

B 大数据丛书
IG DATA SERIES

TEXT MINING APPLICATIONS AND THEORY

文本挖掘

迈克尔·W.贝瑞

MICHAEL W. BERRY

【美】

雅克布·柯岗

JACOB KOGAN

编

文卫东 译

 机械工业出版社
CHINA MACHINE PRESS

WILEY



大数据丛书

文本挖掘

[美] 迈克尔·W. 贝瑞 (Michael W. Berry) 编
雅克布·柯岗 (Jacob Kogan) 译
文卫东 译

机械工业出版社

本书呈现了文本挖掘领域先进的算法，同时从学术界和产业界的角度介绍了文本挖掘。本书涉及的业界学者跨越多个国家，来自多个机构：大学、企业和政府实验室。本书介绍了文本挖掘在多个领域中的自动文本分析和挖掘计算模型，这些领域包括：机器学习、知识发现、自然语言处理和信息检索等。

本书适合作为人工智能、机器学习和自然语言处理等领域相关人员的教科书和参考书。同时，也适合研究人员和从业人员阅读。

Copyright © 2010, John Wiley & Sons, Ltd

All Rights Reserved. This translation published under license. Authorized translation from the English language edition, entitled Text Mining: Applications and Theory, ISBN: 9780470749821, by Michael W. Berry and Jacob Kogan, Published by John Wiley & Sons. No part of this book may be reproduced in any form without the written permission of the original copyrights holder.

本书中文简体字版由 Wiley 授权机械工业出版社独家出版，未经出版者书面允许，本书的任何部分不得以任何方式复制或抄袭。

版权所有，翻印必究。

北京市版权局著作权合同登记 图字：01-2013-2607 号。

图书在版编目 (CIP) 数据

文本挖掘/(美) 迈克尔·W. 贝瑞 (Michael W. Berry), (美) 雅克布·柯岗 (Jacob Kogan) 编; 文卫东译. —北京: 机械工业出版社, 2017. 5

(大数据丛书)

书名原文: Text Mining: Applications and Theory

ISBN 978-7-111-57050-9

I. ①文… II. ①迈…②雅… ③文… III. ①数据采集—研究
IV. ①TP274

中国版本图书馆 CIP 数据核字 (2017) 第 130385 号

机械工业出版社 (北京市百万庄大街 22 号 邮政编码 100037)

策划编辑: 韩效杰 责任编辑: 韩效杰 陈崇昱

责任校对: 佟瑞鑫 封面设计: 路恩中

责任印制: 孙 炜

保定市中国画美凯印刷有限公司印刷

2019 年 1 月第 1 版第 1 次印刷

169mm × 239mm · 11.25 印张 · 222 千字

标准书号: ISBN 978-7-111-57050-9

定价: 49.00 元

凡购本书, 如有缺页、倒页、脱页, 由本社发行部调换

电话服务

网络服务

服务咨询热线: 010-88361066

机工官网: www.cmpbook.com

读者购书热线: 010-68326294

机工官博: weibo.com/cmp1952

010-88379203

金书网: www.golden-book.com

封面防伪标均为盗版

教育服务网: www.cmpedu.com

译者序

随着网络时代的到来，用户可获得的信息包含了从技术资料、商业信息到新闻报道、娱乐资讯等多种类别和形式的文档，这样便构成了一个异常庞大的，具有异构性和开放性等特点的分布式数据库，而这个数据库中存放的一般都是非结构化的文本数据。如何处理文本数据并挖掘数据中所隐含的意义，对政府的政策指导，对企业的精准营销，以及对机构的风险防范等都具有很高的价值。因此，目前社会各界对于文本挖掘的需求非常强烈，文本挖掘技术的应用前景广阔。

文本挖掘就是抽取有效、新颖、有用、可理解、散布在文本文件中的有价值知识，并且利用这些知识更好地组织信息的过程。1998年年底，国家重点研究发展规划首批实施项目中曾明确指出，文本挖掘是“图像、语言、自然语言理解与知识挖掘”中的重要内容。

文本挖掘利用智能算法，如神经网络、基于案例的推理、可能性推理等，并结合文字处理技术，分析大量的非结构化文本源（如文档、电子表格、客户电子邮件、问题查询、网页等），抽取或标记关键字概念、文字间的关系，并按照内容对文档进行分类，获取有用的知识和信息。

本书的编者从产业界与学术界的角度总揽并分析了文本挖掘的最先进算法和模型，分别从关键词提取、分类与聚类、突发事件与趋势预测、文本流处理四个方面深入浅出地总结和分析了其所研究的问题及解决办法，值得一读。

译者长期从事文本挖掘研究与应用，此书也是译者所在课题组学习与参考的重要书籍之一。了解文本挖掘的内容，掌握文本挖掘的方法，以及灵活运用，这些都是文本挖掘领域研究者和应用者的迫切需求，本书既可作为教科书，也可作为参考书。为了向国内读者及时提供高质量的译本，我们课题组人员利用工作之余翻译了此书。在翻译过程中，我们不放过任何一个疑点，尽可能地使用国内通用的专业术语来表述，尽管如此，仍有可能存在一些遗漏的问题和错误，恳请读者在阅读过程中发现问题后不吝指教。

最后要感谢课题组陈振、薛冰在翻译过程中所做的工作，感谢机械工业出版社编辑团队以及其他同仁的帮助。

文卫东

原 书 序

随着数字计算设备的普及和它在通信中的应用，人们对挖掘文本数据的系统和算法的需求与日俱增。因此，非结构化、半结构化与完全结构化文本数据挖掘技术的发展对于学术界和产业界来说都相当重要。2009年5月2日，为期一天的文本挖掘专题研讨会与SIAM第九届数据挖掘国际学术会议一起举行，会议汇集了来自不同学科的研究人员，他们提出了目前在文本挖掘研究与应用中的方法和结果。会议研讨了文本挖掘的新兴领域、机器学习与自然语言处理相结合的技术上的应用、信息提取、信息检索的代数或数学方法等方向。从新的文档分类和聚类模型到主题检测、跟踪和可视化的新方法的发展，在这一领域的许多问题正在解决之中。

来自六个不同国家的、共计超过40位应用数学家和计算机科学家分别代表大学、企业和政府实验室参会。会议通过特邀报告以及会议论文报告这两种形式，讨论了使用机器学习、知识发现、自然语言处理和检索设计计算模型进行自动文本分析和挖掘等技术的应用。会议所提交的大多数特邀论文和投稿论文都已被编进本书。总的来说，这些论文的内容跨越了几个主要的文本挖掘方向上的主题领域：

1. 关键词提取；
2. 分类与聚类；
3. 突发事件与趋势预测；
4. 文本流处理。

本书介绍了目前从产业界与学术界的角度进行文本挖掘的最先进的算法，书中的每一章都是独立的，并含有一系列的参考文献。读者在阅读本书的一些章节之前需要熟悉一些基本的本科数学知识。本书既可供本领域的初学者学习，也可供研究文本挖掘领域的专家参考。

类似研究者所写的文字和读者所使用的文字的内在差异，持续推动了有效的搜索和检索算法与软件在文本挖掘领域的发展。本书展示了人们是如何利用应用数学、计算机科学、机器学习和自然语言处理等领域的最新进展来获取、分类并翻译文本及其上下文的。

迈克尔 W. 贝瑞 雅克布·柯岗

分别于田纳西州的诺克斯维尔和马里兰州的巴尔的摩

目 录

译者序

原书序

第 1 章 独立文档的关键词的自动提取	1
1.1 简介	1
1.1.1 关键词提取方法	1
1.2 快速自动关键词提取	3
1.2.1 候选关键词	3
1.2.2 关键词得分	4
1.2.3 邻接关键词	5
1.2.4 提取关键词	5
1.3 基准评估	6
1.3.1 准确率和召回率评估	6
1.3.2 效率评估	7
1.4 停用词列表生成	9
1.5 新闻消息的评估	12
1.5.1 MPQA 语料库	12
1.5.2 从新闻消息中提取关键词	12
1.6 总结	15
参考文献	16
第 2 章 利用数学方法进行多语言文档聚类	17
2.1 简介	17
2.2 背景	17
2.3 实验设置	18
2.4 多语言 LSA	20
2.5 Tucker1 方法	21
2.6 PARAFAC2 方法	23
2.7 词对齐的 LSA	24
2.8 潜在形态语义分析 (LMSA)	26
2.9 词对齐的 LMSA	27
2.10 对技术和结果的讨论	27
参考文献	29
第 3 章 使用机器学习算法对基于内容的垃圾邮件进行分类	31
3.1 简介	31

3.2	机器学习算法	32
3.2.1	朴素贝叶斯	33
3.2.2	LogitBoost	33
3.2.3	支持向量机	34
3.2.4	增广的潜在语义索引空间	35
3.2.5	径向基函数网络	36
3.3	数据预处理	37
3.3.1	特征选择	37
3.3.2	信息表示	39
3.4	邮件分类的评估	39
3.5	实验	40
3.5.1	使用 PU1 的实验	40
3.5.2	使用 ZH1 的实验	42
3.6	分类器特点	43
3.7	结束语	45
	参考文献	45
第4章	利用非负矩阵分解研究邮件分类问题	47
4.1	简介	47
4.1.1	相关工作	48
4.1.2	概要	49
4.2	研究背景	49
4.2.1	非负矩阵分解	49
4.2.2	计算 NMF 的算法	50
4.2.3	数据集	52
4.2.4	解释	52
4.3	基于特征排序的 NMF 初始化	54
4.3.1	特征子集选择	54
4.3.2	FS 初始化	55
4.4	基于 NMF 的分类方法	57
4.4.1	使用基础特征分类	58
4.4.2	基于 NMF 的一般化 LSI	59
4.5	结束语	65
	参考文献	66
第5章	使用 k-均值算法进行约束聚类	68
5.1	简介	68
5.2	表示法和古典 k -均值算法	69
5.3	具有布莱格曼散度的 k -均值约束聚类算法	70
5.3.1	具有“不能链接”约束关系的二次 k -均值聚类	70
5.3.2	“必须链接”约束关系的移除	73

5.3.3 使用布莱格曼散度进行聚类	75
5.4 smoka 类型约束聚类	77
5.5 球形 k -均值约束聚类	79
5.5.1 仅有“不能链接”约束关系的球形 k -均值聚类算法	80
5.5.2 具有“不能链接”和“必须链接”约束关系的球形 k -均值聚类	82
5.6 数值实验	83
5.6.1 二次 k -均值聚类	84
5.6.2 球形 k -均值聚类	85
5.7 总结	85
参考文献	86
第 6 章 文本可视化技术的研究	88
6.1 文本分析的可视化	88
6.2 标签云图	89
6.3 著作权及其变更的追踪	90
6.4 数据探索和 novel 模式的探索	91
6.5 情绪追踪	92
6.6 可视化分析和 FutureLens	94
6.7 场景发现	94
6.7.1 场景	94
6.7.2 评估策略	95
6.8 早期版本	95
6.9 FutureLens 的特征	96
6.10 场景发现举例: 生态恐怖主义	97
6.11 场景发现举例: 毒品走私	101
6.12 未来的工作	103
参考文献	104
第 7 章 新颖性挖掘的自适应阈值设置	106
7.1 简介	106
7.2 新颖性挖掘中的自适应阈值设置	107
7.2.1 背景	107
7.2.2 动机	108
7.2.3 基于高斯分布的自适应阈值设置	108
7.2.4 实现过程中的问题	112
7.3 实验研究	113
7.3.1 数据集	113
7.3.2 加工实例	113
7.3.3 实验及结果	116
7.4 总结	120
参考文献	121

第 8 章 文本挖掘与网络犯罪	122
8.1 简介	122
8.2 网络欺凌和网络捕食研究的现状	123
8.2.1 获取即时通信和在线聊天	124
8.2.2 当前用于分析的收集	124
8.2.3 对即时通信和在线聊天的分析	125
8.2.4 网络捕食检测	125
8.2.5 网络欺凌检测	129
8.2.6 法律问题	130
8.3 监控聊天的商业软件	131
8.4 结论与未来的方向	132
参考文献	133
第 9 章 文本流中的事件和发展趋势	136
9.1 引言	136
9.2 文本流	138
9.3 特征提取和数据还原	138
9.4 事件监测	139
9.5 趋势检测	142
9.6 事件和趋势描述	143
9.7 相关讨论	147
9.8 总结	147
参考文献	148
第 10 章 在 LDA 主题模型中嵌入语义	150
10.1 简介	150
10.2 背景	150
10.2.1 向量空间模型	151
10.2.2 潜在语义分析	151
10.2.3 概率潜在语义分析	151
10.3 潜在狄利克雷分配	152
10.3.1 图模型和生成过程	153
10.3.2 后验推断	153
10.3.3 在线潜在狄利克雷分配 (OLDA)	154
10.3.4 算例分析	156
10.4 在维基百科中嵌入外部语义	158
10.4.1 相关维基百科文章	158
10.4.2 维基百科影响的主题模型	158
10.5 数据驱动语义的嵌入	159
10.5.1 数据驱动语义嵌入的生成过程	159
10.5.2 嵌入数据驱动语义的 OLDA 算法	160

10.5.3 实验设计	161
10.5.4 实验结果	163
10.6 相关工作	166
10.7 结论与未来工作	166
参考文献	166

第 1 章 独立文档的关键词的自动提取

Stuart Rose, Dave Engel, Nick Cramer 和 Wendy Cowley

1.1 简介

关键词，可简洁代表一个文档的内容，理想的关键词是对一篇文章基本内容的浓缩。由于关键词很容易被定义、校正、记忆和共享，所以它们被广泛应用于信息检索（Information Retrieval, IR）系统中的查询。与数学符号相比，关键词独立于任意语料库，它可以被应用在多语料库和信息检索系统中。

关键词也可以用于改善信息检索系统的功能。Jones 和 Paynter (2002) 发表了 Phrasier 的相关文献，Phrasier 是一个可以把与主文档关键词相关联的文档罗列出来的系统，其中有将关键字锚点作为文档之间的超链接提供给用户的功能，该功能使得用户能够快速访问相关文档。Gutwin 等人 (1999) 则将 Keyphind 描述成把从文档中提取出的关键词作为 IR 的基本框架的系统。关键词也可以用来丰富检索结果。Hulth (2004) 描述 Keegle 是一个可以动态提取关键词的系统，它主要是为谷歌检索页面进行关键词提取。Andrade 和 Valencia (1998) 提出了一个系统，它可以自动地从与已知蛋白质相关的科学文档中提取出关键词来标记蛋白质的功能。

1.1.1 关键词提取方法

尽管关键词对于分析、索引和检索都很有帮助，但大部分的文档却都没有指定关键词。现有的关键词提取方法大都集中于专业人员的手动提取，而这些专业人员很有可能只使用固定的分类方法，或者根据作者的意图来提供典型的关键词列表。因此，目前的研究主要集中在从文档中自动提取关键词的方法上，这种方法为专业的检索工具给出了建议的关键词，或者为不可直接获取的文档产生特征摘要。

一些早期的自动提取关键词的方法主要集中于面向语料库的单个词语的统计分析的评估。Jones (1972) 和 Salton 等人 (1975) 把词汇搜索中出现的正确结果描述为跨语料库的可统计识别的单词。后来的许多关键词提取研究都使用了这种度量方法从而在文档中选择可识别的关键词。比如，Andrade 和 Valencia (1998) 提出的方法就是基于目标文档中词频的分布与参考语料库中相应分布的比较来提取可识别的关键词的。

虽然有很多关键词可以通过统计学的方法提取出来，但是仍然有很多关键词由于在文档中出现的次数过少而不能被提取出来。在语料库中的很多文档中都出现的关键词有时候是不能被统计识别出来的。面向语料库的方法是典型的只对单个词语进行操作的方法，而且单个的词语经常会被应用在不同主题的文档中，这就进一步

局限了统计可识别的词语。

为了避免这些问题，我们把注意力集中在独立文档的关键词提取方法上。这种面向文档的方法可以从任意语料库的文档中把关键词提取出来。面向文档的方法也就因此拥有了与语境无关的特点，从而可以使用一些描述文本流随时间发生变化的分析方法，如 Engel 等人（2009）、Whitney 等人（2009）发表的分析方法。这些面向文档的方法适合于变动的语料库，比如说随着时间变化的摘要集合或者是新闻文章流。另外，这些方法是在单一文档中进行操作的，因此它们可以被很容易地扩展到大量文档中，而且还可以应用于许多不同的场景以增强 IR 系统和分析工具的功能。

在面向文档的关键词处理方法方面，早先的工作主要是利用自然语言处理的方法来识别词性标签（part-of-speech tags, POS tags），该方法结合了监督学习、机器学习算法和统计方法。

Hulth（2003）使用监督机器学习的算法并且选择了四种可识别的特性作为自动提取关键词方法的输入，并且还比较了 noun-phrase（NP）chunks, n-grams, 以及 POS tags 这三种选择方法的功效。

Mihalcea 和 Tarau（2004）描述了一个应用一系列的句法过滤器来确认词性标签（这些标签可以用来选择词语作为关键词）的系统。在一个固定大小的滑动窗口中，这些被选中词的共同出现频率将在词共现图中体现出来。一个基于图的算法 TextRank 被用来进行词排列，词排列是基于词在词共现图中的关联程度，排在前面的词就被选为关键词。在文档中邻接出现的词语结合起来组成多词语的关键词组。Mihalcea 和 Tarau（2004）的报告称 TextRank 算法在只有名词和形容词被选为潜在关键词的时候其效果才是最好的。

Matsuo 和 Ishizuka（2004）用卡方测度来计算在同一句子内的词和短语共同出现的次数，并利用此测量结果选择出文档中的高频词汇。卡方测度用来纠正文档中词共现的某些偏差，为之后用词或短语的排列来选取关键词打下基础。Matsuo 和 Ishizuka（2004）指出，在词频非常小的情况下计算出的偏差其实并不可靠。作者提出了一个用于完整文档的评估方法，并在一个 27 页的文档上进行了实验，结果显示他们的方法对于规模较大的文档效果是比较好的。

在下面的部分中，我们将会重点介绍快速自动关键词提取（Rapid Automatic Keyword Extraction, RAKE）。这种方法可以在无监督的情况下，不限制研究领域和使用语言地为独立文档提取关键词。我们在这里提供了其算法的详细内容和参数结构，并且展示了在标准数据库中实验的结果。结果显示，相对于 TextRank, RAKE 表现出了更高的准确率和相差不多的召回率。然后我们描述了一种生成停用词列表的新方法，并且结合 RAKE 来解决特定领域和语料库的问题。最后，我们把 RAKE 应用在新闻消息上，并定义相关矩阵来对关键词提取进行排他性、必要性和普遍性的评估，使得这个系统在没有人工注释的情况下可以识别出必要且普遍的关键词。

1.2 快速自动关键词提取

在 RAKE 的发展过程中，我们的目的曾经是找到一种关键词提取方法：它是对某个独立文档的操作，而且可以被应用到动态集合；它可以方便地应用于新的领域并且在不同类型的文档中均有较好的表现，特别是对那些不遵从特定语法规则的文章。图 1.1 包含了文章的标题和正文，以及人工提取出的关键词。

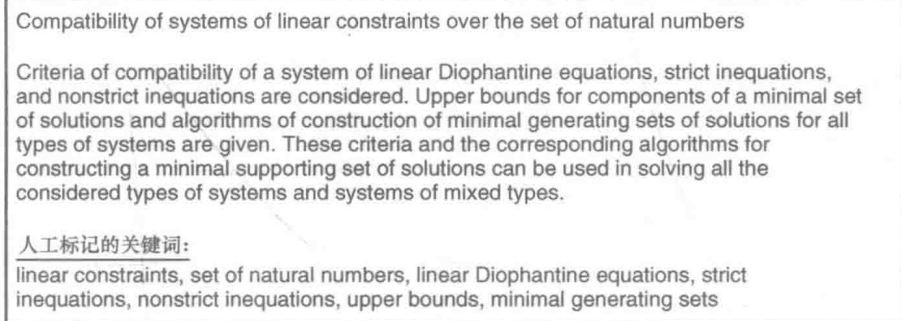


图 1.1 测试集中的样本文摘和人工提取的关键词

RAKE 是基于对不包含标点或者停用词的关键词词频的观察结果，比如说功能词语 and、the 和 of，或者其他不包含重要意义的词语。观察图 1.1 中人工对文摘提取的关键词，这里只有一个关键词包含一个停用词（set of natural numbers 中的 of）。停用词在信息检索指标中是要被去除的词汇，而且它们被认为不含有足够的信息因而不会被纳入不同文本的分析中去。原因是这些词语太过频繁的出现和太过广泛的应用会使其对用户的分析和查询毫无帮助。在一个文档中，富含意义的词汇描述了该文档的内容框架，它们通常被视为内容的关键词。

RAKE 的输入参数包含一系列的停用词、短语分隔符和词分隔符。RAKE 使用停用词和短语分隔符把文档分成一系列的候选关键词。这些候选关键词中的共现词是有意义的，它可以避免我们应用任意大小的滑动窗口去识别共现词。词之间的联系就可以通过一种根据文本的风格和内容自适应的方式来进行度量。词共现的、自适应的和细颗粒度的度量将被用于对候选关键词的评分上。

1.2.1 候选关键词

RAKE 是从把文章分成一些候选关键词集合开始的：首先，文档被指定的词分隔符分成一系列的词集；然后这些词集会被短语分隔符和停用词分成一系列连续的词。位于一个序列上的词将会被标记在文章中的相同位置，而且会被一同标记成为候选关键词。

图 1.2 中的候选关键词是从图 1.1 中的样本文摘中分析出来的。关键词 linear Diophantine equations 是从停用词 of 开始到逗号结束。紧接着的词 strict 则是下一个

候选关键词 strict inequations 的开始。

Compatibility – systems – linear constraints – set – natural numbers – Criteria – compatibility – system – linear Diophantine equations – strict inequations – nonstrict inequations – Upper bounds – components – minimal set – solutions – algorithms – minimal generating sets – solutions – systems – criteria – corresponding algorithms – constructing – minimal supporting set – solving – systems – systems

图 1.2 从样本文摘中提取出的候选关键词

1.2.2 关键词得分

在每一个关键词被确定以后，词共现图就完成了（见图 1.3）。每一个候选词都要被计算得分，它的得分是其每个词项的得分的和。我们基于很多度量方法来计算词项的得分，包括：（1）词频 $\text{freq}(w)$ ；（2）词度 $\text{deg}(w)$ ；（3）词度与词频的比 $\text{deg}(w)/\text{freq}(w)$ 。

	algorithms	bounds	compatibility	components	constraints	constructing	corresponding	criteria	diophantine	equations	generating	inequations	linear	minimal	natural	nonstrict	numbers	set	sets	solving	strict	supporting	system	systems	upper
algorithms	2						1																		
bounds		1																							1
compatibility			2																						
components				1																					
constraints					1								1												
constructing						1																			
corresponding	1						1																		
criteria								2																	
diophantine									1 1			1													
equations									1 1			1													
generating										1		1						1							
inequations												2				1					1				
linear					1				1 1				2												
minimal										1			3					2 1				1			
natural													1		1										
nonstrict											1				1										
numbers														1		1									
set													2				3					1			
sets										1		1						1							
solving																			1						
strict											1										1				
supporting														1				1				1			
system																							1		
systems																								4	
upper		1																							1

图 1.3 样本文摘中的词共现图

图 1.4 中展示了样本文摘的每一个实词的得分。从图中可以看出： $\text{deg}(w)$ 比较青睐于频繁出现的词或者是在较长的候选关键词中出现的词； $\text{deg}(\text{minimal})$ 的得分要比 $\text{deg}(\text{systems})$ 的得分高；高词频的 $\text{freq}(w)$ 不受其共现词的影响； $\text{freq}(\text{systems})$ 要比 $\text{freq}(\text{minimal})$ 的得分高；出现在长候选关键词中的词的 $\text{deg}(w)/\text{freq}(w)$ 都比较高； $\text{deg}(\text{diophantine})/\text{freq}(\text{diophantine})$ 比

deg (linear)/freq (linear)高。每一个候选关键词的得分是通过其成员词项的得分相加而得到的。图 1.5 列出了样本文摘中每一个候选关键词的 $\text{deg}(w)/\text{freq}(w)$ 。

	algorithms	bounds	compatibility	components	constraints	constructing	corresponding	criteria	diophantine	equations	generating	inequations	linear	minimal	natural	nonstrict	numbers	set	sets	solving	strict	supporting	system	systems	upper
deg(w)	3	2	2	1	2	1	2	2	3	3	3	4	5	8	2	2	2	6	3	1	2	3	1	4	2
freq(w)	2	1	2	1	1	1	1	2	1	1	1	2	2	3	1	1	1	3	1	1	1	1	1	4	1
deg(w) / freq(w)	1.5	2	1	1	2	1	2	1	3	3	3	2	2.5	2.7	2	2	2	2	3	1	2	3	1	1	2

图 1.4 词的得分

minimal generating sets (8.7), linear diophantine equations (8.5), minimal supporting set (7.7), minimal set (4.7), linear constraints (4.5), natural numbers (4), strict inequations (4), nonstrict inequations (4), upper bounds (4), corresponding algorithms (3.5), set (2), algorithms (1.5), compatibility (1), systems (1), criteria (1), system (1), components (1), constructing (1), solving (1)

图 1.5 候选关键词和得分

1.2.3 邻接关键词

由于 RAKE 是按照停用词来划分候选关键词的，所以通过这种方法提取出的关键词内部不包含停用词。随着 RAKE 在提取专业术语能力上所展现出的相当大的优势，它能够识别在内部包含停用词的关键词（比如 axis of evil）的优势也越来越明显。找出这些关键词需要找出在同一文档的同一位置上至少邻接出现两次的关键词。一个新的关键词就是由这些关键词和内部停用词组合而成的。这个新的关键词的得分将由其组成词项的得分之和组成。

需要指出的是，一些邻接关联的关键词被提取出来无疑在某种程度上增加了这个关键词的重要性。由于邻接关键词必须在同一文档的同一位置上至少出现两次，因此较长文档的关键词提取比较短文档的关键词提取要简单得多。

1.2.4 提取关键词

在候选关键词被计算得分之后，前 T 个得分最高的关键词将最终被选为关键词，如同 Mihalcea 和 Tarau (2004) 所做的那样，我们取 T 为词共现图中词的数量三分之一。

样本文摘一共包括 28 个实词，所以取 T 为 9。表 1.1 是 RAKE 提取出的关键词和人工标记的关键词对比。我们通过统计方法来测量准确率和召回率，使用 F 值^①来衡量 RAKE 的效果。RAKE 一共提取了 9 个关键词，其中有 6 个是正确的。

① F 值 (F -measure) 是信息检索领域的一种系统性能测试指标，它是综合召回率和准确率的一种系统评价指标。——编辑注

这就是说，在 RAKE 提取的 9 个关键词中有 6 个关键词与人工提取的关键词是相同的。虽然其中的 natural numbers 和人工提取的 set of natural numbers 非常相似，但是为了更加准确，我们把它当成错误的对待。所以提取出的关键词中有三个是错的，我们可以得到 RAKE 的准确率为 67%。把提取出的 6 个正确的关键词与人工提取的 7 个关键词进行对比，就可以得到召回率为 86%，我们计算准确率与召回率的加权平均数就可以得到 F 值为 75%。

表 1.1 RAKE 提取出的关键词与人工标记的关键词对比

RAKE 提取出的关键词	人工标记的关键词
minimal generating sets	minimal generating sets
linear diophantine equations	linear Diophantine equations
minimal supporting set	
minimal set	
linear constraints	linear constraints
natural numbers	
strict inequations	strict inequations
nonstrict inequations	nonstrict inequations
upper bounds	upper bounds
	set of natural numbers

1.3 基准评估

为了评估算法的性能，我们将 RAKE 与在 Hulth (2003)、Mihalcea 和 Tarau (2004) 的报告中进行关键词提取实验的技术摘要集合做对比，主要是为了使它们的结果可以进行直接比较。

1.3.1 准确率和召回率评估

这组技术摘要包括 2000 篇来自计算机科学与信息技术期刊论文的文摘摘要。这些摘要被划分成：1000 篇摘要的训练集、500 篇摘要的验证集和 500 篇摘要的测试集。我们仿照了 Mihalcea 和 Tarau (2004) 中描述的方法，因为 RAKE 不需要训练集，所以就测试集作评估。从每篇摘要中抽取出来的关键词被用来与人工标记的不受控关键词的关联集合中的内容进行比较。

表 1.2 描述了 RAKE 使用 Fox (1989) 的停用词列表，且 T 取图中单词数目的三分之一时的性能。对于每一种方法，对应表中每一行所显示的信息如下：提取关键词的总数量和平均每一篇摘要中的关键词数量；提取正确关键词的总数量和平均每一篇摘要中正确关键词的数量；准确率；召回率； F 值。在 Hulth (2003)、Mihalcea 和 Tarau (2004) 上公布的结果也在比较的范围内。准确率、召回率和 F 值的最高值在表中加粗显示。因为人工标记的关键词并不一定总是出现在摘要中，

所以使用任何技术都不可能达到 100% 的准确率。如果 RAKE 方法使用了基于邻接关键词生成的停用词列表, 则可以获得最高的准确率和 F 值。这组邻接关键词是如图 1.6 所示的列表的一个子集。从 F 值和准确率的角度来看, RAKE 对这组停用词列表产生了最好的结果, 并且有一个相似的召回率。在使用 Fox 的停用词列表时, RAKE 可获得一个高的召回率, 但是准确率会有所降低。

表 1.2 用 RAKE、TextRank [Mihalcea 和 Tarau (2004)] 和监督学习 [Hulth (2003)] 的方法在技术摘要测试集的 500 篇摘要中自动提取关键词的结果

方法	提取的关键词		正确的关键词		准确率/(%)	召回率/(%)	F 值
	总量	均量	总量	均量			
RAKE ($T=0.33$)							
KA 停用词列表 ($df>10$)	6052	12.1	2037	4.1	33.7	41.5	37.2
Fox 停用词列表	7893	15.8	2054	4.2	26	42.2	32.1
TextRank							
Undirected, co-occ. window = 2	6784	13.6	2116	4.2	31.2	43.1	36.2
Undirected, co-occ. window = 3	6715	13.4	1897	3.8	28.2	38.6	32.6
(Hulth 2003)							
Ngram with tag	7815	15.6	1973	3.9	25.2	51.7	33.9
NP chunks with tag	4788	9.6	1421	2.8	29.7	37.2	33
Pattern with tag	7012	14	1523	3	21.7	39.9	28.1

the, and, of, a, in, is, for, to, we, this, are, with, as, on, it, an, that, which, by, using, can, paper, from, be, based, has, was, have, or, at, such, also, but, results, proposed, show, new, these, used, however, our, were, when, one, not, two, study, present, its, sub, both, then, been, they, all, presented, if, each, approach, where, may, some, more, use, between, into, 1, under, while, over, many, through, addition, well, first, will, there, propose, than, their, 2, most, sup, developed, particular, provides, including, other, how, without, during, article, application, only, called, what, since, order, experimental, any

图 1.6 生成的停用词列表中的排名前 100 的词汇

1.3.2 效率评估

由于人们对大型数据中心能源节约的兴趣在逐步增加, 因此我们也评估了使用 RAKE 和 TextRank 方法的相关计算成本。TextRank 将语法的过滤器应用到文档文本中以标记文中的实词, 并且计算一个大小为 2 的窗口的词共现图。图中每个单词的排名都是通过一系列使它收敛于临界值的迭代计算而得到的。

我们设置 TextRank 的阻尼因数 $d=0.85$, 收敛的临界值为 0.0001。我们无法