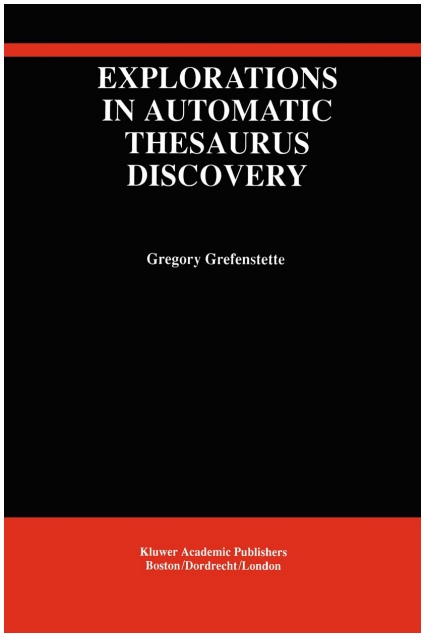

1994, XIII, 305 p.

Printed book

Hardcover

ISBN 978-0-7923-9468-6

- **164,99 € | £145.00**
- ***176,54 € (D) | 181,49 € (A) | CHF 194.50**

G. **Grefenstette**

Explorations in Automatic Thesaurus Discovery

Series: The Springer International Series in Engineering and Computer Science, Vol. 278

Explorations in Automatic Thesaurus Discovery presents an automated method for creating a first-draft thesaurus from raw text. It describes natural processing steps of tokenization, surface syntactic analysis, and syntactic attribute extraction. From these attributes, word and term similarity is calculated and a thesaurus is created showing important common terms and their relation to each other, common verb--noun pairings, common expressions, and word family members.

The techniques are tested on twenty different corpora ranging from baseball newsgroups, assassination archives, medical X-ray reports, abstracts on AIDS, to encyclopedia articles on animals, even on the text of the book itself. The corpora range from 40,000 to 6 million characters of text, and results are presented for each in the Appendix.

The methods described in the book have undergone extensive evaluation. Their time and space complexity are shown to be modest. The results are shown to converge to a stable state as the corpus grows. The similarities calculated are compared to those produced by psychological testing. A method of evaluation using Artificial Synonyms is tested. Gold Standards evaluation show that techniques significantly outperform non-linguistic-based techniques for the most important words in corpora.

Explorations in Automatic Thesaurus Discovery includes applications to the fields of information retrieval using established testbeds, existing thesaural enrichment, semantic analysis. Also included are applications showing how to create, implement, and test a first-draft thesaurus.