CHAPTER 2

LITERATURE REVIEW

This chapter reviews previous research on two aspects of information searching: library collection and user information need. The first section examines various concerns and techniques employed to characterize an individual document object or a collection of such document objects. The core of an information retrieval system is its analytical capability to dissect, digest, and derive knowledge from underlying records. The second section is devoted to investigating user information needs with respect to the information seeking process. In general, information retrieval systems are able to manage underlying information and even derived knowledge. However, such systems' capability remains static in comparison with the dynamic nature of user needs.

2.1 Static Nature of Knowledge in Library Collection

Recent development in research and technology has advanced information retrieval systems to enable them to handle multimedia objects such as images (Ma and Manjunath, 1998), audio (Witten et al., 1999) and video (Wactlar et al., 1999). However, the discussion in this paper mainly focuses on textual documents, which include bibliographical records and full-text articles, although it demonstrates some parallelism with analysis of multimedia systems.

Two levels of document analysis are commonly studied and performed by information scientists and practitioners. The next sub-section reviews how a single document is characterized, while the second sub-section depicts how a document collection is analyzed. Instead of focusing on documents, the last sub-section examines concepts or terms existing in documents.

2.1.1 Characterizing Document Objects

2.1.1.1 Theory of Indexing

The dual purposes of indexing an document are to represent a lengthy and structureless textual record by a set of indexes (*atomic elements*) and to access a set of textual records through their indexes (Salton, 1975). Research on indexing has been focused on defining a set of *good* index terms as well as assigning a set of *good* index terms to a particular document. Ideally, the choice of such *good* index terms should collect all relevant documents to yield high recall and simultaneously distinguish them from irrelevant ones in order to give high precision. However, in reality, the choice of index terms always exhibits the phenomenon of the well-known inverse relationship between recall and precision (Salton, 1975). Achieving high recall is generally at the cost of low precision and vice versa.

Lancaster has shown that the rate of growth of information continues at an exponential pace, while the corresponding rate of growth over the same period of time for number of concepts (index terms) converges logarithmically (Lancaster, 1986). Chen has summarized this phenomenon as a logarithmic vocabulary growth principle that sheds light on the *information overload* problem (Chen, 1994). Nonetheless, as the growth of concepts moves farther into the flat region of the plateau, the once manageable volume of concepts led by the logarithmic vocabulary growth makes it difficult to discriminate information related to each concept, which increases at the same exponential pace. This phenomenon can be easily demonstrated by using a web search engine, which returns hundreds of thousand or more web pages from a simple query (Kirsch, 1998).

Based on the document frequency of index terms, Salton has suggested a model for construction of *good* index terms (Salton, 1975), (Salton and Yu, 1973). The model divides terms into three groups according to their document frequencies: low, medium, and high. Given a document collection, all index terms can be ranked by their document frequencies and listed from the left (low frequency) to the right (high frequency). *Good* terms fall into the medium document frequency range. Terms in the high frequency end are considered the *worst* index terms because they do not have discriminating power. Terms in the low frequency end are called *poor* index terms characterized by poor performance on recall. Even though it was developed with two small collections of 450 and 1,400 documents each, the model created by Salton (1975) generally holds true with the World Wide Web's so far approximately 100 million web pages (Kirsch, 1998), (Schwartz, 1998). Of course, the document frequency range for *good* index terms varies between different document collections. However, not much research has been done to investigate what a *good* range would be. Instead, many search engines, especially web search engines, have adotped various strategies to rank hundreds of thousand retrieved documents and give the best 10 or 20 to users (Schwartz, 1998).

With his model for *good* index terms, Salton has proposed two different mechanisms to convert the *worst* or *poor* terms into *good* ones (Salton, 1975). The purpose is to make a collection consistently made up of *good* index terms. The first mechanism is called the *right-to-left phrase construction.* The idea is to transform high frequency terms (on the right end of the document frequency range) into *units* with lower frequency (toward the left to the middle) in order to improve their precision. Such desirable units are *term phrases.* A classical method to do so is to generate *phrases* consisting of several combined terms (Salton, 1988). A restricted, practical, and automatic version of this classical method is to use consecutive adjacent words to form *phrases* (Chen and Lynch, 1992). For example, in a computer science collection, the terms *program* and *language* may be insufficiently specific, particularly when assigned to a large proportion of the documents in a collection. The phrase *programming language* is more specific and may, when assigned to the documents, lead to improved precision output.

The second mechanism is called the *left-to-right thesaurus transformation* (Salton, 1975). The goal is to transform low frequency terms (on the left side) into *units* of higher frequency (toward the right to the middle) in order to improve their recall. Such units are generated by grouping a number of the low-frequency entities into classes. The term classes are then characterized by frequency properties equivalent to the sum of the frequencies of the individual components. A classical way of combining individual terms into classes is by means of a *thesaurus.* Such a thesaurus specifies a grouping of the vocabulary in which items included in the same class are normally considered to be related in some sense – for example, by being synonymous, or by exhibiting closely similar content characteristics. The success of this method relies heavily on the availability of a *good* thesaurus in a given document collection or domain for a given time frame. In practice, low frequency index terms in a very large collection such as World Wide Web, indeed, give quite desirable performance in terms of precision, with the trade off of forgiving and forgettable document recall. Such low frequency is considered to be in hundreds range, compared with the usual hundreds of thousands range.

In addition to size of collection, number of index term assignment has a direct impact on constructing an index having a *good* range. Traditionally, content providers hire human indexers to assign three to six index terms to a document after reading it, hoping of calibrate the choice/usage of controlled vocabularies and the accountability of each index term. Over time, as information technology becomes more capable and affordable, the number of index terms assigned to each document increases. A certain number of free-text term phrases are included as index terms. An extension of free-text indexing, full-text indexing, is available in some information retrieval systems.

2.1.1.2 Manual Indexing

Manual indexing is part of the summarization process professional indexers typically perform on journal articles (Endres-Niggemeyer and Neugebauer, 1998). The whole process requires a expert summarizer to read an article and perform two tasks: abstracting and indexing. Abstracting includes reading, taking notes, drafting an abstract, revising the draft, and writing the final version (Rowley, 1988). The task of indexing is tightly associated with the classification process, which also involves subject analysis, translation into the indexing language, and construction of a register entry (Langridge, 1989). While abstacting produces a relatively long textual summary in natural language format with sentences or even paragraphs, indexing gives a list of discrete term phrases or concepts to represent core ideas in an article and classification assigns one or more artificial codes according to a classficiation scheme. Each summarized record becomes an entry to a bibliographic information system.

Manual indexing is very cognitively intensive but mechanical (Endres-Niggemeyer and Neugebauer, 1998). A study by Endres-Niggemeyer and Neugebauer found that a well-trained indexer deals with each article independently, reading through it only once, an indexer writes down what is noteworthy and creates an abstract. However, the study falls short on describing in detail how index terms are selected by an expert indexer.

In the field of information science, Bates summarizes a consistent and historical phenomenon that indexers simply index what is in the record (Bates, 1998), directly reflecting the fact that a document is known and visible to an indexer. Factual information can be checked directly and immediately to create an absolutely accurate bibliographic record. In addition, indexers are trained to use a specific indexing system and vocabulary, generally establishing rules for resolving debatable situations such as which term is to be used rather than the other when there are two closely related concepts. The achieved preciseness of manual indexing has the drawback of consistency of relying on human perception and interpretation, as has been reported in various studies (Cooper, 1969), (Sievert and Andrews, 1991), (Chan, 1989).

Because human judgment requires manpower, the number of index terms assigned to a record is normally a matter of policy driven by cost (Plaunt and Norgard, 1998), also including factors like storage space, computing power for the search process, and search performance results after human indexing has assigned some number of authorized index terms to each document as it enters a particular system.

2.1.1.3 Indexing with Controlled Vocabulary and Thesauri

Using controlled vocabulary is a common practice of human indexing. After identifying potential index terms to be assigned to a document, an indexer selects final index terms by consulting a list of controlled vocabularies that have either been constructed by a group of professionals and experts such as medical researchers and practitioners or by an information provider and organizer such as library. For example, Medical Subject Headings (MeSH) has been created by National Institutions of Health (NIH) (Lindberg et al., 1993), (McCray and Nelson, 1995). Library of Congress Subject Headings (LCSH) have been generated by Library of Congress.

In addition to defining controlled vocabularies, semantic relationships between such vocabularies are constructed manually. Semantic relationships commonly include broader term, narrower term, synonym, related to, used for, and uses. In order to achieve the completeness of characterizing a document, some information provider utilizes semantic relationships in thesaurus to automatically bring in related terms of those assigned index terms. This practice can easily expand five assigned index terms to twenty or thirty index terms for a document. Petroleum Abstracts is one example (Finnegan, 1991), (Bailey, 1994). They carefully calibrate the induced terms in the bibliography record.

The use of controlled vocabulary for indexing reduces the wording variation chosen by different indexers. That is, as long as a concept is deemed to be important for a document, a precise index term will be assigned to that document, thereby eliminating problems from morphological variations such as *lung tumor* and *tumor of the lung* (Jacquemin and Tzoukermann, 1999).

The main drawback of using controlled vocabulary is the limitation on assigning new terms to reflect new concepts. Usually there is a delay incorporating new terms into controlled vocabulary as well as thesauri.

2.1.1.4 Free-text Indexing

Free-text indexing relies totally on words or phrases found in a document. The technique can be applied to full-length documents, abstracts, titles, and combinations of them. Nowadays, it currently is usually performed with automatic methods, is also used manually in some bibliographic systems such as INSPEC. When it is employed manually, term phrases (commonly of two or three words) are carefully selected from sentences. Occasionally, four-word or five-word term phrases may be used. Whereas manual effort limits the number of free-text indexes to be associated with a record, the automatic practice commonly demonstrated by popular web search engines uses all non-stop words to index each record. Term-phrase search may be supported by the adjacency of words in records or using some term formation technique to create term phrases from text as indexes (Salton, 1988), (Chen and Lynch, 1992).

The main advantage of free-text indexing is that it allows a document to speak for itself. It captures authors' wording which generally very up-to-date. In the case of term phrases, indexes also are retained in their natural language format, mainly as noun phrases that may represent precise concepts. In addition, automatic free-text indexing provides the most complete index coverage to all records in a information system.

Nonetheless, the completeness of free-text indexing is one dimensional - exact words or phrases inside each document. Two documents with similar content but different vocabularies will have different free-text indexes. Two documents on the same topic may have only a small portion of their manual indexes in common. This leads to diminished recall value when relevance is measured beyond the syntactic level.

On the other hand, the massive volume of free-text indexes may weaken the precision of retrieval. Some words and even term phrases may have multiple meanings. They may exist in two unrelated documents. A retrieval process with many general terms brings unrelated records together to the detriment of precision and the inconvenience of an overwhelming number of retrieved records.

2.1.2 Characterizing Global Knowledge in Document Collections

While analyzing a single document is undertaken to reveal the knowledge it contains by enlisting embedded concepts, analyzing a collection of documents is done to discover the overall but hidden knowledge under the conglomerated effect. The characteristics of the knowledge of a collection identify its relevance, completeness, and proper usage. Because the cumulated information is so voluminous, the characterizing process involves intensive resources and expertise, but the return on investment fortunately is an understandable summary of an ever-growing information in organized scheme.

The following three sub-sections describe different techniques and resources that can be used to characterize document collections. The goal of the first two of these (classification schemes and knowledge discovery) is to explicitly reveal the characteristics of a particular document collection in its entirety. The main difference between them is that classification schemes rely solely on manual effort while knowledge discovery relies heavily on automatic computational power. On the contrary, the third technique converts the global characteristics of a document collection into functions that take the form of ontology and inferencing rules.

2.1.2.1 Classification Schemes and Categorization

Classification or classification systems have a long history of being used organize large amounts of information in a managable manner. Library systems use the

Dewey Decimal System or the *Library of Congress Classification System* to organize their collections physically and conceptually (Kao, 1995), (Kohl, 1986). The *Association for Computing Machinery* (ACM) uses its own classification scheme to organize subject areas of its interest over the past 50 plus years. The *Yahoo!* directory is one of the attempts that have been made to organize vast amount of Internet information.

Classification or categorization is the putting together of like things into their categories. For purposes of explication, we may consider a category system as having both vertical and horizontal dimemsions (Rosch, 1978). The vertical dimension concerns the level of abstraction of the category system while the horizontal dimension focuses on the segmentation of categories at the same level of abstraction. By implication, a category system having these two dimensions is intrinsically hierarchical. That is, the natural knowledge representation of a category system is a tree structure, one of the most readily comprehensible data structures to human beings. Tree structure is commonly used and seen in tables of content, family trees, and organizational charts.

A classification scheme is intended to provide coverage of all known knowledge. Since the 19th century, several classificiation schemes such as Dewey Decimal Classification, the Library of Congress Classification, Universal Decimal Classification, and Reader Interest Classification have been used in libraries to include all kinds of knowledge (Miller and Terwillegar, 1990). However, there has never been a single scheme upon which everyone agrees. In practice, each classification scheme defines its own set of categories at different levels of abstraction (Rosch, 1978). In addition, each scheme has its own choice of vocabulary. The bottom line is not which scheme is correct but which has the flexibility to extend its coverage to include new categories and sub-categories (Miller and Terwillegar, 1990).

On the contrary, categorizing a smaller information set calls for providing maximum information about the underlying collection while requiring the least cognitive effort to obtain given information within it (Rosch, 1978). In the other words, the result of categorization covers only all knowledge found inside a given information collection. Many disciplines and communities have their own category systems, such as ACM's Classfication Scheme and Compendex's Engineering Classification Scheme that offer grand coverage only to their sponsoring communities.

Traditionally, the making of category systems, like that of various classfication schemes, is very labor-intensive. Even though the goal of having a good classification scheme is to minimize cognitive effort needed to distinguish different categories, the massive quantity of defined categories requires appropriate human learning and comprehension in order to locate a classified piece of information. Fortunately, the widely known hierarchical structure provides a natural divide-and-conquer approach to directing users' attention.

In general, physical items like books go into a category in a classification scheme while the intellectual items like topics in a book go into several categories. Such cross listing or cross referencing ensures completeness of coverage by a defined category. In addition, items in the same category provide direct results to a search for similarity to a particular item, simply because only like-items go into the same category.

Although there is no “true” classification scheme or categorization, Wynar (Wynar, 1985) identifies a few criteria for a successful classification scheme. For existing information, a classification system must be inclusive as well as comprehensive. For new information, the system must be flexible and expansible. In all cases, the system must employ terminology that is clear and descriptive, with consistent meaning for both the user and the classifier. This set of criteria resembles the heuristics used by most classification schemes, which may in fact serve as evaluation criteria for automatic classification and categorization methods described in next sub-section.

Classification systems are commonly used in digital libraries and document management systems. However, classfication systems are pragmatically designed for optimum ease of human access. They do not aim at a semantically clear formal model (Abecker et al., 1998).

2.1.2.2 Knowledge Discovery from Large Databases

Similarity between documents can be computed based on the vector space model (Salton et al., 1975), (Salton and Yang, 1973). This similarity computation forms the core technique used to perform various methods of knowledge discovery from large textual databases. Chen (Chen, 1995a) discusses the use of neural networks, symbolic learning, and genetic algorithms to perform automatic characterization on large document collections. Once each document is converted into an index vector, various similarity functions such as Jaccard and Cosine coefficients used by different algorithms will compute the similarity score between a pair of documents (Salton, 1988). The main difference among various algorithms is the clustering method of lumping all similarity scores together to form *meaningful* clusters of documents.

Many clustering or categorization techniques have been applied to the field of information retrieval. These techniques include classical graph data structure and algorithms such as Ward's algorithm (Even, 1979), (Ward, 1963), statistical algorithms such as multi-dimension scaling (MDS) and discriminant analysis (Jain and Dubes, 1988), (McLachlan, 1992), symbolic learning algorithms such as ID3 and AQ15 (Quinlan, 1983), (Michalski et al., 1986), and neural network algorithms such as Kohonen's self-organizing map (SOM) and Hopfield Net (Kohonen, 1995), (Hopfield, 1982), (Lippmann, 1987). They are all capable of making some meaningful categories from document sets ranging from several hundreds to several thousands. However, large scale attempts are limited because of the scalability issue that is related to size of data set and the demanding need for computational resources such as processing power and memory. In the mid-90s, the field of information science started to make use of high performance computing resources to analyze very large textual data collection.

2.1.2.3 Knowledge Bases, Inferencing, and Ontology

Classification scheme and categorization are special cases of knowledge bases. Their knowledge is tailored to provide grand coverage of corresponding document collections. In a broader sense, knowledge bases cover knowledege in different domains or subject areas. Under the notion of knowledge discovery, at the risk of being overly recursive, facilitating knowledge mining requires providing knowledge about knowledge (Rouse et al., 1998). Ontologies and data models are used in knowledge-based and database systems, respectively, to specify the basic assumptions that went into the system's conceptualization (Gruber, 1993).

The idea behind knowledge base and database coupling is first to build a knowledge base to reflect the database and then to access the database through the use of the knowledge base. However, there is no established form of knowledge base. One explanation is that it is not necessary for such derived knowledge to be made explicitly available to users. Since such a knowledge base is a component or function of an integrated system, knowledge may freely appear in many kinds of knowledge representations such as semantic net, production rules, frames, and ontologies (Rich and Knight, 1991). Some knowledge is manually crafted. Examples are the metathesaurus of the Unified Medical Language System (UMLS) project, a thesaurus created by medical experts and practitioners (Lindberg et al., 1993), (Rada and Martin, 1987), (Humphreys and Lindberg, 1989); MYCIN, an expert system engineered by intensive knowledge acquisition in the medical field (Shortliffe, 1976); CYC, an ontology crafted by effort in gathering global and common sense knowledge (Lenat and Guha, 1990), (Lenat et al., 1990). Other knowledge is derived through automatic machine learning methods such as ID3 and Kohonen self-organizing map.

In order to make knowledge bases work with a target document collection, different inferencing mechanisms are employed according to different knowledge representations. In a production rule system, an inference mechanism relies on the control strategy built inside the recognize-act-cycle control module. Production rules, in the form of predicate logic, are brought into the control module based on some induced conditions. Conflict resolution strategies and even heuristics are utilized to decide the next action. The control module also includes a backtracking mechanism to counter any wrong decision made in the recognize-act-cycle.

In semantic or hybrid networks, different search algorithms can be used as inferencing mechanisms to traverse a *network of knowledge.* The simplest form of such networks uses *nodes* to represent concepts and links to represent semantic or probabilistic relationships between each pair of concepts. Built upon this graphic representation, a frame or script is used to extend the content of each node - constraints, exceptions, time and place information (Lehmann, 1992).

2.1.3 Revealing Knowledge in Neighborhood

While characterizing documents shows what they are and how they are related, concepts - basic elements in documents - yield another level of analysis. The following sub-sections show how a list of concepts is defined and how concepts can be used to reveal some “neighborhood” knowledge.

2.1.3.1 Syntactic Mapping: Index List

Syntactic mapping provides the capability to display an index list used in a document collection. For all information retrieval systems, such index lists are given because they are the list of keys in the inverted index to documents.

A common practice is to list a set of terms existing in a system in alphabetical order based on term fragments entered by a user (Kowalski, 1997). The user can then examine the terms on both sides of the neighborhood. This browsing process allows a user to select exact terms for the searching process. It also is feasible to type in some *approximate* pattern with *errors* to retrieve a list of system terms (Wu and Manber, 1992b), (Wu and Manber, 1992a). In addition, it is a useful for some *advanced* searches to list the number of occurrences (document frequency) to decide what terms should be used in a query.

A list of searchable terms is a given component of an information retrieval system. Issues in policy, design, and implementation require inclusion of such list as a reminder service or a dictionary service. In fact, having an index list to aid searchers is a long-time practice in books and on-line help information and offers the least expensive way to reveal some potential search terms in a general neighborhood.

2.1.3.2 Keyword Mapping: Controlled Vocabulary

Controlled vocabularies are created manually by information producers like INSPEC or organizers like the Library of Congress. In addition, controlled vocabularies can be used to assign subject headings to documents, or to support effective search, since controlled vocabulary terms can be precisely submitted as queries. Unless searchers are indexers, however, controlled vocabularies are often foreign to most searchers.

Library of Congress publishes the Library of Congress Subject Headings (LCSH) to cover vocubularies in all subject areas. LCSH is the most comprehensive list of subject headings in the world today. It provides an alphabetical list of subject headings, with cross-references and subdivisions verified by the Library of Congress. As of Spring 1999, it had a total of 234,000 headings and references (Library of Congress. Subject Cataloging Division., 1998). The *USE* relationship directs users to use proper keywords while the *used for* (UF) relationship reminds users what a proper keyword means. However, searchers need to go through a manual browsing process to locate the controlled vocabularies.

In order to simplify the use of such controlled vocabularies, some systems try to map searcher's keywords to an *internal* vocabulary list. For example, the Ovid search engine for databases like CancerLit and Medline from Ovid Tecnologies, Inc. uses a statistical analysis to map which subject headings tend to occur in documents containing a searcher's free text query. Ovid utilizes the database producer's online vocabulary or a thesaurus such as UMLS. Knowledge bases and rules can be used to help users locate appropriate controlled vocabularies from their own search terms (Shoval, 1985).

2.1.3.3 Semantic Mapping: Man-made Thesauri

In addition to providing a controlled vocabulary, LCSH is a thesaurus associating concepts through a set of semantic relationships: narrower term (NT), broader term (BT), and related term (RT) (Library of Congress. Subject Cataloging Division., 1998). Instead of focusing on the syntactic variation of terms, a thesaurus gives the semantic variation of potential search terms. In the other words, a thesaurus brings together in different forms terms for the same or similar concept. Another general domain thesaurus is Roget's Thesaurus (Roget, 1962). Different communities create their own thesauri in some specific domain areas. For example, INSPEC thesaurus is for the domains of physics, computer science and engieering (Institution of Electrical Engineers, 1993); GeoRef thesaurus is for geosciences (Goodman, 1997).

Because it requires a tremendous effort to create a thesaurus, the UMLS project tries to utilize some abstraction analysis to create a general "lightweight", but comprehensive, *Semantic Network* to cover semantic relationships between medical concepts (McCray and Hole, 1990), (Lindberg et al., 1993), (McCray and Nelson, 1995). The Semantic Network contains 132 semantic types and 53 semantic relationships. Semantic types are abstracted from all UMLS terms. Semantic relationships are created for these sets of semantic types. Although the Semantic Network covers all semantic relationship of UMLS terms through semantic types, a semantic relationship for two particular UMLS terms under two corresponding semantic types associated with a semantic relationship may not make sense.

2.1.3.4 Co-occurred Mapping: Automatic Thesauri

Co-occurrence analysis is a statistical algorithm to calculate a co-occurring weight between a pair of terms in a document collection (Salton, 1975). The creation of a list of co-occurring terms is solely dependent on what terms are being *indexed* from all documents. It also has the capability to associate two opposite concepts. For example, the sentence, "A is not B", has two negatively associated terms (*A* and *B*). However, the probability of having many such sentences in a large document collection is very small and chances of such a co-occurrence relationship are insignificant.

Strictly speaking, automatically computed co-occurrence relationships can hardly be considered semantic relationships. Nonetheless, the aggregate effect of such co-occurrence information does give a tremendous contextual information with respect to the underlying document collection. For example, an acronym usually co-occurs with its full name and this context gives insight into a particular concept. As in reading - contextual information helps in understanding an article. Co-occurence analysis can be used to compute both symmetrical and asymmetrical co-occurred weights between pairs of terms. The choice of asymmetrical co-occurrence has an advantage to mnemonic human thinking process (Chen and Lynch, 1992).

2.2 **Dynamic Nature of User Information Need**

When searching, what do users want? Given the most common utilization of web search engines, the *obvious* answer is a list of documents (URLs). But is it a *real* answer? This section does not intend to find the *real* answer. Instead, it will look at the dynamic nature of user need in two aspects - expressing user need and perceiving knowledge. To a certain extent, using information retrieval systems is like communicating with another person. A user needs to express his or her need to a system and then the system will return *some information* to the user. The first of the following two sub-sections discusses user need. The second one examines how users deal with the information returned.

2.2.1 Expressing User Need

The goal of information retrieval is to return as many as possible relevant objects in response to a given information need. This simple goal has two types of complexity: relevance and information need. The question of relevance will be discussed in the section on *perceiving knowledge*.

First, we will discuss what information need is and how it is related to information retrieval systems. We will then look at three different phenomena involved in users' difficulties in expressing their needs. Finally, we will examine whether we can borrow some techniques from information providers to help users to express their information need.

2.2.1.1 Information Need

Information need is merely the reflection of what a searcher wants at a given time. Cooper (Cooper, 1971) describes information need as a searcher's psychological state which is not directly observable or symbolized. Bates (Bates, 1990) also argues that a mismatch between information need and what information systems currently provide is partly due to the static nature of information systems previously discussed, which is complicated by the dynamic nature of information need.

In attempts to overcome the mismatch and more effectively capture the information need, there has been much research in the area of human-computer interaction. In a system-oriented review of over 50 different search-interfaces, Vickery and Vickery (Vickery and Vickery, 1993) conclude that current interfaces may be over-elaborated or over-engineered by being so structured that the information displays and the ordering constraints for entering query are predetermined and unchangeable. They suggest that people might benefit from simple systems that allow a flexible and revealing dialogue between the system and user. In the research on information workspaces and visualization, Rao et al. (Rao et al., 1992) states that even thought searchers may want to interleave access operations and track their progress, current information systems are weak in their support for this process. Their findings simply show that no predefined interface to any information system can fit the diversity of information need.

Regarding difficulties in the development of knowledge based systems to aid the search process, Bates (Bates, 1990) argues that the goal should not be to replace the searcher with a knowledge based system, but rather to design the interface to support the strategic, opportunistic behavior of searchers.

In contrast to a predefined interface, Hendry and Harper (Hendry and Harper, 1997) create a loose and informal information-seeking environment in which searchers can *freely* express their information needs. Such an environment allows a searcher to associate query and result graphically and spatially in a work space. The ability to spatially arrange information, called *secondary notation* (Hendry and Harper, 1997), can help the searcher to comprehend the search process as well as to guide the execution of search activity. However, their findings show that lack of expertise in using such open information-seeking environment is a barrier to expressing a searcher's information need, indicating that user interface is not the sole factor in or solution to elicitate user's information need.

2.2.1.2 Indeterminism

Indeterminism is another phenomenon related to expressing information need (Chen et al., 1994a), (Blair, 1986), (Bates, 1986). Three factors may contribute to searchers' inability to express precisely what they want: system, searcher, and process. The system factor involves the variety in index terms assigned to documents. Lacking of indexing consistency can be further broken down into three areas: inter-indexer inconsistency, intra-indexer inconsistency, and inter-system inconsistency. The inter-indexer inconsistency comes from the observation that different indexers are likely to assign different index terms to the same document. The intra-indexer consistency derives from recognition that an indexer may use different index terms for the same document at different times (Blair, 1986), (Bates, 1986). The inter-system inconsistency comes from the fact that the collections of documents in many information retrieval systems overlap (Chen et al., 1994a). Furthermore, different systems use different indexing policies and controlled vocabularies (Barber et al., 1988).

The second factor for indeterminism arises from the variety in search terms used by searchers (Blair, 1986), (Bates, 1986). Like indexers, different searchers use different search terms to express the same information need, either because of unfamiliarity with the subject area or differences in experience and training.

The process factor is the subtlety and complexity of the search process (Bates, 1986), which is heavily dictated by the search mechanisms provided by different systems. Variations of search mechanism range from query term entering methods, choice of filters, complexity of boolean formulation, and user interface designs.

2.2.1.3 Opportunism

At the heart of search tactics and strategies, a search process involves planning, backtracking, and comparison (Bates, 1979a), (Bates, 1979b). In this connection, drawing from the research in programming environments, Visser (Visser, 1994) finds that support for basic cognitive tasks such as planning, backtracking, and reformulating is considered essential because it allows people to work the way they want to - opportunistically.

Carmel et al. (Carmel et al., 1992) performed a cognitive study which showed that opportunism appears in two out of three kinds of browsing strategies in a hypertext environment: review-browse and scan-browse. The review-browse strategy is to scan and review interesting information in the presence of transient browse goals that represent changing tasks. The scan-browse strategy is to scan for interesting information without review. Marchionini and Shneiderman (Marchionini and Shneiderman, 1988) also define browse as an exploratory, information seeking strategy appropriate for ill-defined problems and for exploring new task domains. As long as indeterminism in expressing information need exists, opportunism stemming from any hint will be a great help to users.

2.2.1.4 Vocabulary Problem

A searcher's experience is phenomenologically different from an indexer's experience (Bates, 1998), giving rise to potential choice of different vocabularies to describe the same object and creating a matching problem between terms provided by indexers and searchers. Even if well trained in an indexing scheme, different indexers might assign different index terms for a given document (Bates, 1986). In another study, Furnas et al. demonstrated that in a setting allowing spontaneous word choice for objects in five domains, two well-trained indexers favored the *same* term with less than 20 percent probability (Furnas et al., 1987). The probability might be even worse for searchers having different levels of domain expertise and system knowledge (Furnas et al., 1983), (Furnas et al., 1987), (Chen, 1994).

As for expressing a query, the dynamic of vocabulary differences also happens between searchers and systems. Vocabularies in systems are, of course, mainly determined by authors, indexers, and information providers, but, in addition, vocabularies in systems are static in terms of their accountability to documents. Every search process from the same or different searchers creates a different vocabulary matching situation to a system so unless searchers know the vocabulary used in the system, the variety of language presents a serious barrier to express query.

2.2.1.5 Recognition with Contextual Information

To alleviate the difficulties in expressing user need, techniques and heuristics can be borrowed from the manual process of generating thesauri. After using some automatic techniques to generate a set of potential terms from a defined domain area, the main task is to select a set of *useful* terms to be included in the new thesaurus (Kowalski, 1997). To aid the selection process, text concordances from documents that cover the domain area are used. A text concordance is an alphabetical listing of terms from a set of documents along with their frequency of occurrence and *references* to documents in which they are found (Salton, 1988). A text concordance is also known as *Key Word Out of Context* (KWOC) (Kowalski, 1997). Related to KWOC, *Key Word In Context* (KWIC) shows meaningful contextual information about a term by listing two fragments of terms before and after the term in the original sentence or document (Luhn, 1960), (Salton, 1988), (Kowalski, 1997), (Wynar, 1985). KWIC is useful in determining the meaning of homographs. For example, the term “chip” could be *wood chip* or *memory chip*. With KWIC, the editor of the thesaurus can read the sentence fragment associated with the term to understand its context and then determine its meaning. However, the KWIC tool is impractical in a large database where the listing of sentence fragments is very long. Nonetheless, contextual information may help a user develop a better understanding of a term’s meaning.

The technique of relevance feedback is similar to KWIC in delivering contextual information to user. The relevance feedback concept is that the new query should be based on an old query modified to increase the weight of terms in relevant documents and decrease the weight of terms that are in non-relevant documents (Rocchio, 1971), (Salton and Buckley, 1990). Under relevant feedback, contextual information is given by a set of retrieved documents instead of a set of sentence fragments. Users need to identify whether each document is relevant to what they want. The system will then use the information of relevance to modify the users’ query. In practice, users often skip the most important step - providing the relevance feedback information.

2.2.2 Perceiving Knowledge

How humans perceive knowledge has been studied by many disciplines such as Philosophy, Psychology, Linguistics, Cognitive Science, Behaviorial Science, etc.

Various competing and sometimes even complementary theories and models have arisen to explorer how a human mind works (O'Brien, 1998).

Indeed, the fundamental is a question about what knowledge is. The same goes for human perception. Our objective here is not to attempt to answer these two philosophical questions. Rather, we would like to use them to look at these issues in information retrieval, or search process in general. What is the user's perspective of knowledge? How does a user perceive retrieved or *derived* knowledge? In the field of information retrieval, many tools and techniques to assist humans to obtain information have been produced from theories and experiments. Some show promising results in experimental settings. However, in practice, they may have made a limited contribution, as witnessed by the observation that popular web search engines often have reverted to technologies developed in 1960s. What is the problem?

In the discussion of knowledge in intelligent systems, Minsky (Minsky, 1968) declares that heuristic programs are able to solve much harder problems than self-organizing systems because heuristic programs are given enough specific factual knowledge about particular problems. They do not have to start from an unstructured basis to evolve everything they will need, given that the requisite *knowledge* is suitably represented. How much of this discussion can be applied to information seekers? Since Minsky attempts to build intelligent machines which model human intelligence, a parallel discussion of analysis of users seems justified. Then, how can we represent strategic knowledge? And, how can we so present strategic knowledge to users to allow them to perceive retrieved information and possibly obtain more knowledge?

Meaning requires time-consuming thought, and the pace of modern life works against affording us time to think: the time to translate data into information and information into knowledge (Wurman, 1989). We have a limited capacity to transmit and process information, so we must distort the flow by being selective. The more there is to select from, in a given time frame, the more distortion must occur. Basically the technology of the production, storage, and transmission of information has outstripped the human strategies available to cope with this great volume of global data (Reeves, 1996). The ability to link relevant but fragmented information into some coherent whole allows for new knowledge. Having an adequate context or contextual knowledge base greatly enhances the possibility of assimilation of new information within pre-existing frames.

This literview review has been restricted to some important areas related to how humans perceive knowledge from information retrieval systems, in particular, how users comprehend knowledge embedded in systems, text documents, and search processes.

2.2.2.1 Computing Relevance?

If an information need can be expressed clearly, can *relevance* be computed? Searchers can consistently and easily determine the relevance of an information object with respect to their information need without being able to enunciate the criteria they use to do so (Blair, 1990). Cooper (Cooper, 1971) also defined the determination of relevance of an information object as a subjective process which reflects the psychological state of the information need. In order to have a better understanding of relevance, Cooper (Cooper, 1971) makes a distinction between so called *logical relevance* and *utility.* A given object is logically relevant to an information need if the object is topically related to the need. On the other hand, utility is purely a pragmatic notion: Is the object *useful* to the searcher? It is obvious that utility is solely determined by the searcher and cannot be computed by any information system. In an attempt to compare different information retrieval models, Bruza and Huibers (Bruza and Huibers, 1996) propose *aboutness*, which is level of topical relatedness between an object and a request computed by an information retrieval system. That is, the most that a system is able to do is to retrieve related but not relevant objects. The relevance only can be identified after the searcher examines the retrieved result. In the other words, the searcher needs to go through the perceiving process on the retrieved materials first and then evaluates their relevance.

Nonetheless, most information retrieval *sessions* stop working when they return a list of retrieved document objects. Such sessions happen in several situations. Observing the information retrieval in the web reveals that a common search model is *query-trial-and-error*, in which the only activity that the searcher can perform is to try giving different query terms to prompt a system to retrieve a small set of top ranked *related* objects. Searchers may not use some systems with sophisticated features such as relevant feedback, refinement, and knowledge base, because of time constraints or a non-intuitive user interface. In addition, the content and format of knowledge employed by the producers of knowledge bases are seldom of interest to the vast majority of searchers (Rouse et al., 1998). Obviously, such a system is along way from fulfilling the information need.

More than thirty years ago Minsky (Minsky, 1968) summarized the problem of defining relevance to users when dealing with large amount of information as follows:

> ... it is hard to find a knowledge-classifying system that works well for many different kinds of problems; it requires immense effort to build a plausible thesaurus that works even within one field. Furthermore, any particular retrieval structure is liable to entail commitments making it difficult to incorporate concepts that appear after the original structure is assembled (Minsky, 1968).

Since then, different research groups and information providers have spent tremendous resources to build domain specific thesauri. In addition to thesauri, different *original knowledge structures* have been created to give comprehensive coverage in different domains. However, the issue of having some *helpful retrieval structure* remains unsolved. Minsky states that:

> The problem-solving abilities of a highly intelligent person lies partly in his superior heuristics for managing his knowledge structure and partly in the structure itself; these are probably somewhat inseparable. In any case, there is no reason to suppose that you can be intelligent except through the use of an adquate, *particular* knowledge or model structure (Minsky, 1968).

Different knowledge structures have been created to aid searchers to *perceive* the relevance of retrieved objects. The following three sections describe them based on the existence (or non-existence) of structure and context of each method and thereby complete the second half of the retrieval journey.

2.2.2.2 Structureless and Contextless: Document List

List structure has long been recognized as a commonly comprehensible structure. The inverted index - the core technology in information retrieval - is a list structure that stores all documents having the same term index. Although the inverted index is an internal data structure in a system, most user interfaces directly present only a document list as a search result. Of course, a majority of these present only a partial list of documents due to the tremandous volume of information in different search engines. It is also true that searchers usually examine only a small portion of the retrieved document list.

From the perspective of helping searchers to perceive retrieved information, a document list provides hardly any structural or contextual aid. Even though documents are presented in some kind of ranked order based on the vector space model, searchers have to evaluate each document independently for its relevance. The list structure merely provides an ease-of-use access method. Besides, the listing does not carry any contextual information related to documents. The only contextual information comes from the query.

2.2.2.3 Structural but Contextless: Dynamic Clustering

Ad-hoc clustering or categorization is often used for converging vast information in a considerably shorter time. The Northern Light web search engine categories search results into some pre-defined and ad-hoc categories as a summary tool (Schwartz, 1998). The convergence process of electronic meeting also utilizes categorization techniques to define categories of concerns and discussion in such meeting (Chen et al., 1994b). These additional analyzed results are usually presented prior to or alongside document lists.

From the searchers' point of view, dynamic clustering is an extra knowledge structure superimposed on the underlying retrieval materials. The searchers benefit from the structure to direct themselves to certain *interesting* categories and then examine documents in those areas. However, the contextual information comes after the searchers examine the dynamic clustering structure. That is, the searchers can expect only a clustering structure but without any hint of what the content would be.

2.2.2.4 Sturctural and Contextual: Path to the Knowledge

According to the mental-logic model on basic inference schemas (Braine and O'Brien, 1998), Lea et al. in the field of psychology performed a series of studies on how humans draw inferences from text comprehension (Lea et al., 1990). One of the studies was on a *Recognition Task* in which subjects were asked to indicate whether the information contained in three newly presented sentences had been presented *explicitly* in the text (story) they had just read, or whether they had to infer that information.

Here is an example extracted from Lea's work (Lea, 1998). The story is:

1. The Borofskys were planning a dinner party.
2. "Alice and Sarah are vegetarians," Mrs. Borofsky said, "so if we invite either one of them, we can't serve meat."
3. "Well, if we invite Harry, we have to invite Alice," Mr. Borofsky said.
4. "And if we invite George, we have to invite Sarah."
5. "We already decided to invite either Harry or George, or maybe both of them," said Mrs. Borosky.

Among the three sentences for the *Recognition Task*, the following one is most interest to the discussion of users' inferencing process with respect to information retrieval:

- In their discussion, the Borofskys concluded that they had to invite Alice or Sarah (or both).

Characteristic of this sentence is that it contains a logical inference that the mental-logic model predicts readers would make while reading the story. Lea et al.'s results indicated that sixty-nine percent of the subjects thought that the inferences had been presented explicitly in the text (story), a result that provides strong evidence that people often do not realize they are making inferences in text comprehension.

In addition, in many discourse situations a distant premise is simply a fact stored in long-term memory, not information that was presented earlier in the text or conversation. The following example illustrates how distant premise information can be retrieved from the participants' world knowledge, as well as from their memory of earlier parts of the discourse (Lea, 1998):

- Bob was asking Barb about her new personal computer.
- "What did you decide about which type of computer to get?"
- "Well," said Barb foolishly, "in the end I decided not to get the Mac because they'll probably be out of business before I fill out the warranty card."
- *Inference:* Barb got an IBM-compatible machine.

In this example, the distant premise is world knowledge stored in both Barb's and Bob's long-term memory: Personal computers are either Mac or IBM-compatible machines. This is the type of distant premise used for making inferences in text comprehension.

In the information retrieval process, a parallel analysis indicates that the distant premise can be captured by a system, even though the knowledge is not complete. Users choose distant premises, both from the computer system and their own knowledge, in order to make inferences to help them comprehend the search space, and possibly the search result. From a philosophical point of view, Lipton (Lipton, 1991) argues that inferences lead to the best possible explanation. Inferences may occur recursively until a comprehensible premise is reached and a partially defined search space can be extended to a comprehensible one through inference. In the other words, the premises coming from inference help shape the scope of the search space. In addition, distant premises give clues that help users comprehend and analyze the underlying search space, which is the document space with respect to the information retrieval.

From the psychological point of view, the activities of knowledge revision can give rise to aggregates for concept formation (Wrobel, 1994). Wrobel (Wrobel, 1994) describes knowledge revision as a necessary activity for knowledge acquisition due to several factors like changing world and application domains, sloppy modeling at the initial stage of knowledge acquisition, and selection bias and incremental learning.

From a user study of inexperienced searchers' learning to use online search engines, a positive effect was found to be the direct result of increased searching success as greater knowledge of vocabulary and search strategy was acquired (Meghabghab, 1995). Knowledge of a vocabulary gives context to the search process while search strategy gives structure to the process.