCH QUESTIONS AND METHODOLOGIES

3.1 Research Questions

Chapter 2 discussed various information analysis methods for building information retrieval systems these range from indexing to classification and from manual to automatic processing. The purpose of such systems is to provide knowledge and mechanisms that allow users to access underlying documents relevant to their individual information needs. However, the information analysis process does not incorporate information need, making the knowledge embedded in systems static and unable to encounter ever-changing user information needs. In addition, because systems usually employ specific search knowledge and system knowledge and possess specific domain knowledge, they raise extra barriers to user access to needed information.

The main difference between today's online information retrieval process and the traditions maintained through hundreds of years of history in library science and information science is that they rarely involve information consultants such as librarians, and information consultants can help users articulate their specific

information needs and eliminate users' having to know the specific searching mechanism and system knowledge of a particular information retrieval system. These changes are directly from the availability of Internet technology and the use of the World Wide Web. During its 40-year history, information retrieval research has implied that the advancement in information analysis and retrieval mechanisms will overcome the deficit of not having help from information consultants. Nonetheless, as discussed in Chapter 2, difficulties in expressing information needs to information retrieval systems and perceiving knowledge contained in the retrieved results persist. Information retrieval process is still a time-consuming and labor-intensive effort.

Information seeking is an intrinsically cognitive process undertaken to fulfill information needs raised by users. It is a semantics-bearing and interactive process in which we can simply ask a colleague for some information or try to retrieve it from some information retrieval system. My research focus is a machine's semantic processing, rather than its mechanical operation, of information retrieval.

The board question is, can knowledge sources be used to help users express their information needs? How to use and how to generate such knowledge sources are the two focuses of this dissertation. The specific question are as follows. Would the *automatic concept exploration process* be able to help users identify more relevant concepts? Would such a process be able to perform more efficient exploration of a concept space than the conventional manual browsing method? If so, which

algorithmic methods - symbolic-based branch-and-bound or neural network-based Hopfield net algorithm - is better in terms of gathering relevant concepts from knowledge sources? Would the *concept space consultation* process provide a *semantic medium* to reduce the *cognitive demand* from users in terms of *elaborating information needs*? Would the *concept exploration process* be able to help users find more relevant documents? With regard to computing scalability, would the technique of computer generation of concept spaces be applicable to very large textual databases? With regard to domain specific knowledge scalability, would concept space generation by technology create satisfactory domain-specific concept associations from corresponding textual databases? How does the quality of concept associations in concept space generated from very large textual databases compare with that of a man-made domain-specific thesaurus?

3.2 Methodologies

The research problems being looked at involved an effort to develop a semantic component which is capable of resembling semantics used by humans and to create a systematic process to make use of the semantic component to aid the information retrieval process. Two methodologies which seemed most appropriate for this research were the systems development approach (Nunamaker et al., 1991) and experimental design.

In the systems development methodology, a problem is first identified and then an automatic or system-based solution is analyzed, designed and implemented, typically using one of the systems analysis and design paradigms such as object-oriented analysis and design. The resulting system solution is then tested in a laboratory or real-world setting.

In order to validate the usefulness of a system or automatic technique, its results must be compared against existing systems, existing standards or existing techniques. The experimental design methodology provides guidelines such as sample size and confidence level in conducting experiments to collect data for measuring against the underlying objective. In information retrieval, precision and recall are two measures commonly used to evaluate the performance of different information retrieval techniques. However, several studies reveals problems with both (Raghavan et al., 1989), (Tague-Sutcliffe, 1992), (Hersh, 1994). The first problem is that the knowledge of the proper estimation of relevancy of all the documents in a large collection is unavailable, which implies that recall cannot be estimated precisely. Relative, instead of absolute, measures must be used for large collections. The second problem is that recall and precision originally were designed to measure the effectiveness over a set of queries processed in batch mode. However, since almost all retrieval systems are processed interactively, the time factor should be considered in determining effectiveness.

3.3 Dissertation Plan

My dissertation aims to investigate the use of semantic-enabled information technologies to help users seek information. In order to facilitate understanding of the research and its context regarding to various relevant information technologies, I present a framework to guide the research and discussion. Its two core semantic-enabled information technologies are *concept consultation system* and *concept association analysis* (as one of the knowledge discovery methods).

3.3.1 A Framework for Semantic-enabled Information Technologies

Based on past research and our experience with various knowledge discovery methods, I developed a framework to depict the different entities involved in knowledge discovery and their intertwining relationships.

The main entities involved in this framework include the underlying *databases* from which knowledge is acquired, *knowledge discovery* methods, the *knowledge bases* discovered by the algorithms and those imported from other domain-specific sources, *information retrieval systems* for accessing documents, and *concept consultation systems* for users. The definitions of databases and knowledge bases are sometimes blurred because of their historical roots in different disciplines such as artificial intelligence, database management systems, and information retrieval. In the context of this research, *databases* refer to online repositories of basic facts

about objects and events in the world, e.g., employee files, transaction records, bibliographic records, etc. and *knowledge bases* are online repositories of high-level, abstract human knowledge represented in terms of heuristics, inferencing rules, problem solving strategies, networks of inter-related concepts (concept space), and so on. These entities and their relationships are shown schematically in Figure 3.1 and discussed in the context of earlier research.

- *Databases:*

  Domain-specific databases that capture information or data of relevance to users' applications can be taken as major sources of knowledge. Since the '80s, many major corporations and government agencies have used databases of drug side effects, retail shopping patterns, tax and welfare frauds, frequent flyer patterns, and so on to identify application-specific knowledge. Frawley et al. (Frawley et al., 1991) presents a good overview of databases used in knowledge discovery.

  The enormous sizes of real-life databases, which have frequently prevented human beings from conducting labor-intensive analysis, and the availability of unused computing cycles in many institutions have prompted the use of computers for knowledge discovery (Parsaye et al., 1989), (Frawley et al., 1991). Massively parallel computers, and even supercomputers, have

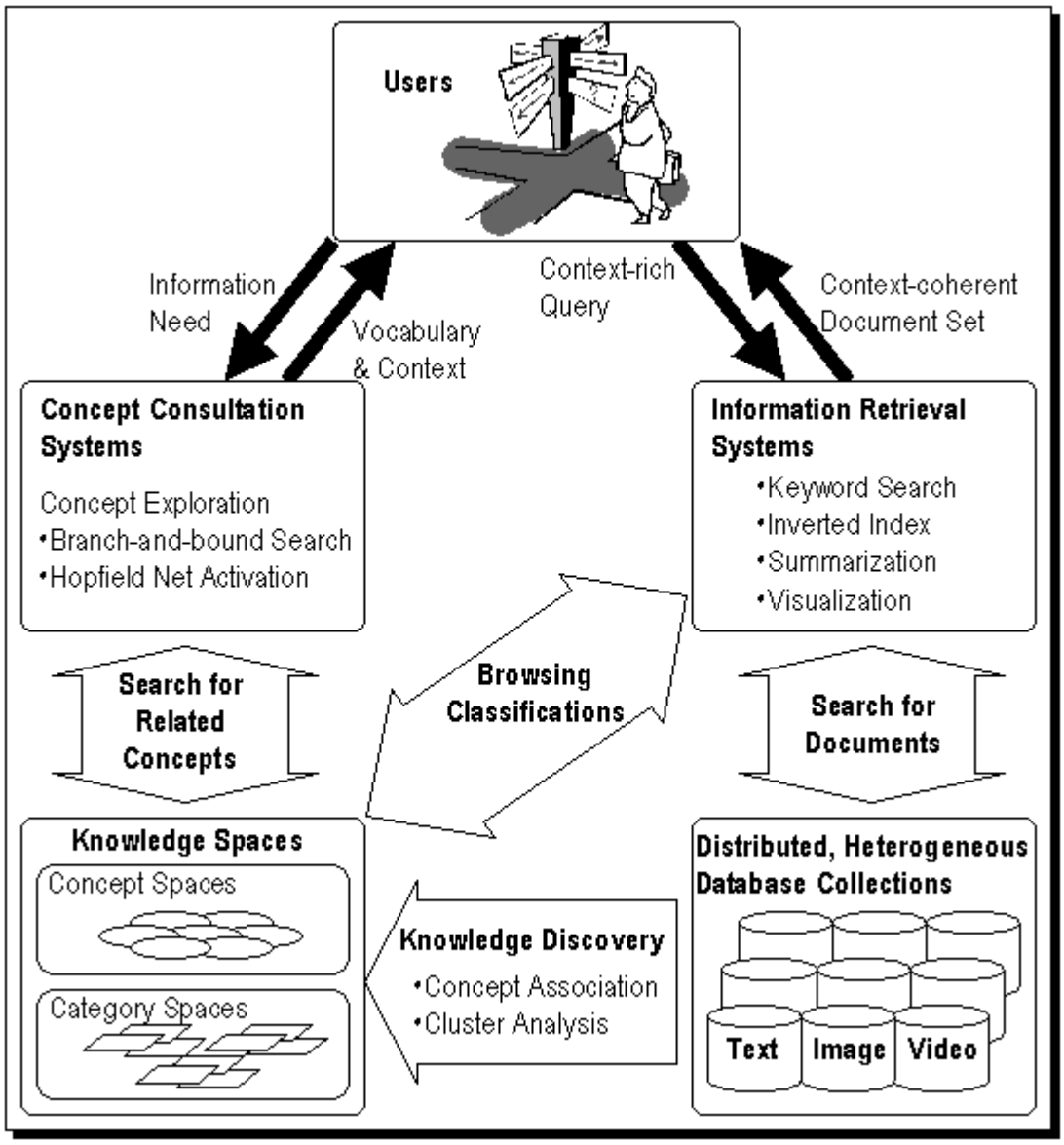

Figure 3.1: A framework for semantic-enabled information technologies

also been considered for analysis of some really large business or scientific databases.

- *Knowledge discovery methods:*

  Various knowledge discovery techniques have been developed over the past few decades by statistics, information science, and artificial intelligence researchers.

  Statistical algorithms typically examine quantitative data for the following purposes (Parsaye et al., 1989): clustering descriptors with common characteristics, e.g., factor analysis, principal components analysis (Morrison, 1976), and cluster analysis (Everitt, 1980); hypothesis testing for differences among different populations, e.g., t-test and analysis of variance (ANOVA) (Montgomery, 1976); trend analysis, e.g., time series analysis (Nelson, 1973), (Morrison, 1976); and correlation between variables, e.g., correlation coefficient and linear/multiple regression analysis (Montgomery, 1976).

  Recently, classical symbolic AI learning algorithms such as ID3 (Quinlan, 1983) and AQ (Michalski and Larson, 1978) and resurgent neural net learning algorithms such as Backpropagation (Rumelhart et al., 1986) have provided new perspectives for knowledge discovery. These techniques allow effective analysis of both qualitative and quantitative data. Unlike the statistical approach, which typically is based on some underlying models, assumptions,

and stringent conditions, many AI-based techniques are more flexible, easier to use, more powerful, and produce output that is more meaningful to users. (For a complete overview of the AI-based learning techniques, readers are referred to (Carbonell et al., 1983), (Dietterich and Michalski, 1983), (Knight, 1990), (Frawley et al., 1991).)

My research focuses on a knowledge discovery method for semantic-enabled technology based on the statistical co-occurrence analysis (Chen and Lynch, 1992) or context analysis (Attar and Fraenkel, 1977) in the field of information retrieval.

- *Discovered knowledge and other knowledge sources:*

  In addition to the mathematical formulas and parameters produced by statistical techniques, symbolic AI-based techniques produce outputs that are based on traditional knowledge representation schemes such as semantic net, frame, decision trees, and logic (Parsaye et al., 1990), (Quinlan, 1983), (Michalski and Stepp, 1983). Because most AI-based knowledge representations are grounded on cognitive research (human memory, problem solving, story understanding, production systems, etc.) (Anderson, 1985a), they are often considered more natural and understandable (for users) than statistical formulas or neural nets.

The amount of knowledge discovered by various knowledge discovery methods (e.g., number of rules produced, number of nodes and links of a system-generated semantic net or neural net, levels and branches of a decision tree, etc.) sometimes may be substantial. For some applications, it may also be necessary to include such domain-specific knowledge bases, expert systems, corporate rules and guidelines, data dictionaries, and external thesauri in an institution's complete knowledge repository – shown in Figure 3.1 as the *Knowledge Space*.

My research focuses on the generation of *Concept Space* as a knowledge source and a part of the semantic-enabled information technologies.

- *Information retrieval systems:*

  Information retrieval systems are commonly known as *search engines* nowadays because of the popularity of several major Internet search engines like Yahoo!, Altavista and Northern Light (Baeza-Yates and Ribeiro-Neto, 1999). The ability to index hundreds of million of web pages does not translate well to the ability to present what users want. Oftentimes, users are overwhelmed by millions of potentially retrievable web pages. The common practice for dealing with this overloading problem is to give the top ten or twenty ranked web pages. Some better solutions work toward information visualization

and summarization (Card et al., 1998), (Shneiderman, 1997), (Tufte, 1990), (Tufte, 1997). over retrieved data set.

- *Concept consultation systems:*

  Concept consultation systems are system-supported *concept exploration* or knowledge management tools for users. In order to manage and utilize the different knowledge bases within a knowledge space, it is necessary to develop some high-level, friendly, and efficient system-supported methods or tools for users. In Figure 3.1, while knowledge discovery methods help generate knowledge bases from databases; concept consultation systems make use of knowledge sources and provide users with *concept exploration* methods to manage and utilize these knowledge bases effectively.

  These concept exploration methods need to be highly *interactive* so that users can explore in the knowledge space freely and efficiently, articulate their conceptual models, and use whatever knowledge is relevant to their applications or tasks. While *knowledge discovery* methods are *data-driven*, *concept exploration* needs to be *user-driven*.

  Kaufman et al. (Kaufman et al., 1991) suggested a *knowledge management* component for knowledge discovery applications that can assist in the operation and use of knowledge bases and is similar to the *database management* function of commercial database management systems (DBMS). They

proposed counterparts of the relational DBMS operators such as SELECT, PROJECT, JOIN, and INTERSECT for knowledge management. However, the proposed functionalities and roles for these operators remained vague and these authors have suggested further research investigating different high-level operators for different representation schemes. My research has been aimed at examining in detail an "ACTIVATE" operator on semantic net and neural net representations that can assist in spreading activation based inferencing. It focuses on the investigation of concept exploration methods as semantic-enabled information technology for users to communicate their information needs.

3.3.2 Structure of Dissertation

Based on the framework shown in Figure 3.1, the next three chapters describe three distinctive but closely interrelated pieces of research to investigate the creation and use of semantic-enabled information technologies to help users seek information.

Chapter four describes the use of the semantic component (concept space) to help a user express an information need. This chapter directly relates to the *concept consultation systems* in the framework. Two algorithms - branch-and-bound and Hopfield Net - are employed to explore within the concept space. The purpose of using these two algorithms is to automatically gather closely related concepts for

presenting to a user who may choose to use any number of related concepts to better express a query. Experiments are also performed to evaluate the performance of each exploration methods and its use.

Chapter five describes the process of identifying and extracting concepts from textual documents as basis for further semantic analysis. A concept is represented by term phrases appearing naturally in sentences. Co-occurrence analysis, one of the context analysis techniques, is then performed on the entire document collection. This context analysis calculates the strength of relationship between each possible pair of co-occurring concepts. Essentially, it automatically builds a thesaurus of concepts, called a concept space, with statistical relationships rather than traditional semantic relationships. Experiments are performed to evaluate the performance of the algorithmical strength of relationship between concepts. This chapter corresponds with the *knowledge discovery* methods and *knowledge sources* components in the framework.

Finally, chapter six presents my conclusions and thoughts for future exploration of semantic research. Appendix A presents a detailed benchmark testing of the two algorithms used for concept exploration. Appendix B presents a sample interactive session of the concept exploration process. Appendix C shows a web-based information retrieval system called *Cancer Space*. Appendix D lists the funding sources and specific personal acknowledgements for various stages of research in this dissertation.