Preface

Satisfying information need (IN) is one of the eternal human problems. However, it is also an eternally new problem because the conditions of life, its content, and humans themselves are constantly changing. Throughout history and its various stages of development, human beings have developed various methods and forms for satisfying IN. In this century, characterized by an unprecedented growth of information, efforts to develop methods for the satisfaction of IN received systematic attention. This eventually led to the creation of information science, all areas of which focus directly or indirectly on satisfying IN.

One of the more important new forms of satisfying IN is the information retrieval (IR) system. These systems were first studied almost half a century ago, and since that time a great number of ideas have been introduced, and extensive experience has been gained as the result of using many real functioning IR systems. These successes, both theoretical and practical, provide the foundation for the theory of IR systems presented in this book. At the same time, because the theory of constructing IR systems is only one area of information science, and this theory relies significantly on a number of fundamental results in information science, this factor will also be taken into consideration in its description.

This book could be viewed as consisting of six parts. The first part (which includes Chapter 1) examines the general principles of constructing any system, particularly an IR system. The second part (Chapters 2 and 3) provides a detailed analysis of basic (key) concepts such as IN, information, information crisis, and the notion of information retrieval. The third part (Chapters 4 and 5) examines the goal, function, and structure of an IR system and information retrieval language (IRL) used in an IR system. The fourth part (Chapters 6, 7, 8, and 9) describes the construction of an IR system, that is, the methods, algorithms, and approaches to the realization of every structural element of the system. The fifth part of the book (Chapters 10 and 11) deals with different approaches to evaluating the results of information retrieval, the IR system itself, and separate components of an IR system. The sixth part (Chapter 12) describes some new

directions in the development of IR systems. The basic content and, in some cases, the orientation of the chapters can be summarized as follows.

Chapter 1 is intended primarily for those who are unfamiliar with the theoretical basis of systems analysis and design. The principles introduced in this chapter, which constitute the essence of the systems approach, are used in many of the subsequent chapters, and understanding these principles is often critical for understanding the text.

Chapter 2 provides an overview of the nature and properties of IN, that is, the need that an IR system must satisfy. It should be emphasized that in constructing an IR system, the knowledge of IN properties is very important because the quality of IN satisfaction depends on how fully and precisely the system takes these properties into account. It is for this reason that during the construction of every structural element of the system (the realization of every process in an IR system), the prime focus should be on taking into account the relevant properties of IN.

Some features of a search phenomenon, which do not depend on what is being searched, form the basis for creating any search system. These general features are described in Chapter 3. In the same chapter, a search for information is analyzed using these general features.

Chapter 4 gives a more precise description of the notion of the IR system and begins with a description of how to create IR systems. Following basic ideas of the systems approach (see Chapter 1), the goal of an IR system and its function are analyzed, and its structure is determined.

Chapter 5 examines the languages used for information retrieval. In addition to providing an analysis of theoretical questions, the chapter focuses particular attention on some practical approaches to creating such languages. The chapter describes a number of more popular criteria of similarity.

Chapter 6 begins the description of the construction of IR systems that is continued through Chapter 9. This chapter describes automatic methods for the indexing of documents. It also provides an example of a possible algorithm that is currently used in a functioning system.

Chapter 7 explores various approaches to automating the indexing of search requests. A large portion of this chapter contains a detailed analysis of existing problems, and at the end an algorithm for constructing query formulations in Boolean form is described in detail.

In describing Chapter 8, it is important to mention an approach used in this book. This chapter addresses questions dealing with storage and access to information. Because the methods for storage and access used in constructing IR systems are well known in computer science (within the context of courses on data structures and file organization, for example), we targeted the presentation to those readers who are not familiar with the material typically taught in these courses. For this reason, we avoided overloading the text with technical details.

Chapter 9 deals with the questions of realization in an IR system of a mechanism that allows the system to adapt to the user's IN. Furthermore, we solve another problem within the framework of adaptation—optimal search for an individual user. The chapter also describes a number of algorithms for solving these problems.

To understand the material on evaluating IR systems (described in Chapters 10 and 11), the reader must have some mathematical background. Further simplification of the described ideas and approaches without loss of coherence does not seem possible.

Chapter 12 considers different attempts for the further development of IR systems that use ideas borrowed from research on artificial intelligence. This chapter also presents one of the more promising new directions for perfecting IR systems based on a more complete accounting of the properties of IN.

Generally, in this book we attempted to write all chapters so that they would follow the same structure. Each chapter was to contain a definition of every object, phenomenon, and process; its general properties; and its role and place in information retrieval. Theoretical questions on the creation of IR systems are dealt with comprehensively and in detail in practically all 12 chapters of the book. Limitations can be found only in the presentation of methods used to construct an IR system, especially in Chapters 7 and 9. This may be explained by the high number of existing publications that deal with realizing different processes in an IR system in relation to all publications in the field of IR systems. Nearly a third of the publications in information science deal with information retrieval. Therefore, we found it impossible to include all approaches and methods for all components of an IR system, and hence we had to choose which methods would be included in the book. Because the purpose of the book (reflected by its title) is to describe the development of automated information retrieval systems, we did not include any manual methods and approaches to constructing different components of an IR system. In other words, we selected only the material that helped us demonstrate the importance and the possibility of a full automatization of any existing process in the system.

We should also point out some additional considerations in selecting the methods included in Chapters 7 and 9. In these chapters we examine algorithms used in the systems with Boolean searches. We did this for the following reasons. First, almost all functioning IR systems today use the Boolean search method. Because one of our objectives is to bring readers closer to the curent practice of using IR systems and to orient them toward the future development of the systems, not only is it expedient to discuss Boolean-oriented algorithms, but it is pragmatically justified.

Second, we are not aware of any books that discuss fully automatic methods for indexing search requests and feedback in functioning systems (that is, systems that conduct Boolean searches). This book intends to fill this void and thereby benefit those who already work with IR systems.

In addition, please note that the automated methods of the processes mentioned that are oriented toward other approaches (for example, the vector-space approach) are thoroughly addressed in existing literature (see, for example, Salton and McGill, *Introduction to Modern Information Retrieval,* which is fully cited at the end of Chapter 3). Therefore, we decided not to repeat this material in this book.

Because we did not attempt to write our book as an encyclopedia of the subject area, we excluded some technical and applied questions regarding IR systems (questions that would address such applications as reusable software and information filtering, among others). We tried to focus on theoretical questions regarding the creation of IR systems, which are independent of specific methods to be used in practically realized IR systems.

The text should be useful for computer science as well as information science and library science students. It is probably most appropriate for seniors and graduate students. Also, professional readers with academic or practical interest in the information retrieval field will, we hope, find it interesting and useful.

The authors express their gratitude to Professor Harold Borko of the University of California, Los Angeles, and to Professor Donald Kraft of Louisiana State University, whose involvement was one of the important factors in bringing this book to its published form. Many thanks to Michael Belkin, president of MCC International, for all of the friendly support he provided while the book was being written. Similarly, we are grateful to all those at Academic Press who participated in the book's publication. Finally, a big thanks to our families. They provided the coffee (in the case of two authors and tea with lemon for the third author) as well as encouragement and occasional inspiration during the long hours that were spent writing the book.

Valery I. Frants
Jacob Shapiro
Vladimir G. Voiskunskii