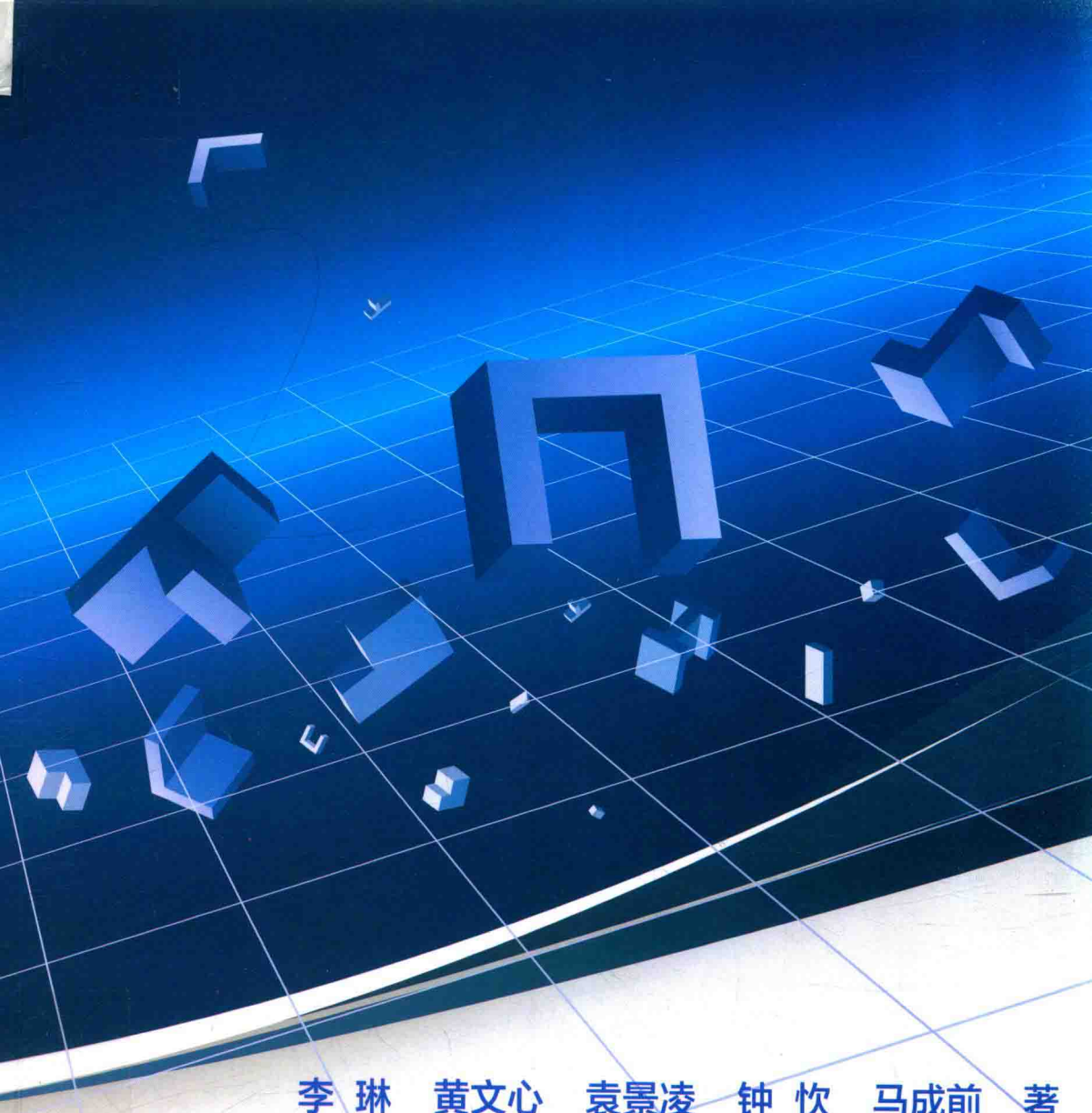

李琳 黄文心 袁景凌 钟忺 马成前 著

Web大数据的分析与推荐方法

Analysis and Recommendation Method for Web Big Data

科学出版社

Web 大数据的分析与推荐方法

Analysis and Recommendation Method for Web Big Data

李 琳 黄文心 袁景凌 钟 忺 马成前 著

科学出版社
北京

内 容 简 介

本书利用当前最热门的社交网络媒体微博等进行大数据文本分析，并在此基础上，提出基于文本分析的推荐方法，多层级推荐方法，融合评分矩阵的推荐方法，基于社团聚类的推荐方法，基于用户点击行为的混合推荐方法，融合隐性特征的群组推荐方法，分布式群组推荐方法。同时给出一种 Web 查询词推荐服务，让用户更精确地查找并定位到所要搜索的相关网页。

本书可供 IT 领域硕士、博士研究生，大数据分析处理工程技术人员阅读参考。

图书在版编目(CIP)数据

Web 大数据的分析与推荐方法/李琳等著. —北京：科学出版社，2018.5

ISBN 978-7-03-057272-1

Ⅰ.①W… Ⅱ.①李… Ⅲ.①数据处理 Ⅳ.①TP274

中国版本图书馆 CIP 数据核字(2018)第 082438 号

责任编辑：杜 权 / 责任校对：董艳辉

责任印制：彭 超 / 封面设计：苏 波

科学出版社 出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

武汉中科兴业印务有限公司印刷

科学出版社发行 各地新华书店经销

*

开本：B5(720×1000)

2018 年 5 月第 一 版 印张：12 1/4

2018 年 5 月第一次印刷 字数：250 000

定价：70.00 元

(如有印装质量问题，我社负责调换)

前　言

近些年来，随着信息技术的不断发展和 Internet 的推广应用，产生了海量信息。成千的电影、上万的书籍、亿万的网页已经成为内容丰富的大数据[1]。网络产品及应用程序产生的 Web 大数据涉及生产与生活的各个领域。Web 大数据，不仅数据规模大、数据形式多样、数据更新快，而且具有丰富的数据价值[2]。通过数据挖掘方法发现数据呈现的模式后，可以针对性地为生产与生活提供更合适的服务。然而，Web 大数据信息的这些特点同样会造成严重的信息过载，以至于用户无法直接接触到自身最需要的信息。因此，在大数据时代，各领域及行业应用平台迫切需要根据用户需要，定制化地展现更合适的信息，为用户提供更便捷有效的信息获取方式[3]。目前，解决信息过载问题主要有两类方法：搜索引擎和推荐系统。推荐系统凭借其充分挖掘大数据信息的能力，即使在用户没有明确目的的情况下，也能发现用户感兴趣或以后会感兴趣的信息[4]。因此，为了更好地自动发现用户需要的信息，推荐系统在大数据时代得到了广泛的应用，成为解决信息过载问题强有力的工具。

推荐系统的核心是推荐算法。大数据时代，推荐算法作为最有用的数据挖掘算法之一，具有以下多个优势：①大多推荐算法的可扩展性强，能够并行化执行，因此可以处理大数据；②无须用户提供明确需求就可以解决信息过载，挖掘出用户感兴趣的甚至潜在感兴趣（推荐给用户后，用户才发现自身感兴趣）的物品[5]；③推荐平台选择合适的推荐算法后，也可便捷地根据用户的反馈，调整和改进已有推荐算法；④能够和聚类算法、分类算法、模式挖掘算法等其他数据挖掘算法结合使用，在推荐、预测、分类等多方面发挥作用[6]。

推荐系统在电子商务、电影与视频、音乐与电台、社交网络、书籍、基于位置的服务以及广告等多个领域发挥着十分重要的作用。推荐系统中最普遍的两个应用场景是以用户为核心的个性化推荐场景和以物品或对象为核心的相关推荐场景[7]。个性化推荐场景能够从用户的基本特征或行为模式中发现用户的偏好和品味，从而为不同偏好用户定制化地推荐相应物品，保证推荐平台能够满足用户对不同品味的需求[8]。相关推荐场景能够根据物品的特征和用户对物品的反应，为用户展现、查看物品的相关物品[9]。这样，在用户对某物

品感兴趣时，可以很容易找到相似或者相关的物品，避免了用户从大量信息中筛选信息的过程。推荐平台应用推荐系统的意义在于，如果用户经常从推荐系统中获得自身需要的信息，他们会逐渐信任和依靠推荐系统，并加强与推荐系统的交互，从而促进推荐系统的良性循环。

本书就是在这一大背景下产生的。本书总结作者近几年的相关研究成果，分别从原理、方法、应用及实验分析等方面进行介绍与讨论。主要包括微博大数据分析与推荐方法、Web 大数据多层级推荐方法、融合评分矩阵和评论文本的推荐方法、基于社团聚类的推荐方法、基于用户行为的混合推荐方法、融合隐性特征的群组用户推荐方法和分布式群组推荐方法等。

本书由武汉理工大学的李琳博士、袁景凌博士、钟忺博士、马成前博士和武汉大学的黄文心博士共同撰写。要感谢参考文献作者的贡献，感谢钟珞教授全面的指导与支持。

限于作者水平，不足之处在所难免，诚望读者批评指正。

作　者

2018 年 1 月于武汉

目　录

第 1 章　绪论 …… 1
第 2 章　微博大数据分析与推荐方法 …… 4
2.1　新浪微话题的媒体特征分析 …… 4
2.1.1　微博活跃度 …… 4
2.1.2　微话题的演变趋势 …… 6
2.1.3　基于 LDA 的语义抽取 …… 9
2.2　基于新鲜方面的 Web 查询词推荐服务 …… 11
2.2.1　查询词推荐流程 …… 11
2.2.2　查询词推荐算法 …… 12
2.2.3　数据集的选取与数据评估方法 …… 13
2.2.4　实验结果与分析 …… 14
第 3 章　Web 大数据多层级推荐方法 …… 21
3.1　单层级相关推荐 …… 21
3.1.1　相关推荐场景及基础算法分析 …… 21
3.1.2　基于热度融合的相关推荐 …… 24
3.1.3　实验结果与分析 …… 32
3.2　多层级相关推荐 …… 38
3.2.1　基于资源传播的相关推荐 …… 38
3.2.2　基于用户反馈的多层级相关推荐 …… 44
3.2.3　实验结果与分析 …… 49
第 4 章　融合评分矩阵和评论文本的推荐方法 …… 54
4.1　基于评分数据的矩阵分解模型 …… 54
4.1.1　传统的评分矩阵分解模型 …… 54
4.1.2　邻域影响的矩阵分解模型 …… 56
4.1.3　传统模型实验结果分析 …… 57
4.2　融合评分与评论的 HFPT 及 DLMF 算法 …… 61
4.2.1　基于评论主题偏好的 HFPT 算法 …… 62

4.2.2 融合用户偏好与商品特性的 DLMF 算法 …… 69
4.2.3 实验结果与分析 …… 76
第 5 章 基于社团聚类的推荐方法 …… 83
5.1 社团结构以及社团发现算法 …… 83
5.2 基于用户偏好聚类的社团发现算法 …… 84
5.2.1 用户兴趣偏好建模 …… 85
5.2.2 CDPC 算法流程 …… 85
5.3 基于社团聚类的兴趣偏好建模算法 …… 88
5.3.1 CDCF 算法的提出 …… 88
5.3.2 CDCF 算法流程 …… 89
5.3.3 CDCF 算法实验 …… 90
5.4 社团聚类与多源数据融合建模的兴趣点推荐算法 …… 92
5.4.1 SoGeoSco 建模过程 …… 93
5.4.2 社团聚类与多源数据融合建模的 SoGeoSco 模型 …… 98
第 6 章 基于用户行为的混合推荐方法 …… 105
6.1 基于加权的混合模型 …… 105
6.1.1 基于 SVD 的矩阵分解模型 …… 105
6.1.2 ListWise 优化后的矩阵分解模型 …… 107
6.1.3 基于用户的协同过滤模型 …… 109
6.1.4 线性加权混合 …… 110
6.2 基于 Stacking 的混合推荐 …… 111
6.2.1 初级学习器选择 …… 112
6.2.2 次级学习器选择 …… 113
6.3 实验设计与结果分析 …… 114
6.3.1 实验评价指标 …… 114
6.3.2 数据集选取与处理 …… 114
6.3.3 实验环境 …… 117
6.3.4 实验设计与分析 …… 117
6.3.5 实验结果对比 …… 123
第 7 章 融合隐性特征的群组用户推荐方法 …… 125
7.1 融合隐性特征的个人推荐算法 …… 125
7.1.1 个人推荐的常用方法 …… 125
7.1.2 SVD++推荐算法模型 …… 128

7.2　融合隐性特征的群组推荐算法 …… 131
7.2.1　群组推荐生成方式 …… 131
7.2.2　群组融合策略 …… 133
7.2.3　融合隐性特征的群组推荐算法 …… 135
7.3　群组推荐实验分析 …… 137
7.3.1　数据准备 …… 137
7.3.2　融合隐性特征的个人推荐算法实验 …… 139
7.3.3　融合隐性特征的群组推荐算法实验 …… 144
第 8 章　分布式群组推荐方法 …… 150
8.1　并行架构及算法描述 …… 150
8.1.1　LUALS-WR 算法描述 …… 150
8.1.2　基于 LU 分解的特征向量更新 …… 153
8.2　分布式矩阵分解模型求解 …… 157
8.2.1　SGD 和 ALS …… 157
8.2.2　基于 MapReduce 的分割策略 …… 160
8.2.3　实验结果与分析 …… 161
8.3　Follow 社交关系的群组推荐方法 …… 165
8.3.1　群组推荐方法 …… 166
8.3.2　偏好融合策略 …… 168
8.4　实验结果与分析 …… 172
8.4.1　实验数据预处理 …… 172
8.4.2　实验方案设计 …… 172
8.4.3　实验分析 …… 173
参考文献 …… 179

第1章 绪 论

在大数据和云计算席卷全球的当今，推荐系统和推荐算法得到了广泛而又迅速的发展。推荐系统通过对大数据进行数据挖掘和知识发现，创造了极大的价值，使推荐算法受到国内外大量专家、学者、研究员等的广泛关注。随着互联网作用的不断扩大，包括电影、音乐、电子商务等在内的各 Web 数据平台都或多或少地采用推荐算法来提高用户满意度[10-14]。在电子商务平台，Zhao 等[15]针对用户购买力设计的推荐算法和 McAuley 等[16]提出的基于物品样式的推荐算法有着异曲同工之效，都能从单因素的角度发现适合用户的物品。Reddy 等[17]将奇异值分解(singular value decomposition，SVD)技术用于音乐推荐，以资深歌手的角度为用户推荐符合其品味的音乐作品。Diao 等[18]将电影网站包含的电影信息、用户评分、用户评论等在内的 Web 大数据用于推荐系统，提出 JMARS 推荐算法来为观赏电影的用户推荐他们更感兴趣的电影。

以 QQ、微信为代表的聊天工具，以及以 Facebook、微博为代表的信息展示与交互工具使社交网络成为大众生活中密不可分的一部分。社交网络由用户、物品、用户属性、物品标签等 Web 数据元以及它们之间的相互关联组成。作为信息密集且飞速更新的代表，社交网络中的推荐系统发挥着举足轻重的作用。其最大的特点在于，传统推荐算法加入社交网络中的关联信息后，可以进行更合理的物品推荐、好友推荐以及标签推荐[19]。

虽然推荐系统已经能够在解决信息过载方面发挥重要的作用，但由于 Web 大数据的复杂性，传统推荐算法在处理 Web 大数据方面还有许多不足。因此，各种改进策略用于完善推荐算法。Li 等[20]用 SVD 矩阵分解技术改进传统推荐算法，从而帮助推荐平台缓解 Web 大数据过于稀疏的问题。于洪等[21]将时间权重信息与用户-项目-属性三分图相结合，根据用户时间权重建立用户积性模型，从而能在一定程度上缓解推荐算法中新项目的冷启动问题。Kluver 等[22]融合传统的基于用户的协同过滤(UserCF)、基于物品的协同过滤(ItemCF)和 SVD 推荐算法，完善了新用户的推荐问题。根据不同领域或场景的实际需要，不断有专家将聚类技术[23]、排序学习技术[24]、深度学习技

术[25]和逻辑回归技术[26]等数据挖掘技术方法与传统推荐算法相结合，设计更符合场景的推荐模型，以提高推荐效果。

多样化推荐结果是优化推荐算法的一个重要方向。当前的多数推荐算法是以提高推荐集合的准确性为主要目标的。但是，这样设计的推荐算法具有很大的局限性。适当提高推荐集合的多样性，不仅能够为用户提供更丰富的物品推荐，而且能够反过来提高推荐结果的准确率等其他性能，改善整体的推荐效果[27]。根据用户的反馈，改进推荐模型是另一个应用前途广泛的推荐算法改进方向。Aiolli[28]结合反馈信息，对 top-n 推荐中的物品排名进行重计算，在很大程度上提高了相关物品预测的准确性。Yi 等[29]根据用户浏览不同内容的停留时间，分析用户反馈信息，并将用户反馈信息融入协同过滤模型，从而提供令用户更加满意的推荐。Volkovs 等[30]根据用户对历史推荐结果的点击、查看、播放操作行为，计算用户对推荐结果满意与否的二元反馈值，并用于更新相似度矩阵，从而在提高推荐准确性的同时加速推荐结果的计算。

对于物品相关推荐场景，通用化推荐往往无法发挥用户特性。因此，许多专家采用基于群体的推荐算法，根据用户特征划分用户群，以群体为单位生成相应的群推荐列表。Chen 等[31]将基于群体的推荐算法用于以 Flickr 为代表的图片分享网站，与通用化推荐算法相比，其多个方面的性能有所提高。Zhang 等[32]将群推荐方法与潜在因子模型相结合，从而更好地利用相同群内用户的位置特征之间的联系，进行合适的兴趣点(point of interest，POI)推荐。尽管群推荐技术已有不错的发展，但相关推荐场景的群推荐还有很大的改善空间。

如今微博正在迅猛发展，越来越多的国内外学者开始将其作为研究和关注的焦点。各行业的科研人员在现有的社交网络基础上，对微博相关理论以及实践的开展进行了进一步研究。随着在线社交网络服务的兴起，更多的人开始研究其数据特征。国内外有大量的学术文章是关于微博的，尤其是 Twitter。Newman 等[33]对整个 Twitter 空间以及信息扩散进行了定量分析和研究。2007 年，Java 等[34]对 Twitter 进行了初步分析，数据集大约有 76 000 个用户和 1 000 000 条推文，他们发现了基于用户意图对话题的用户集群。Krishnamurthy 等[35]根据粉丝和所关注人数量之间的关系分析了用户的特征。Jansen 等[36]也对 Twitter 的口碑进行了初步分析。在 2010 年进一步讨论了 Twitter 的拓扑特征以及其作为新的信息媒介传播和分享的能力[37]。然而随着新浪微博的迅速崛起，越来越多的研究者开始着手对中文微博的媒体特征做出分析和研究。Liu 等[38]将基于翻译的方法和基于频率的方法结合

起来对关键字进行抽取，他们抽取来自新浪微博用户的关键字。

Web搜索引擎在过去的十年里极大地提高了人们获知信息的方式。当一个用户在搜索框里键入一个查询词时，大部分搜索引擎是通过查询词的历史记录提供推荐服务帮助用户给出搜索结果[39]。用户可以快速地选择一个已推荐的完整词(在某些时候，也可以直接替代)。这样，用户就不需要完整地键入整个查询词。目前的搜索引擎主要是根据用户输入的查询词在历史记录里进行查询检索，然后将其与用户查询相关的结果返回给用户，但是，在大多数情况下得到的检索结果并不能够完全准确地表达出用户的意图，尤其是对于那些网络上新鲜出炉的信息。用户如果对检索返回的结果不满意，则可进行再次搜索，输入新的查询词，直到找到最满意的结果，并将其返回。为了方便用户查询，近几年来很多商业的搜索引擎如Google、Bing、Baidu等都给出了查询词推荐以方便用户进行准确的搜索，进一步提高了搜索引擎的可用性。

国内外有很多研究工作是关于查询词推荐的。最初的工作主要集中在对当前用户的查询词去识别历史的相似查询词。Baeza-Yates等[40]在搜索日志里呈现了群集查询。给定一个初始查询词，来自群集的相似查询词将会被识别，这是基于向量相似度衡量标准之上的，然后将其推荐给用户。Baroui-Ebrahimi等[41]根据词频，统计那些出现在过去用户提交的查询词，然后将其推荐给用户。Gao等[42]描述了针对跨语言的信息检索的查询词推荐算法。最近，Broder等[43]提出了稀少查询词的在线扩充方法，Song等[44]也研究了在日志里基于稀少查询词的推荐去挖掘潜在的反馈信息。Bhatia等[45]为了从一个既不使用查询词日志，也不利用文档的语料库里抽取一些候选词作为查询词推荐，提出了一个基于概率的算法机制。

传统的一些方法主要是依据用户之前搜索过的信息，利用大量过去使用的数据去提供可能的查询词推荐。尽管有很多著作[46-53]使用的是查询日志来给出查询词推荐，但是仍然存在一些困难。目前，虽然有很多研究者在挖掘关联查询词方面进行了大量研究，但大多数都是基于结果文档以及用户查询日志的方法。而对于Web上涌现出来的新鲜内容很难理想地给出合理的关联查询词推荐，因为这些新词很少能够在用户查询日志或者结果文档中反映出来。

本书主要介绍近几年在Web大数据分析及推荐方面的一些研究成果，包括微博大数据分析与推荐方法、Web大数据多层级推荐方法、融合评分矩阵和评论文本的推荐方法、基于社团聚类的推荐方法、基于用户行为的混合推荐方法、融合隐性特征的群组用户推荐方法，以及分布式群组推荐方法。

第 2 章　微博大数据分析与推荐方法

2.1　新浪微话题的媒体特征分析

微博作为一种广泛使用的媒介平台，其多样化的特征满足了人们的信息、人际关系以及一些新需求。用户越来越注重通过一些热门的微话题来传播他们的想法和意见。针对某一个具体话题，到底有多少条微博，有多少用户参与互动，什么话题在某一段时间内很热门，人们关心最多的又是哪一个话题以及每一个话题最活跃的是什么时期等这些问题成为大众关注的热点。本章用真实的数据对这些微博媒体特征进行全面的分析[54-58]。

2.1.1　微博活跃度

2.1.1.1　用户数与微博数

2012 年 3 月底～2012 年 6 月，一共有 43 967 个用户参与到 14 个话题中，微博总数达到了 55 768 条，其中不包括重复的微博数。统计的结果如表 2.1 所示，反映了这些话题的整体分布情况。从表 2.1 可以看出，话题“电信版 iPhone4s 即将开售”的用户涨到 4921 个，微博数大约为 8038 条，但是另一个话题“领导干部专用平板电脑”，它的用户数只有 326，微博数也只有 1538 条。由这些数据可以看出人们更多地关注“电信版 iPhone4s 即将开售”这个话题，这个话题在当时的流行度很高。在我们统计数据的这段时间，“柯达申请破产保护”这个话题拥有最多的微博数和用户数，因此可以说这个话题是这 14 个话题中最热门的一个。

2.1.1.2　用户参与度与用户活跃度

为了衡量某个话题在这 14 个话题中用户的参与程度以及测试用户在微话题下的活跃度，决定分别对这 14 个话题的每一个话题进行用户参与度以及用户活跃度测试。其用户参与度以及用户活跃度的定义为

表 2.1　来自新浪微博统计结果

话题	新 iPad 香港开售	身绑 25 部 iPhone 被抓	苹果推出新一代 iPad	苹果 CEO 年薪 24 亿	苹果 App Store	CES2012	HTC 被判侵犯苹果专利
用户数/个	4 893	1 020	5 569	4 111	5 289	3 792	934
微博数/条	6 043	1 242	6 889	5 600	6 760	6 126	1 237
@微博数/条	823	169	1 415	660	1 401	1 313	146
总的微博数/条	6 874	1 419	8 313	6 324	8 175	7 453	1 385
实际微博数/条	6 866	1 411	8 304	6 320	8 161	7 439	1 383
垃圾微博数/条	8	8	9	4	4	14	2
用户参与度	0.111 3	0.023 2	0.126 7	0.093 5	0.120 3	0.086 3	0.021 2
用户活跃度	1.235 0	1.217 7	1.237 0	1.362 2	1.278 1	1.615 5	1.324 4
话题热度	0.124 6	0.025 6	0.142 0	0.115 5	0.139 4	0.126 3	0.025 5
话题活跃度	0.809 7	0.821 3	0.808 4	0.734 1	0.782 4	0.619 0	0.755 1
话题	领导干部专用平板电脑	柯达申请破产保护	华为秀出你节日新生活	电信版 iPhone4S 即将开售	Windows8 预览版发布	iOS5.0.1 完美越狱	Facebook 宣布收购
用户数/个	326	5 648	933	4 921	4 537	1 063	3 971
微博数/条	1 283	7 137	1 116	6 714	6 516	1 645	5 046
@微博数/条	255	1 931	332	1 324	607	282	590
总的微博数/条	1 556	9 084	1 448	8 085	7 125	1 932	5 639
实际微博数/条	1 538	9 068	1 448	8 038	7 123	1 927	5 636
垃圾微博数/条	18	16	0	47	2	5	3
用户参与度	0.007 4	0.128 5	0.021 2	0.111 9	0.103 2	0.024 2	0.090 3
用户活跃度	3.935 6	1.263 6	1.196 1	1.364 4	1.436 2	1.547 5	1.270 7
话题热度	0.026 5	0.147 2	0.023 0	0.138 4	0.134 4	0.033 9	0.104 0
话题活跃度	0.254 1	0.791 4	0.836 0	0.732 9	0.696 3	0.646 2	0.786 9

$$\text{UserP}=\text{UT}/U \tag{2.1}$$

$$\text{UserA}=T/\text{UT} \tag{2.2}$$

式中：UserP 为用户参与度；UT 为某个话题的用户数；U 为用户总数；UserA 为用户活跃度；T 为某话题的微博数量。在计算中并没有将转发的微博数考虑进去，而只是对某话题自身的微博进行研究。从实验结果可以看出，在这 14 个话题中，微话题“柯达申请破产保护”的用户参与度最高，达到0.1285，用

户活跃度却为 1.2636。还发现，除极少数的话题外，大部分微话题的用户活跃度相似，其中，微话题“领导干部专用平板电脑”的结果最明显，达到了 3.9356。但是对比发现，其用户参与度最低，只有 0.0074。这说明虽然相对于其他话题，参与的人数较少，但是用户的活跃程度却很高。

2.1.1.3 微话题热度与微话题活跃度

为了探讨某一个话题在这 14 个话题的热度情况，定义了微话题热度这个指标。为了测试某个话题的活跃程度，也定义了微话题活跃度这个指标，其定义为

$$\text{Hot-degree}=T/W \tag{2.3}$$

$$\text{Active-degree}=U/T \tag{2.4}$$

式中：Hot-degree 为话题热度；T 为某个话题的微博数；W 为总的微博数；Active-degree 为话题活跃度；U 为某个话题的用户数。这里也没有将转发的微博数考虑进去，而只是对某话题自身的微博进行研究。由实验结果可以看出，在这 14 个话题中，微话题“柯达申请破产保护”的话题热度最高，达到了 0.1472，这与实际的统计结果正好相符，其总的微博数在这 14 个话题中也是最高的。尽管微话题“华为秀出你节日新生活”的话题热度最低，但是对于其他微话题，其话题活跃度最高，达到了 0.8360。相当于一条微博由0.8360个用户所发，可见该话题很受关注。

2.1.2 微话题的演变趋势

一个话题从刚开始到中间直至话题结束，各个时间段到底能吸引多少条微博和多少个用户呢？什么时间是一个话题的高峰时期呢？什么时间是一个话题的低迷时期呢？带着这些问题，讨论一段时间内用户和微博的分布情况。图 2.1～图 2.14 绘制了每一个话题每隔 20 天用户数和微博数的变化情况。

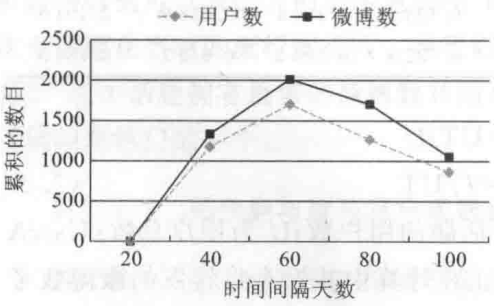

图 2.1 新 iPad 香港开售

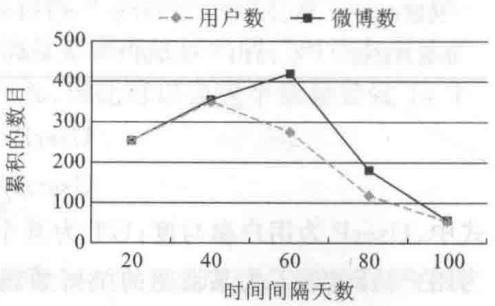

图 2.2 身绑 25 部 iPhone 被抓

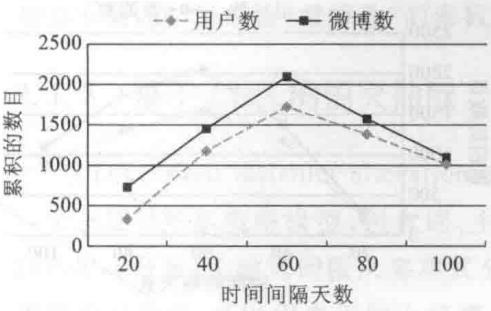

图 2.3　苹果推出新一代 iPad

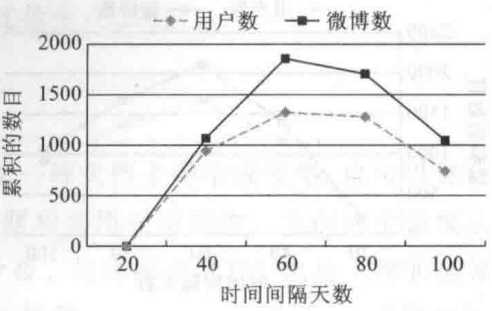

图 2.4　苹果 CEO 年薪 24 亿

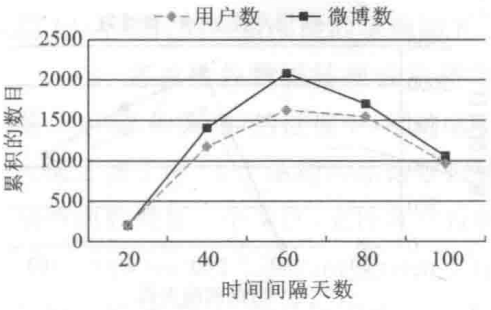

图 2.5　苹果 App Store

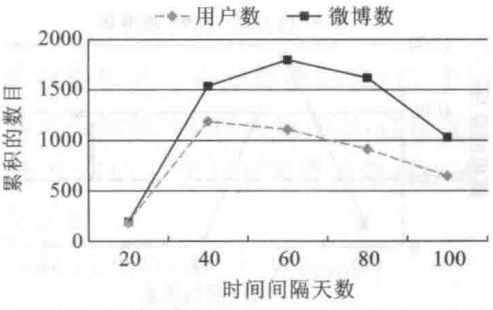

图 2.6　CES2012

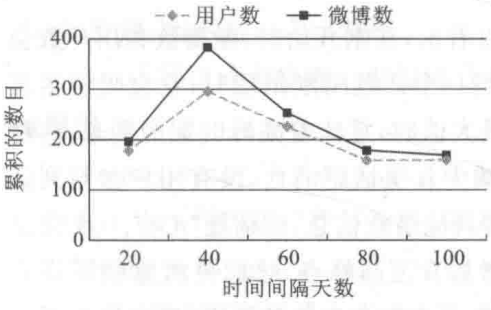

图 2.7　HTC 被判侵犯苹果专利

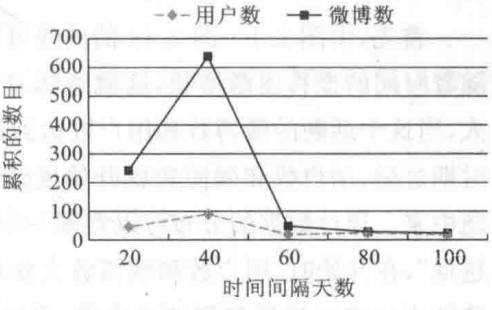

图 2.8　领导干部专用平板电脑

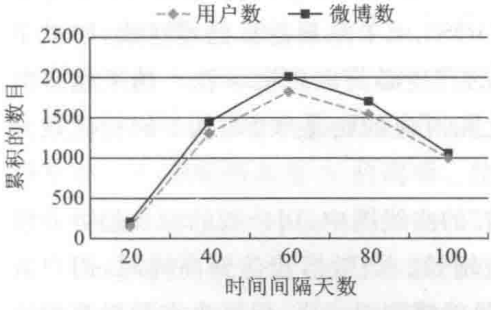

图 2.9　柯达申请破产保护

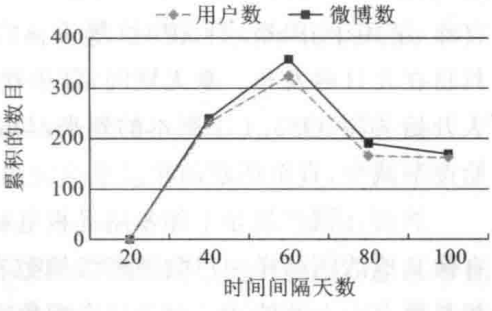

图 2.10　华为秀出你节日新生活

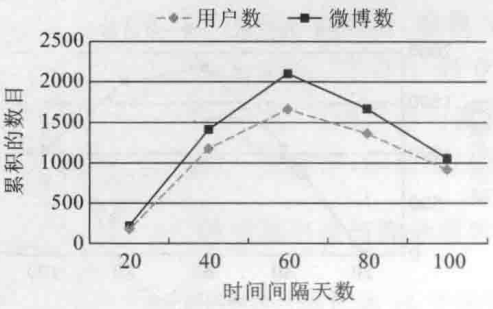

图 2.11 电信版 iPhone4s 即将开售

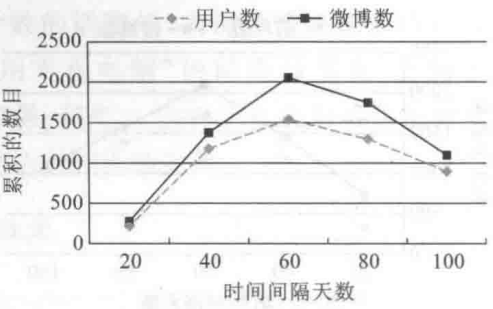

图 2.12 Windows8 预览版发布

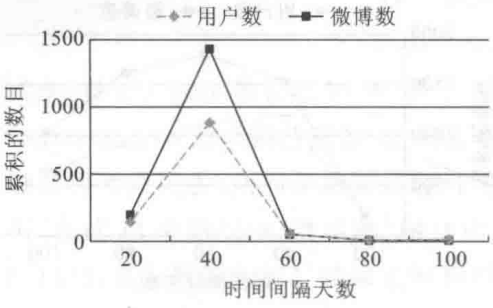

图 2.13 iOS5.0.1 完美越狱

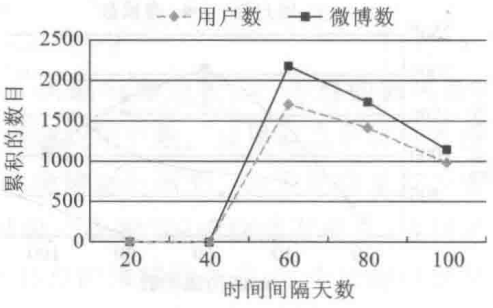

图 2.14 Facebook 宣布收购

首先，由图 2.1～图 2.14 的曲线可以看出，在刚开始时，微博数和用户数会随着时间的推移逐渐增长，这就意味着当一个话题刚刚出现时，它会吸引很多人，当这个话题的微博数和用户数达到最大值时，意味着高峰时期的到来，高峰时期过后，用户数和微博数就开始逐渐减少直至话题消亡，没有用户参与到话题中来。通过数据的分布可以发现，一些其他潜在信息，如话题“iOS5.0.1 完美越狱”，在开始时，用户数和微博数大幅增加直至高峰点，然后突然急剧下降直到结束。由于这组数据波动太大，可以猜测肯定存在某种原因，事实也证明了这一现象。2012 年 5 月 20 日，著名 iPhone 越狱黑客 Pod2g 在其 Twitter 上宣布，适用于 iPhone4s、iPad3 等设备的 iOS5.1.1 完美越狱已经达成，越狱工具将在几日内发布。毫无疑问，新事物的产生必将取代旧事物。越来越多的人开始关注 iOS5.1.1 版本的到来，与之相对应的则是 iOS5.0.1 的粉丝数开始逐渐减少，直至话题消亡。

然而，话题“领导干部专用平板电脑”的曲线图中，用户数的增长趋势并没有像其他话题那样用户数随着微博数在增长。当微博数达到高峰时，用户数却并没有太大的波动。对于这个现象，虽然感到很奇怪，但这也在预料范围之内，因为对于这个话题，它仅仅反映的是活跃用户的情况，微博的用户关注度

却并不高，其“用户数/微博数”的参数比值为 0.2541。

2.1.3　基于 LDA 的语义抽取

LDA(latent dirichlet allocation)是一种文档主题生成模型，也可以称为一个三层贝叶斯概率模型，包含词、主题和文档三层结构。文档到主题服从 Dirichlet 分布，主题到词服从多项式分布。与此同时，LDA 也是一种非监督机器学习技术，可以用来识别大规模文档集(document collection)或语料库(corpus)中潜在的主题信息。这里是用 C 语言来实现 LDA 的 EM 算法。运用 LDA-C 代码所做实验的步骤如下。

(1) 将原始的数据转换成所需要的数据格式。在分析数据集(14 个话题，74 662 条微博)的过程中，词的索引和文档的索引矩阵将被创建，这样就极大地方便了将 14 个话题的原始数据转换为 LDA-C 实现所需要的数据格式。最终的数据是一个文件，文件每一行的格式如下：

[M][term_1]:[count][term_2]:[count]...[term_N]:[count]

其中，[M]是指 14 个话题中某一个话题的不同词项数目；[count]是指在这个话题中某个词项出现的次数。需要注意的是[term_1]代表的是一个词项的索引，是整数，而不是一个字符串。

(2) 话题估计和推理。在编译代码之后，通过执行如下命令就可以估计出模型：

lda est[alpha][k][settings][data][random/seeded/ *][directory]

对一组不同的数据执行推理，需要执行以下命令：

lda inf[settings][model][data][name]

在此，需要强调的是变分推理所用的数据是估计模型产生的数据。在变分推理结束之后，会产生一个[name]. gamma 的文件，它是每一个话题的变分 Dirichlet 参数，接下来就会生成一个以. beta 结尾的文件，里面显示的是每一个话题的前 N 个词项，这个文件需要打印出来。

(3) 打印话题。在这里使用的是 python 脚本 topics. py 来打印. beta 文件中每一个话题排名前 N 的词项。使用的命令如下：

python topics. py〈betafile〉〈vocabfile〉〈nwords〉

在实验中，通常将 alpha 参数的值设为 1，参数 k 代表的是根据这 14 个话题抽取和估计出来的 k 个最热门的话题。然后，通过采取不同的 k 值，就能看到这 k 个热门话题在 14 个话题下的概率分布情况。最后，选取概率分布大于