PREFACE

This thesis provides an overview of the state of the art in Information Retrieval (IR), both commercial practice and research. More specifically, it deals with the retrieval of unstructured and semi-structured text documents or messages. It doesn’t claim to be exhaustive. If it were exhaustive today, it wouldn’t be exhaustive tomorrow given the dynamic nature of the field.

The thesis is divided into five broad areas.

**Chapter one** discusses the basic concepts and definitions, e.g., what IR means, what its goals are, what entities it attempts to retrieve, the criteria by which IR systems are evaluated (and the limitations of those criteria), and how IR differs from retrieval via a traditional DBMS.

**Chapter two** discusses each of the major approaches to the generation of queries and their interpretation, by Information Retrieval engines: classical boolean, extended boolean, vector space, probabilistic, and semantic/natural language processing (NLP). It also discusses IR “querying” from the perspective of routing and classification of an incoming stream of documents (as distinct from their retrieval from fixed collections). (In the routing context, queries are often called *topics* or *classifications.*) Finally, it discusses methods of clustering documents within a collection as a form of unsupervised classification, and as an aid to efficient retrieval.

**Chapter three** discusses the automatic and interactive expansion and refinement of user-generated queries, e.g., on the basis of user “relevance” feedback. Additionally, it discusses the re-use of queries.

**Chapter four** discusses the “fusion” of streams of output documents resulting from multiple, parallel retrievals into a single ranked stream that can be presented to the user. Two kinds of fusion are discussed: (1) A given query may be issued to multiple document collections using a common IR method. The documents retrieved from each of those collections must be merged into a single stream, ideally the same stream that would have resulted if these separate collections had been integrated
into a single collection. (2) The same query may be executed by multiple IR methods (or the same information need may be formulated as multiple queries). In this way, a single query or information need may result in multiple retrievals being applied to the same document collection, each retrieval returning a different set of documents or a different ranking of the documents retrieved. Again, the results of these multiple retrievals must be merged and ranked for presentation to the user.

Chapter five discusses user interaction with IR systems, i.e., system aid in the formulation and refinement of queries, system display of data and retrieved results in ways that aid user understanding, user browsing of (and interaction with) displayed data and results, etc.

Chapter six discusses the ANSI/NISO standard Z39.50, initially developed by the library community for searching and retrieving bibliographic records, now emerging as a generic standard for communicating with diverse IR engines. The discussion includes both the existing 1995 standard and a proposed extension to the Z39.50 query capability, the type 102 query, which reflects enhancements in IR technology, especially the ability to retrieve documents ranked by the likelihood that they satisfy the user's information need.

Chapter seven illustrates the state of the art by discussing briefly six actual IR systems — four commercial and two research.

Chapter eight discusses Web information retrieval, including general concepts, research approaches, and representative commercial Web IR engines.