Mathematical Foundations of Information Retrieval. Sandor Dominich. Dordrecht: Kluwer; 2001: 284 pp. Price: $107.00 (ISBN: 0-7923-6861-4.)

The book under review presents a unified, mathematical description of information retrieval (IR) models. It concerns a personal view of the author, partly based on his Ph.D. work Dominich (1993) and later research of the author. As such it does not interact very much with similar efforts of other researchers but, nevertheless, the presented mathematical model is interesting and shows the relation between the different existing IR models. To give the reader of this review an idea of what it is all about we give a (simplified) description of the model from which, for example, vector IR (also called similarity IR) and probabilistic IR can be derived. Dominich's model is called "Classical Information Retrieval" (CIR), and goes as follows (essentially).

1. We have a set of documents $D = \{\tilde{o}_j | j \in J\}$, where $\tilde{o}_j$ is to be interpreted as the fuzzy set description of a document, for example, with $\mu_{\tilde{o}_j}(t_k) \in [0,1]$, being the "importance" of index term t_k ($k = 1, \ldots, N$) in document $\tilde{o}_j$.
2. We have a set of criteria $A = \{\tilde{a}_1, \ldots, \tilde{a}_C\}$, where each $\tilde{a}_i = \{((q, \tilde{o}_j), \mu_{\tilde{a}_i}(q, \tilde{o}_j)) | j \in J\}$ giving the relation between a query q and each of the documents $\tilde{o}_j$, $\mu_{\tilde{a}_i}(q, \tilde{o}_j) \in [0,1]$, to be specified later on (this will yield the different IR models).
3. We have an α_i-cut of criterion $\tilde{a}_i$: $a_{\alpha_i} = \{\tilde{o}_j \in D | \mu_{\tilde{a}_i}(q, \tilde{o}_i) > \alpha_i\}$, $\alpha_i \geq 0$, and finally a mapping:
4. $R: D \rightarrow \mathcal{P}(D) = 2^D$ (the set of all subsets of D), called retrieval.

To show that CIR comprises similarity IR (SIR), also called vector IR (VIR) as well as probability IR (PIR), Dominich gives the following argument. For SIR = VIR, take $C = 1$, $\mu_{\tilde{a}_1}$ symmetric and $R(q) = a_{\alpha_1} = \{\tilde{o} | \mu_{\tilde{a}_1}(q, \tilde{o}) > \alpha_1\}$. In other words, documents $\tilde{o}$ are retrieved, following a query q, if their similarity measure (e.g., vector inproduct) with q exceeds a certain threshold α_1. For PIR, take $C = 2$ and $\mathbf{R(q)} = \{\tilde{o} | \mu_{\tilde{a}_1}(q, \tilde{o}), \geq \mu_{\tilde{a}_2}(q, \tilde{o})\ \mu_{\tilde{a}_1}(q, \tilde{o}) > \alpha_1\}$. In other words, documents $\tilde{o}$ are retrieved, following a query q, if the probability that $\tilde{o}$ is relevant to q is larger than the probability that $\tilde{o}$ is irrelevant to q *and* that the probability of relevance of $\tilde{o}$ to q exceeds a certain threshold α_1.

Dominich then proceeds by giving an elaborate mathematical theory of CIR, SIR = VIR and PIR followed by an extension to "Interactive Information Retrieval" (IIR or I^2R) comprising CIR; hence, also SIR = VIR and PIR. In I^2R the connections between documents are made explicit and are used, and each time a new document is added the connections are recalculated. This is also done in case a query is formulated that is considered as a new document. These dynamic connections are modeled using neural network theory.

Dominich's models do not include Boolean retrieval, and this might be considered as a surprise: on each of the models described by the author, one can "superimpose" a Boolean structure by applying intersections and unions to retrievals that were obtained in the models. It is the reviewer's opinion that the Boolean model can be incorporated in, for example, SIR by taking C = 1 and applying a certain threshold (yielding retrievals going from pure OR relations to pure AND relations and anything in between) on the similarity measure between a query q and a document $\tilde{o}$ being the fraction of keywords of q that are in $\tilde{o}$, for example, $q = \{k_1, k_2, k_3, k_4\}$ $\tilde{o} = \{k_2, k_3, k_4, k_5, k_6\}$, then $\mu_{\tilde{a}_1}(q, \tilde{o}) = \frac{3}{4}$.

The last major point that is addressed by Dominich is the issue of relevance effectiveness (incl. relevance feedback) in IR. Also, this topic is studied in a mathematical way, and starts from the trivial but very interesting result that

$$\frac{\varphi\pi}{\rho(1 - \pi)} = \text{a constant},$$

where φ = fallout, π = precision, and ρ = recall, given the number of relevant documents to a problem. Hence, this yields a surface in three-dimensional space, called the effectiveness surface of IR, which is the same for every IR system! This surface is used to determine optimality in IR as well as relevance feedback (improving ρ and/or π under the constraint that $\varphi\pi/\rho(1 - \pi)$ is a constant). Feedback is introduced as a recursion function.

We will now give an overview of the book, chapter by chapter. Chapter 1 discusses basic IR topics such as vector retrieval, relevance (feedback), and gives an intuitive introduction to the mathematical development of IR to follow. Remarkable is the large Chapter 2 where possibly all mathematical tools, needed in the mathematical theory of IR, are mentioned. Topics include logics, set theory, relations, functions, families of sets, aspects of algebra, calculus, differential equations, probability theory, fuzzy sets, metric spaces, topology, graph theory, matroid theory, recursion, and complexity theory and neural networks. Although it is a luxury to have all these mathematical tools unified in a book of this level, it is the reviewer's impression that it is not really necessary: or one is familiar with the notions, in which case one does not need it here, or one is not, in which case the introduced notions are described in a too short way which does not allow a real understanding. In addition, all these notions are very well covered in the mathematical literature at the undergraduate level, and hence easily reachable in a scientific library. Chapter 3 discusses existing IR models: classical models (Boolean, vector space, probabilistic), nonclassical models (information logic, situation theory, interaction), and alternative models [cluster, fuzzy, LSI (Latent Semantic Indexing), AI (artificial intelligence)]. Chapter 4 then gives the mathematical theory as described above. Chapter 5 discusses relevance effectiveness, also described above. The final chapter discusses CIR and decision making, data fusion, interaction, and situation theory models. The three appendices present complex IR examples of VIR, PIR, and I^2R. They are not very instructive, in the reviewer's opinion.

The book is written in a clear way, and is very didactical: many pages are devoted to explain the goal of the book, the audience, interlinkings between the chapters. Also, possible course descriptions that can be constructed using the book are given, and there is a (too) extensive pure math chapter (70 pages!). For a book of this status there are too many misprints, and the index is far from complete. The reference list is very complete, but in some cases (e.g., Halmos, . . .) reference is given to the German language version, while an English version certainly exists, which would have been more logical for this English language book.

I would not recommend the reviewed work as a student text book in IR, but parts (e.g., Chapters 1 and 3) can be used as such. Also, the price of the book ($107) is too high for a student book (in fact, I think the price is too high in any application). Certainly, the book is stimulating further research on IR models. It is a refreshing presentation of mathematical theory of IR, unifying many mathematical and IR models. The book must be read by anyone interested in theoretical IR.

JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY, 53(5):409–411, 2002

Prof. Dr. L. Egghe
Universitaire Campus,
B-3590 Diepenbeek, Belgium.
e-mail: leo.egghe@luc.ac.be

DOI: 10.1002/asi.10027

References

Dominich, S. (1993). The formulation of the interaction information retrieval Model as a new and complementary framework for information retrieval. Ph.D. Thesis, Hungarian Academy of Sciences, Budapest, Hungary (in English).