References

Leech, Geoffrey. 1997. Grammatical tagging. In Roger Garside, Geoffrey Leech, and Tony McEnery, editors, *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Longman, London, pages 19–33.

Leech, Geoffrey and Elizabeth Eyes. 1997. Syntactic annotation: Treebanks. In Roger Garside, Geoffrey Leech, and Tony McEnery, editors, *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Longman, London, pages 34–52.

Mathematical Foundations of Information Retrieval

Sándor Dominich
(University of Veszprém)

Dordrecht: Kluwer Academic Publishers (Mathematical modelling: Theory and applications, edited by R. Lowen, volume 12), 2001, xxi+284 pp; hardbound, ISBN 0-7923-6861-4, $107.00, £67.00, €99.00

While libraries have long stored informative material for later use, information retrieval as we now know it did not begin to coalesce as a discipline until the 1960s and early 1970s, with advances in commercial systems and influential research publications by scholars including Lancaster, Maron, Salton, and Sparck Jones. While commercial systems of the time most commonly accepted Boolean queries as input, describing the relationships desired between human-assigned index terms in the documents to be retrieved, researchers began developing models and software consistent with term-weighting systems; these evolved into the methods used by today's search engines in ranking documents. While many of the term-weighting and automatic indexing schemes were initially rather simple, they have grown in complexity, based on developments in retrieval and linguistic theory and years of experimentation. While more sophisticated linguistic methods have been studied in retrieval contexts nearly as long as retrieval itself has been studied, the relative level of satisfaction with the performance of retrieval systems using simple automated indexing has kept the linguistic focus of retrieval researchers on individual terms, usually assuming a "bag of terms" model.

Dominich summarizes many of the mathematical foundations of various information retrieval models. Chapter 2 provides the core mathematical material in the book. A wide range of concepts is presented, with a section for each of the following: logic, set theory, relations, functions, families of sets, algebra, calculus, differential equations, vectors, probability, fuzzy sets, metric spaces, topology, graph theory, matroid theory, recursion and complexity theory, and artificial neural networks. The sections are typically broken down into formal definitions, theorems, and examples. The definitions and theorems are clear and relatively easy to understand. Those seeking longer or deeper mathematical expositions on these topics will need to go to the mathematical literature; however, in most cases, the material provided by Dominich will be adequate for linguists and retrieval specialists trying to understand retrieval models. The chapter has little on the relative strengths, weaknesses, and consequences of the adoption of particular mathematical paradigms, which may be frustrating to those asking "Why?" The numeric or symbolic examples provided at the ends of many sections are brief but very useful.

Chapter 3 addresses retrieval models, with one-third of the chapter addressing traditional text retrieval models (Boolean, vector, and probabilistic), one-third addressing "nonclassical" models that have yet to see much commercial use but would be of interest to philosophers and linguists (e.g., ideas based on the works of Barwise and Devlin), and one-third addressing "alternative" models of information retrieval, including a presentation on latent semantic indexing (four pages) and natural language processing (one page). Other brief discussions of techniques using linguistic information are spread throughout the chapters.

Chapter 4 addresses how information retrieval, as modeled in Chapter 3, may be based rigorously on the mathematics presented in Chapter 2. Much of this constitutes original research. It is helpful to see the work of a variety of authors brought together into the uniform presentation provided by Dominich. Chapter 5 formally examines retrieval effectiveness. This presentation goes into far more depth than do most information retrieval books and provides a helpful summary of the foundations of the

area. The appendices contain algorithms and MathCAD code for several retrieval models.

The book uses a section-numbering style where the chapter number isn't at the beginning of each section number, nor is the chapter number in the header or the footer for most pages. This makes navigating through the book unnecessarily difficult.

Mathematical Foundations of Information Retrieval will be useful to those computational linguists who want an accessible yet mathematically rigorous presentation of the foundations of information retrieval algorithms and models.—*Robert Losee, University of North Carolina*