Computer Science and Application 计算机科学与应用, 2016, 6(9), 531-538
Published Online September 2016 in Hans. http://www.hanspub.org/journal/csa
http://dx.doi.org/10.12677/csa.2016.69066

Hans汉斯

Weibo Recommendation System Based on Semantic Analysis

Mengdi Zhai, Sipei Wu, Yanjuan Liu

Yunnan University of Finance and Economics, Kunming Yunnan
Email: 540532394@qq.com

Received: Aug. 10th, 2016; accepted: Aug. 25th, 2016; published: Aug. 30th, 2016

Abstract

In this paper, we get Weibo's theme by semantic analysis. In order to make the subject more accuracy, we create the dictionary of Weibo. Then, we use the method of word segmentation, keyword extraction to analysis the text in advance, and we build Weibo recommendation system based on association rules, recommend the theme that users are interested in more accurately.

Keywords

Dictionary of Weibo, Cut Words, Key Words, Association Rules, Recommendation System, Weibo

基于语义分析的微博推荐系统

翟梦迪，吴思霈，刘雁娟

云南财经大学，云南 昆明
Email: 540532394@qq.com

收稿日期：2016年8月10日；录用日期：2016年8月25日；发布日期：2016年8月30日

摘 要

本文基于语义分析对博文内容进行主题提取，为使主题词更加准确构造的微博词典，通过分词、关键词

文章引用: 翟梦迪, 吴思霈, 刘雁娟. 基于语义分析的微博推荐系统[J]. 计算机科学与应用, 2016, 6(9): 531-538.
http://dx.doi.org/10.12677/csa.2016.69066

提取对文本进行预处理，进而构建基于关联规则的微博推荐系统，为用户所感兴趣的内容进行精准推荐。

关键词

微博词典，分词，特征词，关联规则，推荐系统，微博

1. 研究背景

近年来，微博一直呈现飞速发展的态势，深受国内外网民的追捧，已成为了一种极具影响力的新型媒体。同时作为国内最受欢迎的社交平台，微博每天产生着数以万计的文字、图片、视频、语音信息，这些海量的微博数据必然包含丰富的知识资源与商业价值，因此，如何在海量的微博数据中挖掘并推荐用户感兴趣的内容，是值得深入的议题。

为了达到微博用户感兴趣内容的有效实时推荐，有必要对微博内容进行文本分析，以获取用户兴趣集。本文基于语义分析利用多种模型构建推荐系统，进而为用户所感兴趣的内容提供精准推荐。

2. 国内外文献综述

目前，很多学者在分析微博数据方面，都涉及到了微博内容的文本分析。有一些学者利用微博里面固有的标签信息来进行用户兴趣模型构建，但是有很多的用户在注册微博时不会填写自己的兴趣标签，因而此法具有很大的缺陷。王宁宁[1]等人基于用户标签来为用户推荐微博内容，首先利用 TextRank 提取用户已发微博中的关键词，用以代表用户的兴趣标签，接着，根据微博效应函数与生命周期构建待推微博列表，通过计算用户标签在待推微博列表中的次数，将出现次数排在前 N-Top 的微博推荐给用户。马慧芳等人[2]提出了基于多标签关联关系的微博推荐算法。此法通过挖掘被同一用户标注的多标签的内在关联以及被不同用户标注的多标签外在关联来构建用户的兴趣集，并对用户兴趣集实时更新。彭泽环等[3]在总结影响用户微博兴趣的基础上，应用潜在因素模型提出了社区热点微博推荐系统。王晟等[4]提出了基于贝叶斯个性化排序的微博推荐算法，克服了传统推荐算法在处理用户活跃程度低、数据稀疏和兴趣动态变化等特点的缺点。李凌云等[5]基于异常检测算法、地理位置定位算法、相关事件推荐算法和事件相关度算法，实现了微博事件实时监测系统。徐雅斌等[6]采用支持向量机进行文本分类，发现用户的兴趣偏好，采用改进的协同过滤算法，结合用户兴趣偏好和推荐信任域来为用户推荐微博。

但是，在微博的日常使用中，最能表达用户兴趣特点的便是其博文内容，所以本文基于对博文的分析，构建推荐系统，并且构造微博专用词典，以增加关键词获取的准确率。

3. 研究内容与研究框架

本文将对语义分析和推荐系统的理论知识进行系统描述，在良好理解的基础之上，以 2014 年新浪微博部分用户的博文内容为训练文本集，将当前较热的语义分析技术应用其中，分析出博文所涉及的主题并进行概括，利用概括的主题进而构建推荐系统。

首先，在已有文本数据的基础之上构建微博专用词典；接下来是数据预处理阶段，进行分词、去停用词、词性标注、名词和动词提取、特征词提取；最后，利用整理出的每个用户感兴趣的领域，进而构建基于 Apriori 关联规则的推荐系统，对用户所关注的话题进行精准推荐。

4. 理论介绍

4.1. 中文分词

Python 中的 jieba 模块是目前较为常用的中文分词方法之一，分词效果较好，也可以进行词性标注、

关键词提取等等工作，其分词思路如下：

1) 加载模块内字典 dict.txt，词典中包括以人民日报为语料库训练得到的词和对应词频；

2) 从内存的词典中构建该句子的 DAG(有向无环图)，这一步是为了给出一句话中所有可能的词的划分；

3) 对于词典中未收录的词，使用 HMM 模型的 viterbi 算法尝试分词处理，该算法中设给定词串 $W=w1,w2,\cdots,wk$，$Si(i=1,2,\cdots,N)$ 表示词性状态(共有 N 种取值，其中 N 为词性符号的总数,可以通过语料库统计出来)，$t=1,2,\cdots,k$ 表示词的序号(对应 HMM 中的时间变量)，Viterbi 变量 $v(i,t)$ 表示从 $w1$ 的词性标记集合到 wt 的词性标记为 Si 的最佳路径概率值，存在的递归关系是 $v(i,t)=\max\left[v(i,t-1)*aij\right]*bj(wt)$，其中 $1\le t\le k$，$1\le i\le N$，$1\le j\le N$，aij 表示词性 Si 到词性 Sj 的转移概率，对应上述 $P(ti|ti-1)$，$bj(wt)$ 表示 wt 被标注为词性 Sj 的概率，即 HMM 中的发射概率，对应上述 $P(wi|ti)$，这两种概率值均可以由语料库计算。每次选择概率最大的路径往下搜索，最后得到一个最大的概率值，再回溯，因此需要另一个变量用于记录到达 Si 的最大概率路径；

4) 已经收录词和未收录词全部分词完毕后，使用 dp 寻找 DAG 的最大概率路径，dp 即为动态规划，动态规划算法通常用于求解具有某种最优性质的问题[7]；

5) 输出分词的结果。

4.2. 去停用词

构建停用词词典，其中包括“的”，“你”，“那么”等等与主题表现无关的词汇，利用代码实现过滤，只保留下与主题关联较大的词，以增加关键词提取的准确度。

4.3. 关键词提取

TF-IDF 是微博文本提取关键词，信息检索领域的成熟技术，TF 代表词频，表示词语在文本中出现的频率，词频描述某个词在一篇微博中出现的频数，考虑到文档的长度，为了防止较长的文本得到更高的相关度权值，因此，要对文档的词语进行归一化，在这里就直接计算词语在文本分词并降噪后的出现频率作为归一化处理，即：

$$TF(i,j)=\frac{n(i,j)}{N}$$

其中，表示词语 i 在文挡 j 中出现的频数，表示文档 j 中词语的总数。

IDF 代表反文档频率，旨在降低所有文档中都会出现的关键词的权重，其思想是，哪些常见的词语对区分文本没有作用，应该给仅出现在某些文档中的词语赋予更高的权重，例如，电信用户的投诉文本，每一个文档中都会出现“用户反馈”这句话，因此“用户”与“反馈”出现的频率就会很大，但实际上这两个词并不是文本的关键信息，并会为提取关键信息造成干扰，因此，通过反文档频率可以降低这类词汇对语义分析的干扰。设 D 为所有可能被推荐的文档的数量，d 为 N 中的某词语在所有文档 D 中出现的次数，一般计算为：

$$IDF(i)=\log\frac{D}{d(i)}$$

而在文档中关键词应被赋予的权重 TF-IDF 是这两个频率的乘积：

$$TF-IDF=TF(i,j)\times IDF(i)=\frac{n(i,j)}{N_j}\times\log\frac{D}{d(i)}$$

此公式表示，某一特定文件内的高频词语频率，以及该词语在整个文件集合中的低频率文件，可以

产生出高权重的 TF-IDF。TF-IDF 倾向于过滤掉常见的词语，进而选择出具有代表性的关键词[8]。

5. 实证分析

5.1. 数据来源及获取

本文利用已有的在 2014-05-03 至 2014-05-11 采集的 50,168 条微博博文内容作为训练数据，在此基础之上，进行关键词提取，并达到最终构建基于关联规则的推荐系统的目标。

5.2. 数据预处理

微博博文内容存在很大的随意性、缩略性，并且其中涉及的词汇在 jieba 自带的字典中并不一定存在，例如对“易烊千玺”、“张继科”、“幻城”这样的词汇无法很好的进行分词。所以，人工构建微博专用词典 weibodict.txt。在分词的过程中，将微博专用词典加入，以增加分词的准确性。部分分词结果如表 1。

由表 1 可知，按照词频降序排列得到的结果噪音很大，“的”“了”“是”这种词对于主题的获取意义不大，但出现次数较多；而“小米”“同桌的你”“房价”这种词对于反映用户兴趣是十分重要的，应突出其关键位置，所以需要进行下一步去停用词的处理。并且经前人分析可知，在一段话中最能代表主题含义的是名词和动词。所以，对结果进行需再一次加工，只提取出于主题相关的名词和动词。处理结果如表 2。

接下来，基于以上结果利用 TF-IDF 算法进行全文特征词提取，选取 15 个特征词，确定博文所涉及主题。利用词云的形式进行展示如图 1。

我们可以总结出，所采集的全部文本涉及的主题分别是房价、魅族、小米、火箭队、林书豪、恒大、韩剧、雾霾、同桌的你、公务员、贪官、转基因。当然我们还需要知道，每一个用户 ID 所涉及的主题内容，接下来，对每一个用户 ID 进行上述预处理，并进行关键词提取，部分结果如表 3。

利用上述结果，接下来便可以进行推荐系统的构造。

5.3. 推荐系统

5.3.1. 协同过滤推荐算法

上一章节，基于关键词把博文分成了 12 类，在本章中，首先计算用户发表博文类别的频率作为用户

Table 1. Word frequency table of weibo
表 1. 微博词频表

词汇	出现次数
的	182,815
了	65,653
是	43,093
不	18,657
小米	18,285
同桌的你	12,601
房价	10,924
我们	10,770
公务员	10,032
...	...

Table 2. Word frequency table
表 2. 词频表

词汇	出现次数	词性
小米	18,285	n
同桌的你	12,601	n
中国	12,404	ns
房价	10,924	n
公务员	10,032	n
转基因	8894	n
小米	18,285	n
同桌的你	12,601	n
中国	12,404	ns
...	...	...

Table 3. System resulting data of standard experiment
表 3. 用户 ID 及其对应主题词

用户 ID	主题词
1750763123	同桌的你
3477580923	恒大
3962426507	恒大
3811084974	贪官
2091273435	公务员
1974576991	公务员
3259396652	小米
3342224310	小米
2211601325	林书豪
...	...

Table 4. The favorite degree of user to examples
表 4. 用户对类别喜爱程度样例

	类别 1	类别 2	类别 3	类别 4	类别 5
用户 1	0.5	0.2	0.1	0.15	0.15
用户 2	0.1	0.7	0	0.2	0
用户 3	0	0.15	0.2	0.25	0.4
用户 4	0.1	0.3	0.2	0	0.6

Figure 1. Word Cloud
图 1. 词云图

对该类别的喜爱程度，如表 4，在用户 1 发表的微博中，类别 1 的微博占到了 50%，那么该用户对类别 1 的喜爱程度为 0.5，而用户 3 并没有发表过关于类别 1 的博文，那么他对类别 1 的喜爱程度是 0。

根据用户对各类别的喜好程度，向用户推荐数据存储库中用户感兴趣类别的相关微博。另外，基于用户本身对于物品的喜爱程度，本文根据余弦公式找出当前用户的相似用户，然后将相似用户喜欢的物品推荐给当前用户。计算上来说就是把一个用户对当前所有物品的评分作为一个输入向量，计算所有用户之间的相似度，从而找出与该用户相似的用户。根据这个相似用户对物品的喜爱程度来预测当前用户对没有评分的物品的喜爱程度。如图所示，用户喜欢物品和物品，用户喜欢物品、物品和物品，计算出用户与用户相似，就把用户喜欢的物品推荐给用户。

应用协同过滤的优点在于该技术并不需要知道关于要推荐的物品的任何信息，避免了系统提供详细信息且实时更新的物品描述信息，节省了成本。但是另一方面，如果想根据物品的特性和用户的特殊偏好非常直观的选择推荐物品，单纯的用协同过滤方法是不可能实现的，因此，下面我们将采用基于内容的推荐算法进一步为用户推荐。

5.3.2. 基于关联规则的推荐算法

一些文献会做出基于物品(微博内容)的推荐算法，但是在本文中考虑到用户发表一篇微博并不一定代表他同样希望关注相似的微博，因此，文本对于微博的推荐主要采取基于关联规则的推荐算法，挖掘微博主题之间的关联性，分析用户的潜在兴趣。关联规则中涉及两个重要的概念：1) 最小支持度，级事件 A 和事件 B 同时发生的频率；2) 最小置信度：即事件 A、B 同时发生的频率与事件 A 发生的频率之比[9]。

该算法具体示例如下，如表 5~8 所示，五个用户分别对 5 类微博感兴趣，每个用户的兴趣集合形成了各个项集，计算出项集中的各个项的支持度形成一阶候选项集，假如设置该算法的阈值为 3，大于该阈值的频繁项集被选出，小于该阈值的频繁项集被淘汰。那么 B、C、E 三个类将形成一阶频繁项集，而要研究哪些兴趣类别可能会同时出现则依照阈值选取出了二阶频繁项集，即兴趣类别 B、C，B、E，C、E 会被同时关注。同理，三阶频繁项集为 B、C、E，在此不再过多说明。

本文各类别的支持度、规则支持度、可部署性如表 9。

从结果分析可以得出，关注“小米”的用户占 25.2%，其中会有 14.7%的用户同时关注“魅族”，那么还未关注的用户占 10.5%，此项关联所占用户群体比例较大，14.7%的规则关联度也比较高，剩下的 10.5%的可部署性的用户推荐是合理、可行的；从“房价”与“贪官”这一项集得到的指标显示，该项集的关联性较强，具有比较大的推荐价值。而“房地产”与“林书豪”这个推荐组合虽然支持度与其项集的可部署性较高，但是共同关注这两个主题的用户较少，因此，没有太大的推荐价值。

本章基于前面章节做出的文本分析介绍了两种适用的推荐算法，分别是基于物品的推荐算法与基于关联规则的推荐算法，并做出了分析。

Table 5. User’s interest set
表 5. 用户兴趣集

ID	项集
001	A C D
002	B C E
003	A B C E
004	B E
005	B C

Table 6. The first-order candidates set
表 6. 一阶候选项集

ID	项集
001	A C D
002	B C E
003	A B C E
004	B E
005	B C

Table 7. The first-order candidates set
表 7. 一阶大项集

项集	支持度
{B}	4
{C}	4
{E}	3

Table 8. The second-order candidates set
表 8. 二阶候选项集

项集	支持度
{B C}	3
{B E}	3
{C E}	2

Table 9. The index of recommended results
表 9. 推荐结果指标

前项	后项	支持度	规则支持度	可部署性
小米	魅族	30.2%	14.7%	10.5%
房地产	林书豪	30.4%	5.3%	25.1%
雾霾	房地产	15.3%	17.0%	7.3%
房价	贪官	30.4%	19.4%	11.0%
小米	公务员	30.2%	15.5%	14.7%
…	…	…	…	…

6. 总结

本文做出了基于微博用户文本语义分析向微博用户进行推荐的研究，文中仅利用用户的微博文本信息建立用户兴趣模型，减小了用户个人信息的获取难度与用户个人信息的不完整性造成的计算难度，旨在增强用户体验、提高用户满意度，创造收益。

本文首先对 4237 个用户的共 50,168 条微博文本进行了分词、基于 TF-IDF 算法提取各个微博文本的关键词形成用户兴趣主题。在第三章节中之后，根据已经生成的用户兴趣主题做出了基于协同过滤的推荐算法向用户推荐兴趣相投的好友、基于微博内容的推荐算法向用户推荐与自己所发表的微博内容相似的微博、基于关联规则的推荐算法找出关联性强的兴趣集合挖掘用户的潜在兴趣，引导用户去关注其他主题，从而可以挖掘新客户群体。最终，本文以一种纯文本的挖掘方法对 40%的用户群进行了有效的推荐。

基金项目

本文为云南财经大学研究生创新基金项目：基于语义分析的微博推荐系统(2016~2017)。

参考文献 (References)

[1] 王宁宁, 鲁燃, 王智昊, 刘承运. 基于用户标签的微博推荐算法[J]. 计算机应用研究, 2017(1).

[2] 马慧芳, 贾美惠子, 李晓红, 鲁小勇. 一种基于标签关联关系的微博推荐方法[J]. 计算机工程, 2016, 42(4): 197-201.

[3] 彭泽环, 孙乐, 韩先培, 陈波. 社区热点微博推荐研究[J]. 计算机研究与发展, 2015, 52(5): 1014-1021.

[4] 王晟, 王子琪, 张铭. 个性化微博推荐算法[J]. 计算机科学与探索, 2012, 6(10): 895-902.

[5] 李凌云, 敖吉, 乔治, 李剑. 基于微博的安全事件实时监测框架研究[J]. 信息网络安全, 2015(1): 16-23.

[6] 徐雅斌, 刘超, 武装. 基于用户兴趣和推荐信任域的微博推荐[J]. 电信科学, 2015, 31(1): 7-14.

[7] 百度百科, http://baike.baidu.com/

[8] Jannach, D., Zanker, M., Felfernig, A. and Friedrich, G. (2013) 推荐系统[M]. 北京: 人民邮电出版社.

[9] 吴喜之. 复杂数据统计方法——基于 R 的应用[M]. 北京: 中国人民大学出版社.