第一图书网, tushu007.com

<<文本挖掘>>

图书基本信息

书名：<<文本挖掘>>

13位ISBN编号：9787115205353

10位ISBN编号：7115205353

出版时间：2009-8

出版时间：人民邮电出版社

作者：（以）费尔德曼,（美）桑格

页数：410

字数：506000

版权说明：本站所提供下载的PDF图书仅提供预览和简介，请支持正版图书。

更多资源请访问：http://www.tushu007.com

第一图书网, tushu007.com

前言

The information age has made it easy to store large amounts of data. The prolifera- tion of documents available on the Web, on corporate intranets, on news wires, and elsewhere is overwhelming. However, although the amount of data available to us is constantly increasing, our ability to absorb and process this information remains constant. Search engines only exacerbate the problem by making more and more documents available in a matter of a few key strokes. Text mining is a new and exciting research area that tries to solve the information overload problem by using techniques from data mining, machine learning, natural language processing (NLP), information retrieval (IR), and knowledge management. Text mining involves the preprocessing of document collections (text categorization, information extraction, term extraction), the storage of the intermediate represen- tations, the techniques to analyze these intermediate representations (such as distri- bution analysis, clustering, trend analysis, and association rules), and visualization of the results.

This book presents a general theory of text mining along with the main tech- niques behind it. We offer a generalized architecture for text mining and outline the algorithms and data structures typically used by text mining systemg The book is aimed at the advanced undergraduate students, graduate students, academic researchers, and professional practitioners interested in complete cov- erage of the text mining field. We have included all the topics critical to people who plan to develop text mining systems or to use them. In particular, we have covered preprocessing techniques such as text categorization, text clustering, and information extraction and analysis techniques such as association rules and link analysis. The book tries to blend together theory and practice; we have attempted to provide many real-life scenarios that show how the different techniques are used in practice. When writing the book we tried to make it as self-contained as possible and have compiled a comprehensive bibliography for each topic so that the reader can expand his or her knowledge accordingly.

第一图书网, tushu007.com

<<文本挖掘>>

内容概要

本书是一部文本挖掘领域名著，作者为世界知名的权威学者。
书中涵盖了核心文本挖掘操作、文本挖掘预处理技术、分类、聚类、信息提取、信息提取的概率模型、预处理应用、可视化方法、链接分析、文本挖掘应用等内容，很好地结合了文本挖掘的理论和实践。

本书非常适合文本挖掘、信息检索领域的研究人员和实践者阅读，也适合作为高等院校计算机及相关专业研究生的数据挖掘和知识发现等课程的教材。

第一图书网, tushu007.com

作者简介

Ronen Feldman 机器学习、数据挖掘和非结构化数据管理的先驱人物。
以色列Bar-Ilan大学数学与计算机科学系高级讲师、数据挖掘实验室主任，Clearforest公司（主要为企业和政府机构开发下一代文本挖掘应用）合作创始人、董事长，现在还是纽约大学Stern商学院的副教授。

第一图书网, tushu007.com

<<文本挖掘>>

书籍目录

第一图书网, tushu007.com

章节摘录

Similarity Functions for Simple Concept Association Graphs Similarity functions often form an essential part of working with simple concept asso- ciation graphs, allowing a user to view relations between concepts according to differ- ing weighting measures. Association rules involving sets （or concepts） A and B that have been described in detail in Chapter II are often introduced into a graph format in an undirected way and specified by a support and a confidence threshold. A fixed confidence threshold is often not very reasonable because it is independent of the sup- port from the RHS of the rule. As a result, an association should have a significantly higher confidence than the share of the RHS in the whole context to be considered as interesting. Significance is measured by a statistical test （e.g., t-test or chi-square）. With this addition, the relation given by an association rule is undirected. An asso- ciation between two sets A and B in the direction AB implies also the association B A. This equivalence can be explained by the fact that the construct of a statisti- cally significant association is different from implication （which might be suggested by the notation AB）. It can easily be derived that if B is overproportionaUy represented in A, then A is also overproportionally represented in B. As an example of differences of similarity functions, one can compare the undi- rected connection graphs given by statistically significant association rules with the graphs based on the cosine function. The latter relies on the cosine of two vectors and is efficiently applied for continuous, ordinal, and also binary attributes. In case of documents and concept sets, a binary vector is associated to a concept set with the vector elements corresponding to documents. An element holds the value 1 if all the concepts of the set appear in the document. Table X.1 （Feldman, Kloesgen, and Zilberstein 1997b）, which offers a quick summary of some common similarity functions, shows that the cosine similarity function in this binary case reduces to the fraction built by the support of the union of the two concept sets and the geometrical mean of the support of the two sets. A connection between two sets of concepts is related to a threshold for the cosine similarity （e.g., 10%）. This means that the two concept sets are connected if the support of the document subset that holds all the concepts of both sets is larger than 10 percent of the geometrical mean of the support values of the two concept sets.

第一图书网, tushu007.com

媒体关注与评论

“……我购买了这本书。
这本书绝对是非常值得拥有的参考书。
”　　——L.Venkata Subramaniam，IBM印度研究实验室　　“一本由该领域最重要专家编写的文本挖掘导论。
这本书写得非常好。
完美地结合了文本挖掘的理论和实践，既适合研究人员又适合实践者……极力推荐那些没有任何计算语言学背景而想钻研文本挖掘领域的人阅读本书。
”　　——Rada Mihalcea，北得克萨斯大学　　文本挖掘已经成为令人兴奋的新兴研究领域。
本书由世界知名的权威学者编写，除了讲解核心文本挖掘和链路检测算法及技术之外，还介绍了高级预处理技术。
并考虑了知识表示方面的因素以及可视化方法。
此外。
书中还探讨了有关技术在实践中的应用，很好地兼顾了文本挖掘的理论和实践

第一图书网, tushu007.com

版权说明

本站所提供下载的PDF图书仅提供预览和简介，请支持正版图书。

更多资源请访问:http://www.tushu007.com