The postscript version of this chapter.

Table of Contents.

Chapter 2

Information Retrieval Models

2.1 Introduction

The purpose of this chapter is two-fold: First, we want to set the stage for the problems in information retrieval that we try to address in this thesis. Second, we want to give the reader a quick overview of the major textual retrieval methods, because the InfoCrystal can help to visualize the output from any of them. We begin by providing a general model of the information retrieval process. We then briefly describe the major retrieval methods and characterize them in terms of their strengths and shortcomings.

2.2 General Model of Information Retrieval

The goal of **information retrieval** (IR) is to provide users with those documents that will satisfy their information need. We use the word "document" as a general term that could also include non-textual information, such as multimedia objects. Figure 4.1 provides a general overview of the information retrieval process, which has been adapted from Lancaster and Warner (1993). Users have to formulate their information need in a form that can be understood by the retrieval mechanism. There are several steps involved in this translation process that we will briefly discuss below. Likewise, the contents of large document collections need to be described in a form that allows the retrieval mechanism to identify the potentially relevant documents quickly. In both cases, information may be lost in the transformation process leading to a computer-usable representation. Hence, the matching process is inherently imperfect.

Information seeking is a form of problem solving [Marcus 1994, Marchionini 1992]. It proceeds according to the interaction among eight subprocesses: problem recognition and acceptance, problem definition, search system selection, query formulation, query execution, examination of results (including relevance feedback), information extraction, and reflection/iteration/termination. To be able to perform effective searches, users have to develop the following expertise: knowledge about various sources of information, skills in defining search problems and applying search strategies, and competence in using electronic search tools.

Marchionini (1992) contends that some sort of spreadsheet is needed that supports users in the problem definition as well as other information seeking tasks. The InfoCrystal is such a spreadsheet because it assists users in the formulation of their information needs and the exploration of the retrieved documents, using the a visual interface that supports a "what-if" functionality. He further predicts that advances in computing power and speed, together with improved information retrieval procedures, will continue to blur the distinctions between problem articulation and examination of results. The InfoCrystal is both a visual query language and a tool for visualizing retrieval results.

The information need can be understood as forming a pyramid, where only its peak is made visible by users in the form of a conceptual query (see Figure 2.1). The conceptual query captures the key concepts and the relationships among them. It is the result of a conceptual analysis that operates on the information need, which may be well or vaguely defined in the user's mind. This analysis can be challenging, because users are faced with the general "vocabulary problem" as they are trying to translate their information need into a conceptual query. This problem refers to the fact that a single word can have more than one meaning, and, conversely, the same concept can be described by surprisingly many different words. Furnas, Landauer, Gomez and Dumais (1983) have shown that two people use the same main word to describe an object only 10 to 20% of the time. Further, the concepts used to represent the documents can be different from the concepts used by the user. The conceptual query can take the form of a natural language statement, a list of concepts that can have degrees of importance assigned to them, or it can be statement that coordinates the concepts using Boolean operators. Finally, the conceptual query has to be translated into a query surrogate that can be understood by the retrieval system.

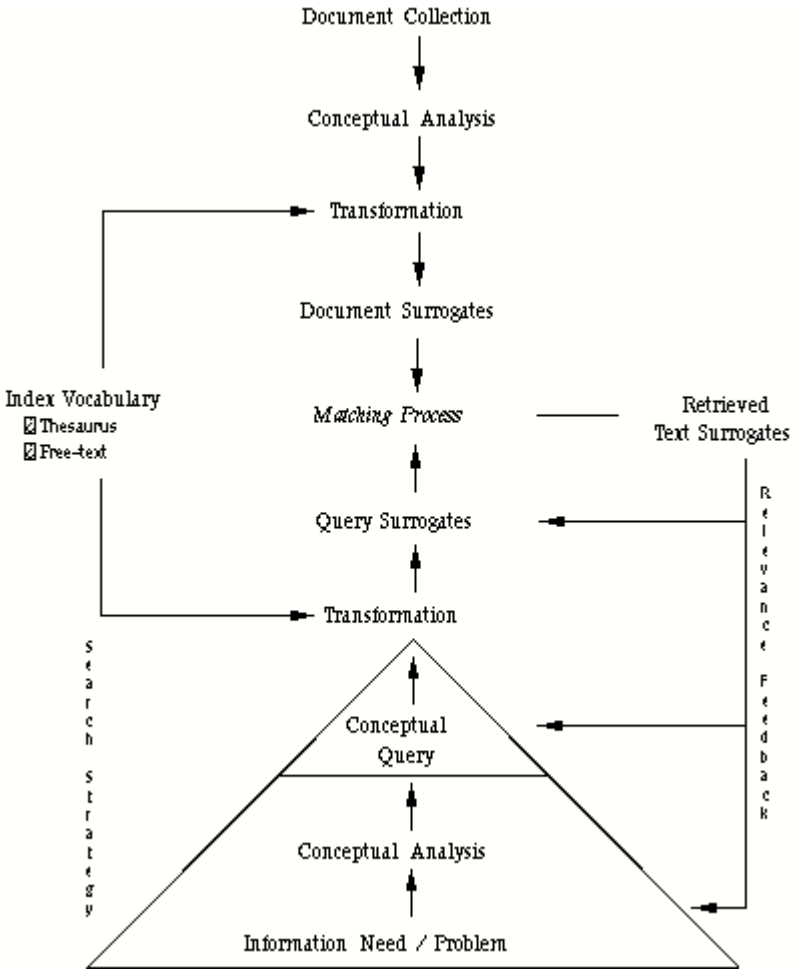

Figure 2.1: represents a general model of the information retrieval process, where both the user's information need and the document collection have to be translated into the form of surrogates to enable the matching process to be performed. This figure has been adapted from Lancaster and Warner (1993).

Similarly, the meanings of documents need to be represented in the form of text surrogates that can be processed by computer. A typical surrogate can consist of a set of index terms or descriptors. The text surrogate can consist of multiple fields, such as the title, abstract, descriptor fields to capture the meaning of a document at different levels of resolution or focusing on different characteristic aspects of a document. Once the specified query has been executed by IR system, a user is presented with the retrieved document surrogates. Either the user is satisfied by the retrieved information or he will evaluate the retrieved documents and modify the query to initiate a further search. The process of query modification based on user evaluation of the retrieved documents is known as relevance feedback [Lancaster and Warner 1993]. Information retrieval is an inherently interactive process, and the users can change direction by modifying the query surrogate, the conceptual query or their understanding of their information need.

It is worth noting here the results, which have been obtained in studies investigating the information-seeking process, that describe information retrieval in terms of the cognitive and affective symptoms commonly experienced by a library user. The findings by Kuhlthau et al. (1990) indicate that thoughts about the information need become clearer and more focused as users move through the search process. Similarly, uncertainty, confusion, and frustration are nearly universal experiences in the early stages of the search process, and they decrease as the search process progresses and feelings of being confident, satisfied, sure and relieved increase. The studies also indicate that cognitive attributes may affect the search process. User's expectations of the information system and the search process may influence the way they approach searching and therefore affect the intellectual access to information.

Analytical search strategies require the formulation of specific, well-structured queries and a systematic, iterative search for information, whereas browsing involves the generation of broad query terms and a scanning of much larger sets of information in a relatively unstructured fashion. Campagnoni et al. (1989) have found in information retrieval studies in hypertext systems that the predominant search strategy is "browsing" rather than "analytical search". Many users, especially novices, are unwilling or unable to precisely formulate their search objectives, and browsing places less cognitive load on them. Furthermore, their research showed that search strategy is only one dimension of effective information retrieval; individual differences in visual skill appear to play an equally important role.

These two studies argue for information displays that provide a spatial overview of the data elements and that simultaneously provide rich visual cues about the content of the individual data elements. Such a representation is less likely to increase the anxiety that is a natural part of the early stages of the search process and it caters for a browsing interaction style, which is appropriate especially in the beginning, when many users are unable to precisely formulate their search objectives.

2.3 Major Information Retrieval Models

The following major models have been developed to retrieve information: the **Boolean** model, the **Statistical** model, which includes the vector space and the probabilistic retrieval model, and the **Linguistic and Knowledge-based** models. The first model is often referred to as the "exact match" model; the latter ones as the "best match" models [Belkin and Croft 1992]. The material presented here is based on the textbooks by Lancaster and Warner (1992) as well as Frakes and Baeza-Yates (1992), the review article by Belkin and Croft (1992), and discussions with Richard Marcus, my thesis advisor and mentor in the field of information retrieval.

Queries generally are less than perfect in two respects: First, they retrieve some irrelevant documents. Second, they do not retrieve all the relevant documents. The following two measures are usually used to evaluate the effectiveness of a retrieval method. The first one, called the *precision rate*, is equal to the proportion of the retrieved documents that are actually relevant. The second one, called the *recall rate*, is equal to the proportion of all relevant documents that are actually retrieved. If searchers want to raise precision, then they have to narrow their queries. If searchers want to raise recall, then they broaden their query. In general, there is an inverse relationship between precision and recall. Users need help to become knowledgeable in how to manage the precision and recall trade-off for their particular information need [Marcus 1991].

2.3.1.1 Standard Boolean

In Table 2.1 we summarize the defining characteristics of the standard Boolean approach and list its key advantages and disadvantages. It has the following strengths: 1) It is easy to implement and it is computationally efficient [Frakes and Baeza-Yates 1992]. Hence, it is the standard model for the current large-scale, operational retrieval systems and many of the major on-line information services use it. 2) It enables users to express structural and conceptual constraints to describe important linguistic features [Marcus 1991]. Users find that synonym specifications (reflected by OR-clauses) and phrases (represented by proximity relations) are useful in the formulation of queries [Cooper 1988, Marcus 1991]. 3) The Boolean approach possesses a great expressive power and clarity. Boolean retrieval is very effective if a query requires an exhaustive and unambiguous selection. 4) The Boolean method offers a multitude of techniques to broaden or narrow a query. 5) The Boolean approach can be especially effective in the later stages of the search process, because of the clarity and exactness with which relationships between concepts can be represented.

The standard Boolean approach has the following shortcomings: 1) Users find it difficult to construct effective Boolean queries for several reasons [Cooper 1988, Fox and Koll 1988, Belkin and Croft 1992]. Users are using the natural language terms AND, OR or NOT that have a different meaning when used in a query. Thus, users will make errors when they form a Boolean query, because they resort to their knowledge of English.

	Standard Boolean
Goal	• Capture conceptual structure and contextual information
Methods	• Coordination: AND, OR, NOT • Proximity • Fields • Stemming / Truncation
(+)	• Easy to implement • Computationally efficient => all the major on-line databases use it • Expressiveness and Clarity Synonym specifications (OR-clauses) and phrases (AND-clauses).
(-)	• Difficult to construct Boolean queries. • All or nothing AND too severe, and OR does not differentiate enough. • Difficult to control output: Null output <-> Overload. • No ranking • No weighting of index or query terms • No uncertainty measure

Table 2.1: summarizes the defining characteristics of the standard Boolean approach and list the its key advantages and disadvantages.

For example, in ordinary conversation a noun phrase of the form "A and B" usually refers to more entities than would "A" alone, whereas when used in the context of information retrieval it refers to fewer documents than would be retrieved by "A" alone. Hence, one of the common mistakes made by users is to substitute the AND logical operator for the OR logical operator when translating an English sentence to a Boolean query. Furthermore, to form complex queries, users must be familiar with the rules of precedence and the use of parentheses. Novice users have difficulty using parentheses, especially nested parentheses. Finally, users are overwhelmed by the multitude of ways a query can be structured or modified, because of the combinatorial explosion of feasible queries as the number of concepts increases. In particular, users have difficulty identifying and applying the different strategies that are available for narrowing or broadening a Boolean query [Marcus 1991, Lancaster and Warner 1993]. 2) Only documents that satisfy a query exactly are retrieved. On the one hand, the AND operator is too severe because it does not distinguish between the case when none of the concepts are satisfied and the case where all except one are satisfied. Hence, no or very few documents are retrieved when more than three and four criteria are combined with the Boolean operator AND (referred to as the Null Output problem). On the other hand, the OR operator does not reflect how many concepts have been satisfied. Hence, often too many documents are retrieved (the Output Overload problem). 3) It is difficult to control the number of retrieved documents. Users are often faced with the null-output or the information overload problem and they are at loss of how to modify the query to retrieve the reasonable number documents. 4) The traditional Boolean approach does not provide a relevance ranking of the retrieved documents, although modern Boolean approaches can make use of the degree of coordination, field level and degree of stemming present to rank them [Marcus 1991]. 5) It does not represent the degree of uncertainty or error due the vocabulary problem [Belkin and Croft 1992].

2.3.1.2 Narrowing and Broadening Techniques

As mentioned earlier, a Boolean query can be described in terms of the following four operations: degree and type of coordination, proximity constraints, field specifications and degree of stemming as expressed in terms of word/string specifications. If users want to (re)formulate a Boolean query then they need to make informed choices along these four dimensions to create a query that is sufficiently broad or narrow depending on their information needs. Most narrowing techniques lower recall as well as raise precision, and most broadening techniques lower precision as well as raise recall. Any query can be reformulated to achieve the desired precision or recall characteristics, but generally it is difficult to achieve both. Each of the four kinds of operations in the query formulation has particular operators, some of which tend to have a narrowing or broadening effect. For each operator with a narrowing effect, there is one or more inverse operators with a broadening effect [Marcus 1991]. Hence, users require help to gain an understanding of how changes along these four dimensions will affect the broadness or narrowness of a query.

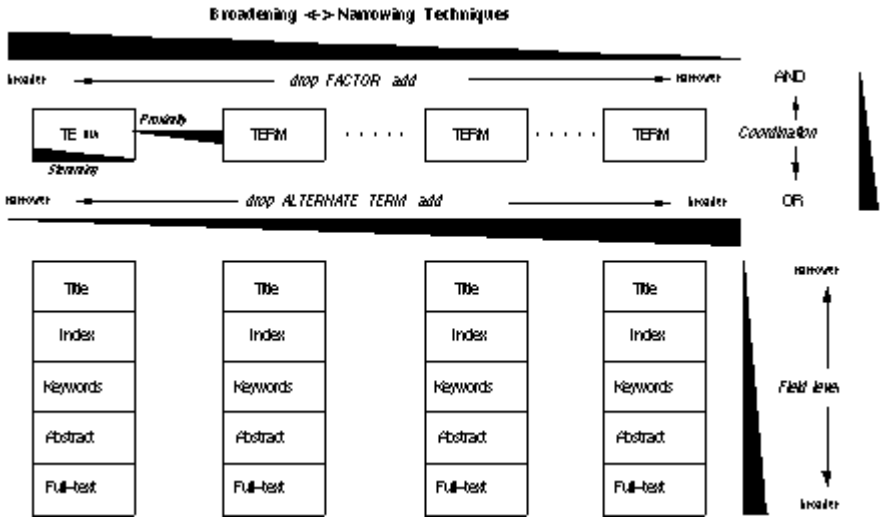

Figure 2.2: captures how coordination, proximity, field level and stemming affect the broadness or narrowness of a Boolean query. By moving in the direction in which the wedges are expanding the query is broadened.

Figure 2.2 shows how the four dimensions affect the broadness or narrowness of a query: 1) *Coordination*: the different Boolean operators AND, OR and NOT have the following effects when used to add a further concept to a query: a) the AND operator narrows a query; b) the OR broadens it; c) the effect of the NOT depends on whether it is combined with an AND or OR operator. Typically, in searching textual databases, the NOT is connected to the AND, in which case it has a narrowing effect like the AND operator. 2) *Proximity*: The closer together two terms have to appear in a document, the more narrow and precise the query. The most stringent proximity constraint requires the two terms to be adjacent. 3) *Field level*: current document records have fields associated with them, such as the "Title", "Index", "Abstract" or "Full-text" field: a) the more fields that are searched, the broader the query; b) the individual fields have varying degrees of precision associated with them, where the "title" field is the most specific and the "full-text" field is the most general. 4) *Stemming*: The shorter the prefix that is used in truncation-based searching, the broader the query. By reducing a term to its morphological stem and using it as a prefix, users can retrieve many terms that are conceptually related to the original term [Marcus 1991].

Using Figure 2.2, we can easily read off how to broaden query. We just need to move in the direction in which the wedges are expanding: we use the OR operator (rather than the AND), impose no proximity constraints, search over all fields and apply a great deal of stemming. Similarly, we can formulate a very narrow query by moving in the direction in which the wedges are contracting: we use the AND operator (rather than the OR), impose proximity constraints, restrict the search to the title field and perform exact rather than truncated word matches. In Chapter 4 we will show how Figure 2.2 indicates how the broadness or narrowness of a Boolean query could be visualized.

2.3.1.3 Smart Boolean

There have been attempts to help users overcome some of the disadvantages of the traditional Boolean discussed above. We will now describe such a method, called *Smart Boolean*, developed by Marcus [1991, 1994] that tries to help users construct and modify a Boolean query as well as make better choices along the four dimensions that characterize a Boolean query. We are not attempting to provide an in-depth description of the Smart Boolean method, but to use it as a good example that illustrates some of the possible ways to make Boolean retrieval more user-friendly and effective. Table 2.2 provides a summary of the key features of the Smart Boolean approach.

Users start by specifying a natural language statement that is automatically translated into a Boolean Topic representation that consists of a list of factors or concepts, which are automatically coordinated using the AND operator. If the user at the initial stage can or wants to include synonyms, then they are coordinated using the OR operator. Hence, the Boolean Topic representation connects the different factors using the AND operator, where the factors can consist of single terms or several synonyms connected by the OR operator. One of the goals of the Smart Boolean approach is to make use of the structural knowledge contained in the text surrogates, where the different fields represent contexts of useful information. Further, the Smart Boolean approach wants to use the fact that related concepts can share a common stem. For example, the concepts "computers" and "computing" have the common stem comput*.

	Smart Boolean
Goal	• Structure search (re-)formulation process. • Use structural and contextual knowledge-bases and clarity of Boolean expressions.
Methods	• Natural language statement is automatically translated into Boolean Topic Representation • Boolean Topic Representation: ANDs of ORs of concepts — Keyword/stem, all fields • Conceptual info. -> Coordination and Add/Drop Factor • Contextual info. -> Proximity • Structural info. -> Field levels • Synonym or word relationships -> Stemming/Truncation overlap => all this information can be used to rank documents • Techniques to Broaden and Narrow query
(+)	• No need for Boolean operators => Convert operator-free statement into ANDs of ORs • Assist user in query (re)formulation: by asking users targeted questions to automatically modify the query. • "Why irrelevant?" -> activates narrowing methods. • "Broaden by Dropping Factors" to estimate recall.
(-)	• How to visualize ? • Conceptual query representation (BTR) • Query modification techniques and their effects • Structured relevance feedback

Table 2.2: summarizes the defining characteristics of the Smart Boolean approach and list the its key advantages and disadvantages.

The initial strategy of the Smart Boolean approach is to start out with the broadest possible query within the constraints of how the factors and their synonyms have been coordinated. Hence, it modifies the Boolean Topic representation into the query surrogate by using only the stems of the concepts and searches for them over all the fields. Once the query surrogate has been performed, users are guided in the process of evaluating the retrieved document surrogates. They choose from a list of reasons to indicate why they consider certain documents as relevant. Similarly, they can indicate why other documents are not relevant by interacting with a list of possible reasons. This user feedback is used by the Smart Boolean system to automatically modify the Boolean Topic representation or the query surrogate, whatever is more appropriate. The Smart Boolean approach offers a rich set of strategies for modifying a query based on the received relevance feedback or the expressed need to narrow or broaden the query. The Smart Boolean retrieval paradigm has been implemented in the form of a system called CONIT, which is one of the earliest expert retrieval systems that was able to demonstrate that ordinary users, assisted by such a system, could perform equally well as experienced search intermediaries [Marcus 1983]. However, users have to navigate through a series of menus listing different choices, where it might be hard for them to appreciate the implications of some of these choices. A key limitation of the previous versions of the CONIT system has been that lacked a visual interface. The most recent version has a graphical interface and it uses the tiling metaphor suggested by Anick et al. (1991), and discussed in section 10.4, to visualize Boolean coordination [Marcus 1994]. This visualization approach suffers from the limitation that it enables users to visualize specific queries, whereas we will propose a visual interface that represents all whole range of related Boolean queries in a single display, making changes in Boolean coordination more user-friendly. Further, the different strategies of modifying a query in CONIT require a better visualization metaphor to enable users to make use these search heuristics. In Chapter 4 we show how some of these modification techniques can be visualized.

2.3.1.4 Extended Boolean Models

Several methods have been developed to extend the Boolean model to address the following issues: 1) The Boolean operators are too strict and ways need to be found to soften them. 2) The standard Boolean approach has no provision for ranking. The Smart Boolean approach and the methods described in this section provide users with relevance ranking [Fox and Koll 1988, Marcus 1991]. 3) The Boolean model does not support the assignment of weights to the query or document terms. We will briefly discuss the *P-norm* and the *Fuzzy Logic* approaches that extend the Boolean model to address the above issues.

	Extended Boolean Models
Goal	• Less strict Boolean operators • Ranked output
Methods	
	• **Fuzzy logic** [OR -> max], [AND -> min] and [NOT -> 1 - max] (-) Lack of sensitivity of min and max: min(0.2, 0.8) = min(0.2, 0.3).

Table 2.3: summarizes the defining characteristics of the Extended Boolean approach and list the its key advantages and disadvantages.

The **P-norm** method developed by Fox (1983) allows query and document terms to have weights, which have been computed by using term frequency statistics with the proper normalization procedures. These normalized weights can be used to rank the documents in the order of decreasing distance from the point (0, 0, ... , 0) for an

OR query, and in order of increasing distance from the point (1, 1, ... , 1) for an AND query. Further, the Boolean operators have a coefficient P associated with them to indicate the degree of strictness of the operator (from 1 for least strict to infinity for most strict, i.e., the Boolean case). The P-norm uses a distance-based measure and the coefficient P determines the degree of exponentiation to be used. The exponentiation is an expensive computation, especially for P-values greater than one.

In **Fuzzy Set theory**, an element has a varying degree of membership to a set instead of the traditional binary membership choice. The weight of an index term for a given document reflects the degree to which this term describes the content of a document. Hence, this weight reflects the degree of membership of the document in the fuzzy set associated with the term in question. The degree of membership for union and intersection of two fuzzy sets is equal to the maximum and minimum, respectively, of the degrees of membership of the elements of the two sets. In the "Mixed Min and Max" model developed by Fox and Sharat (1986) the Boolean operators are softened by considering the query-document similarity to be a linear combination of the min and max weights of the documents.

2.3.2 Statistical Model

The *vector space* and *probabilistic* models are the two major examples of the statistical retrieval approach. Both models use statistical information in the form of term frequencies to determine the relevance of documents with respect to a query. Although they differ in the way they use the term frequencies, both produce as their output a list of documents ranked by their estimated relevance. The statistical retrieval models address some of the problems of Boolean retrieval methods, but they have disadvantages of their own. Table 2.4 provides summary of the key features of the vector space and probabilistic approaches. We will also describe *Latent Semantic Indexing* and *clustering* approaches that are based on statistical retrieval approaches, but their objective is to respond to what the user's query did not say, could not say, but somehow made manifest [Furnas et al. 1983, Cutting et al. 1991].

2.3.2.1 Vector Space Model

The **vector space model** represents the documents and queries as vectors in a multidimensional space, whose dimensions are the terms used to build an index to represent the documents [Salton 1983]. The creation of an index involves lexical scanning to identify the significant terms, where morphological analysis reduces different word forms to common "stems", and the occurrence of those stems is computed. Query and document surrogates are compared by comparing their vectors, using, for example, the cosine similarity measure. In this model, the terms of a query surrogate can be weighted to take into account their importance, and they are computed by using the statistical distributions of the terms in the collection and in the documents [Salton 1983]. The vector space model can assign a high ranking score to a document that contains only a few of the query terms if these terms occur infrequently in the collection but frequently in the document. The vector space model makes the following assumptions: 1) The more similar a document vector is to a query vector, the more likely it is that the document is relevant to that query. 2) The words used to define the dimensions of the space are orthogonal or independent. While it is a reasonable first approximation, the assumption that words are pairwise independent is not realistic.

2.3.2.2 Probabilistic Model

The **probabilistic retrieval** model is based on the Probability Ranking Principle, which states that an information retrieval system is supposed to rank the documents based on their probability of relevance to the query, given all the evidence available [Belkin and Croft 1992]. The principle takes into account that there is uncertainty in the representation of the information need and the documents. There can be a variety of sources of evidence that are used by the probabilistic retrieval methods, and the most common one is the statistical distribution of the terms in both the relevant and non-relevant documents.

We will now describe the state-of-art system developed by Turtle and Croft (1991) that uses Bayesian inference networks to rank documents by using multiple sources of evidence to compute the conditional probability P(Info need|document) that an information need is satisfied by a given document. An inference network consists

of a directed acyclic dependency graph, where edges represent conditional dependency or causal relations between propositions represented by the nodes. The inference network consists of a document network, a concept representation network that represents indexing vocabulary, and a query network representing the information need. The concept representation network is the interface between documents and queries. To compute the rank of a document, the inference network is instantiated and the resulting probabilities are propagated through the network to derive a probability associated with the node representing the information need. These probabilities are used to rank documents.

The statistical approaches have the following strengths: 1) They provide users with a relevance ranking of the retrieved documents. Hence, they enable users to control the output by setting a relevance threshold or by specifying a certain number of documents to display. 2) Queries can be easier to formulate because users do not have to learn a query language and can use natural language. 3) The uncertainty inherent in the choice of query concepts can be represented. However, the statistical approaches have the following shortcomings: 1) They have a limited expressive power. For example, the NOT operation can not be represented because only positive weights are used. It can be proven that only 2N*N of the 22N possible Boolean queries can be generated by the statistical approaches that use weighted linear sums to rank the documents. This result follows from the analysis of Linear Threshold Networks or Boolean Perceptrons [Anthony and Biggs 1992]. For example, the very common and important Boolean query ((A and B) or (C and D)) can not be represented by a vector space query (see section 5.4 for a proof). Hence, the statistical approaches do not have the expressive power of the Boolean approach. 3) The statistical approach lacks the structure to express important linguistic features such as phrases. Proximity constraints are also difficult to express, a feature that is of great use for experienced searchers. 4) The computation of the relevance scores can be computationally expensive. 5) A ranked linear list provides users with a limited view of the information space and it does not directly suggest how to modify a query if the need arises [Spoerri 1993, Hearst 1994]. 6) The queries have to contain a large number of words to improve the retrieval performance. As is the case for the Boolean approach, users are faced with the problem of having to choose the appropriate words that are also used in the relevant documents.

Table 2.4 summarizes the advantages and disadvantages that are specific to the vector space and probabilistic model, respectively. This table also shows the formulas that are commonly used to compute the term weights. The two central quantities used are the inverse term frequency in a collection (*idf*), and the frequencies of a term i in a document j (*freq(i,j)*). In the probabilistic model, the weight computation also considers how often a term appears in the relevant and irrelevant documents, but this presupposes that the relevant documents are known or that these frequencies can be reliably estimated.

Statistical	**Vector Space**	**Probabilistic**
Motivation	Simplify query formulation Ability to control output	Address uncertainty in query representations
Goal	Rank the output based on	
	Similarity	Probability of Relevance
Methods	Cosine measure	Use of different models

Source	Query Term Statistics
	Vector-Space: • similarity(Q,D) = Σ (w_{iq} x w_{ij}) / "normalizer" where w_{iq} = (0.5 + 0.5 $freq_{iq}$ / $maxfreq_q$) x **idf**(i) w_{ij} = $freq_{ij}$ x **idf**(i) • inverse term freq. in collection **idf**(i) = $\log_2$ (N–n(i)) / n(i). Probabilistic: • term weight = log [(r_t / R–r_t) / ((n_t –r_t) / ((N–n_t) – (R–r_t)))] ="(hits / misses) / (false alarms/correct misses)" • similarity $_{jk}$ = Σ (C + **idf**(i)) x **tf**(i,j) where **tf**(i,j) = K + (1–K) (freq(i,j) / maxfreq(j)).
Issues	• How to express NOT ? Proximity searches ? • Limited expressive power • Computationally intensive • Assumes that terms are independent. • Estimation of needed probabilities • Prior knowledge needed. • Independence assumption • Boolean relations lost. • Which model is best ? • Lack of structure to represent important linguistic features • How to better visualize the retrieved set ?

Table 2.4: summarizes the defining characteristics of the statistical retrieval approach, which includes the vector space and the probabilistic model and we list the their key advantages and disadvantages.

If users provide the retrieval system with relevance feedback, then this information is used by the statistical approaches to recompute the weights as follows: the weights of the query terms in the relevant documents are increased, whereas the weights of the query terms that do not appear in the relevant documents are decreased [Salton and Buckley 1990]. There are multiple ways of computing and updating the weights, where each has its advantages and disadvantages. We do not discuss these formulas in more detail, because research on relevance feedback has shown that significant effectiveness improvements can be gained by using quite simple feedback techniques [Salton and Buckley 1990]. Furthermore, what is important to this thesis is that the statistical retrieval approach generates a ranked list, however how this ranking has been computed in detail is immaterial for the purpose of this thesis.

2.3.2.3 Latent Semantic Indexing

Several statistical and AI techniques have been used in association with domain semantics to extend the vector space model to help overcome some of the retrieval problems described above, such as the "dependence problem" or the "vocabulary problem". One such method is **Latent Semantic Indexing** (LSI). In LSI the associations among terms and documents are calculated and exploited in the retrieval process. The assumption is that there is some "latent" structure in the pattern of word usage across documents and that statistical techniques can be used to estimate this latent structure. An advantage of this approach is that queries can retrieve documents even if they have no words in common. The LSI technique captures deeper associative structure than simple term-to-term correlations and is completely automatic. The only difference between LSI and vector space methods is that LSI represents terms and documents in a reduced dimensional space of the derived indexing dimensions. As with the vector space method, differential term weighting and relevance feedback can improve LSI performance substantially.

Foltz and Dumais (1992) compared four retrieval methods that are based on the vector-space model. The four methods were the result of crossing two factors, the first factor being whether the retrieval method used Latent Semantic Indexing or keyword matching, and the second factor being whether the profile was based on words or phrases provided by the user (Word profile), or documents that the user had previously rated as relevant (Document profile). The LSI match-document profile method proved to be the most successful of the four methods. This method combines the advantages of both LSI and the document profile. The document profile provides a simple, but effective, representation of the user's interests. Indicating just a few documents that are of interest is as effective as generating a long list of words and phrases that describe one's interest. Document profiles have an added advantage over word profiles: users can just indicate documents they find relevant without having to generate a description of their interests.

2.3.3 Linguistic and Knowledge-based Approaches

In the simplest form of automatic text retrieval, users enter a string of keywords that are used to search the inverted indexes of the document keywords. This approach retrieves documents based solely on the presence or absence of exact single word strings as specified by the logical representation of the query. Clearly this approach will miss many relevant documents because it does not capture the complete or deep meaning of the user's query. The Smart Boolean approach and the statistical retrieval approaches, each in their specific way, try to address this problem (see Table 2.5). Linguistic and knowledge-based approaches have also been developed to address this problem by performing a morphological, syntactic and semantic analysis to retrieve documents more effectively [Lancaster and Warner 1993]. In a morphological analysis, roots and affixes are analyzed to determine the part of speech (noun, verb, adjective etc.) of the words. Next complete phrases have to be parsed using some form of syntactic analysis. Finally, the linguistic methods have to resolve word ambiguities and/or generate relevant synonyms or quasi-synonyms based on the semantic relationships between words. The development of a sophisticated linguistic retrieval system is difficult and it requires complex knowledge bases of semantic information and retrieval heuristics. Hence these systems often require techniques that are commonly referred to as artificial intelligence or expert systems techniques.

2.3.3.1 DR-LINK Retrieval System

We will now describe in some detail the DR-LINK system developed by Liddy et al., because it represents an exemplary linguistic retrieval system. DR-LINK is based on the principle that retrieval should take place at the conceptual level and not at the word level. Liddy et al. attempt to retrieve documents on the basis of what people mean in their query and not just what they say in their query. DR-LINK system employs sophisticated, linguistic text processing techniques to capture the conceptual information in documents. Liddy et al. have developed a modular system that represents and matches text at the lexical, syntactic, semantic, and the discourse levels of language. Some of the modules that have been incorporated are: The Text Structurer is based on discourse linguistic theory that suggests that texts of a particular type have a predictable structure which serves as an indication where certain information can be found. The Subject Field Coder uses an established semantic coding scheme from a machine-readable dictionary to tag each word with its disambiguated subject code (e.g., computer science, economics) and to then produce a fixed-length, subject-based vector representation of the document and the query. The Proper Noun Interpreter uses a variety of processing heuristics and knowledge bases to produce: a canonical representation of each proper noun; a classification of each proper noun into thirty-seven categories; and an expansion of group nouns into their constituent proper noun members. The Complex Nominal Phraser provides means for precise matching of complex semantic constructs when expressed as either adjacent nouns or a non-predicating adjective and noun pair. Finally, The Natural Language Query Constructor takes as input a natural language query and produces a formal query that reflects the appropriate logical combination of text structure, proper noun, and complex nominal requirements of the user's information need. This module interprets a query into pattern-action rules that translate each sentence into a first-order logic assertion, reflecting the Boolean-like requirements of queries.

Linguistic Level	Boolean Retrieval	Statistical	Linguistic and Knowledge-based
Lexical	Stop word list	Stop word list	Lexicon
Morphological	Truncation symbol	Stemming	Morphological analysis
Syntactic	Proximity operators	Statistical phrases	Grammatical phrases
Semantic	Thesaurus	Clusters of co-occurring words	Network of words/phrases in semantic relationships

Table 2.5: characterizes the major retrieval methods in terms of how deal with lexical, morphological, syntactic and semantic issues.

To summarize, the DR-LINK retrieval system represents content at the conceptual level rather than at the word level to reflect the multiple levels of human language comprehension. The text representation combines the lexical, syntactic, semantic, and discourse levels of understanding to predict the relevance of a document. DR-LINK accepts natural language statements, which it translates into a precise Boolean representation of the user's relevance requirements. It also produces a summary-level, semantic vector representations of queries and documents to provide a ranking of the documents.

2.4 Conclusion

There is a growing discrepancy between the retrieval approach used by existing commercial retrieval systems and the approaches investigated and promoted by a large segment of the information retrieval research community. The former is based on the Boolean or Exact Matching retrieval model, whereas the latter ones subscribe to statistical and linguistic approaches, also referred to as the Partial Matching approaches. First, the major criticism leveled against the Boolean approach is that its queries are difficult to formulate. Second, the Boolean approach makes it possible to represent structural and contextual information that would be very difficult to represent using the statistical approaches. Third, the Partial Matching approaches provide users with a ranked output, but these ranked lists obscure

---

Key Problems	Possible Solutions
Selection of Search Vocabulary	• Thesaurus • Latent Semantic Indexing
Search strategy (re)formulation	• Smart Boolean • Statistical & Linguistic Approaches • Thesaurus • Graphical Interfaces
Information Overload	• Ranking • Clustering • Visualization

Table 2.6: lists some of the key problems in the field of information retrieval and possible solutions.

---

valuable information. Fourth, recent retrieval experiments have shown that the Exact and Partial matching approaches are complementary and should therefore be combined [Belkin et al. 1993].

In Table 2.6 we summarize some of the key problems in the field of information retrieval and possible solutions to them. We will attempt to show in this thesis: 1) how visualization can offer ways to address these problems; 2) how to formulate and modify a query; 3) how to deal with large sets of retrieved documents, commonly referred to as the information overload problem. In particular, this thesis overcomes one of the major "bottlenecks" of the Boolean approach by showing how Boolean coordination and its diverse narrowing and broadening techniques can be visualized, thereby making it more user-friendly without limiting its expressive power. Further, this thesis shows how both the Exact and Partial Matching approaches can be visualized in the same visual framework to enable users to make effective use of their respective strengths.