利用概率主题模型的微博热点话题发现方法①

米文丽[1], 孙曰昕[2]

[1](陇东学院 信息工程学院, 庆阳 745000)
[2](西北师范大学 计算机科学与工程学院, 兰州 730070)

摘 要: 微博具有长度短、实时传播、结构复杂以及变形词多等特点, 传统的向量空间模型(VSM)文本表示方法和隐含语义分析(LSA)无法很好的对其进行建模. 提出了一种基于概率潜在语义分析(pLSA)和 K 均值聚类(Kmeans)的二阶段聚类算法, 此外通过定义微博热度分析和排序, 有效地支持微博热点话题发现. 实验表明, 此方法能有效地进行话题聚类并检测出热点话题.

关键词: 概率潜在语义分析; 话题发现; 微博; Kmeans

Microblog Hot Topics Discovery Method Based on Probabilistic Topic Model

MI Wen-Li[1] , SUN Yue-Xin[2]

[1](College of Information Engineering, Longdong University, Qingyang 745000, China)
[2](College of Computer Science & Engineering, Northwest Normal University, Lanzhou 730070, China)

Abstract: Microblog has the characteristic of short length, complex structure and words deformation. Therefore, traditional vector space model (VSM) and latent semantic analysis (LSA) are not suitable for modeling them. In this paper, a two stage clustering algorithm based on probabilistic latent semantic analysis (pLSA) and Kmeans clustering (Kmeans) is proposed. Besides, this paper also presents the definition of popularity and mechanism of sorting the topics. Experiments show that our method can effectively cluster topics and be applied to microblog hot topic detection.

Key words: probabilistic latent semantic analysis; topic detection; microblog; Kmeans

近年来,在互联网上蓬勃发展的微博客(微博)越来越多地引起了人们的关注. 微博从传统的社交网络中脱胎而出,在拥有了独立的服务平台后逐渐演化为一种新的信息发布形式.

然而, 微博数据主要由普通用户产生, 无论是用词、形式还是具体内容的质量都参差不齐, 给话题发现带来很大困难. 目前话题发现研究主要集中在新闻类数据上, 社会网络上(含微博)话题检测的研究相对较少. 大多数专家和学者都在“Twitter”英文微博数据进行了相关研究, 如 Pal 等人提出一种寻找 Twitter 网络中特定话题的关键人物的算法[1]; 文献[2-3]在大规模 Twitter 数据集上, 用 LDA(Latent Dirichlet Allocation)模型来建模挖掘话题; Ramage 等[4]构造了一个半监督学习模型 L-LDA 将用户和 Twitter 特性化来个性化用户信息需求; Teevan[5]等人, 通过分析大量的 Twitter 上的检索日志和传统搜索引擎上的检索日志, 对微博上的搜索和传统的 Web 搜索做了一个完善而全面的对比, 发现 Twitter 用户倾向于去搜索时间相关的信息, 比如爆炸性的新闻和一些当前的流行趋势; Neil[6]认为 Twitter 是对整个社会事实的反应,可以从中窥探社会这个庞大的机体,同时作者通过一个清晰的结构图展示了 Twitter 上帖子的互动、转发和话题的转换; 日本学者 Takeshi 等人[7]通过日本地震相关微博进行语义分析和位置检测而建立了一套地震报告系统, 此系统将微博的及时性作为区别于其他社会化媒体的重要特征. 相对于英文微博, 在中文微博研究方面相关文献较少, 孙胜平[8]提出了基于 SP&HA 聚类的微博客话题检测方法,利用 VSM 建模后衡量文本相似度, 最后用层次聚类算法实现话题检测; 郑斐然[9]采用向量空间模型在线检测中文微博消息中的关键字, 并对

①收稿时间:2013-12-18;收到修改稿时间:2014-01-14

其聚类来找到新闻话题,但传统的向量空间模型(Vector Space Model, VSM)[10]忽视了中文的“同义”、“多义”及高维向量问题, 因而在微博话题发现过程中, 其发现话题的准确率和速度不尽如人意. 针对中文微博文本的特点, 以及传统VSM模型匹配特征词的局限性, 有学者采用隐含语义分析(Latent Semantic Analysis, LSA)对中文微博建模来发现热点话题[11], 但是 LSA 中的 SVD 方法会将词文档矩阵分解为含有负数的近似矩阵, 这于词文档分布明显不符.

本文提出了一种两阶段聚类算法来支持微博热点话题发现. 首先基于 pLSA(probabilistic Latent Semantic Analysis)对微博进行主题建模, 然后通过 Kmeans 聚类来产生话题. 此外通过话题热度的定义与排序来有效地发现当前热点.

1 理论基础

pLSA 基本思想

概率潜在语义分析与隐含语义分析的不同是: 后者是以共现表(就是共现的矩阵)的奇异值分解的形式表现的, 而前者则是基于多项式分布和条件分布的混合分布来建模共现的概率. 建模过程就是对词和文档同时进行处理.

如前所述, pLSA[12]引入了一个隐含的主题 Z, 通过期望最大化的方法估计在给定文档的情况下主题的分布和在给定主题的情况下词项的分布来建立主题模型.

在 PLSA 中, 使用最大似然估计来训练相关参数. 最大似然估计中比较常用的算法是期望最大化算法. 期望最大化算法[13,14]分为两步:

1) Expectation Step——隐含参数的估计

E 步需要估计的参数是:

$$P(z_k|d_i,w_j)=\frac{P(w_j|z_k)P(z_k|d_i)}{\sum_{k=1}^{K}P(w_j|z_k)P(z_k|d_i)} \quad (1)$$

2) Maximization Step——确定实际参数, 然后根据实际参数做最大似然估计.

M 步需要估计的参数是:

$$P(w_j|z_k)=\frac{\sum_{i=1}^{M}n(d_i,w_j)P(z_k|d_i,w_j)}{\sum_{j=1}^{M}\sum_{i=1}^{M}n(d_i,w_j)P(z_k|d_i,w_j)} \quad (2)$$

$$P(z_k|d_i)=\frac{\sum_{j=1}^{M}n(d_i,w_j)P(z_k|d_i,w_j)}{n(d_i)} \quad (3)$$

通过将估计出的参数带入 log 似然函数近似下界求值, 似然函数形式如下:

$$E[L]=\sum_{i=1}^{M}\sum_{j=1}^{M}n(d_i,w_j)\sum_{k=1}^{K}P(z_k|d_i,w_j)\log[P(w_j|z_k)P(z_k|d_i)] \quad (4)$$

当近似下界变化不大时停止迭代. EM 算法终止, 此时得到的值即为最终结果.

1.2 聚类方法

Kmeans 方法是经典的基于划分的聚类方法. 其基本思想为:对于给定的聚类数目 K, 首先随机选择 K 个文本作初始的类质心, 然后根据每个文本与各个类质心的相似度, 将它赋给最相似的类. 然后重新计算每个类的质心. 不断迭代以上过程,直到准则函数收敛.

具体来说, 假定将样本集分为 c 个类别, 算法描述如下:

1)适当选择 c 个类的初始中心;

2)在第 k 次迭代中, 对任意一个样本, 求其到 c 各中心的距离, 将该样本归到距离最短的中心所在的类;

3)利用均值等方法更新该类的中心值;

4)对于所有的 c 个聚类中心, 如果利用 2)3)的迭代法更新后, 值保持不变, 则迭代结束, 否则继续迭代.

该算法具有简单, 收敛速度快的优点, 然而此种聚类方法对初始聚类数 K 比较敏感. 此外, 算法的关键在于初始中心的选择和距离公式.

本文提出的微博热点话题发现的算法是先由 pLSA 获取文档-主题矩阵, 将每个主题下概率最高的文档作为 Kmeans 的聚类中心, 有效地解决了 Kmeans 算法对聚类中心敏感的问题.

2 基于概率潜在语义分析的微博热点话题发现方法

2.1 基本思想和处理流程

本文首先对微博数据集进行过滤、分词、去停用词和简繁转换等处理, 然后构造词-文档矩阵, 用 pLSA 方法对原始矩阵建立主题模型, 用 EM 算法估计出 p(zk|di)和 p(wj|zk), 然后将两个矩阵相乘得到带有隐含语义的词-文档矩阵. 此矩阵作为下阶段聚类输

入, 将矩阵每个主题下概率最大的文档作为 Kmeans 的聚类中心进行聚类得到最终结果. 算法流程图如下所示:

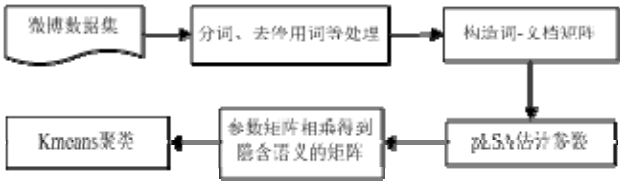

图 1 算法流程

2.2 方法描述

第 1 步: 对微博进行预处理

1)对于含有 “@用户”的微博, 由于该字段是微博转发者为了使某用户得到此条微博而加入微博文本中, 该字段对此条微博的内容并没有实质的影响, 相反会对后序建立主题模型时, 增加字典的长度, 影响最后聚类结果的准确率, 删除该字段.

2)对于含有 “RT @”字段的微博, 该字段之前的文本表示微博用户对某条微博的回复, 并不是真正的微博内容, 仅仅保留最后一个 “RT @”字段后面的微博内容, 这部分文本才是真正的微博.

3)对含有 “#某主题(某用户)#”的微博, 该字段属于噪声数据, 在预处理中将其过滤.

4)对含有繁体中文的微博, 本文根据《现代汉语通用汉字表》收集了 6600 个常见的简体中文及其对应的繁体中文字, 将微博数据中的繁体中文转换成简体中文.

5)调用 stammer 分词系统将降噪处理后的微博数据分词, 同时去掉其中包含的停用词, 停用词表采用新浪提供的 1028 个停用词.

第 2 步: 构造词-文档矩阵

对分词后的文本建索引得到词-文档矩阵 A=[aij]m×n, 其中 i, j 表示第 i 个词在第 j 篇文档中的权重. 由于微博文本短小、数目大, 单个文本中出现的词条非常有限, 因此, A 一般为稀疏矩阵. A 中特征词条权重有多种不同的计算方法, 本文采用目前最常用且效果最理想的 TF·IDF[15] 法, 计算方法如下:

$$a_{ij}=\frac{tf_{ij}\times\log_2\left(N/n_i+0.01\right)}{\sqrt{\sum_{i=1}^{n}\left(tf_{ij}\times\log_2\left(N/n_i+0.01\right)\right)^2}} \quad (5)$$

tf_{ij} 表示第 j 个文档中第 i 个词出现的频度, N 为文本集的总文本数, n_i 为含有词条 i 的文本个数.

第 3 步: 构建主题模型

将构造的词-文档矩阵文档用 pLSA 进行参数估计得到和 $p(w_j|z_k)$两个矩阵, 然后将两个矩阵相乘得到两阶段聚类的输入. 将 $p(z_k|d_i)$矩阵每个主题下概率最大的文档作为 Kmeans 的聚类中心, 在经过多次实验后, 得到使似然函数最大的 K 值作为将要抽取的微博主题数.

第 4 步: Kmeans 算法聚类

依据前面得到的 K 值和聚类中心, 执行 Kmeans 算法, 得到热点话题的聚类结果.

第 5 步: 按热度由大到小对微博排序

由前四步得到了微博的聚类结果, 但是每个类别下的微博数量仍然比较多, 需要提取出其中比较热门的话题. 本文采用了一条微博热度的计算公式[8]:

$$HT_i=N\_com_i+\sqrt{N\_rel_i}-\log(fan_i+1) \quad (6)$$

其中 HT_i 表示第 i 条微博的热度, N_com_i 表示第 i 条微博的评论数, N_rel_i 表示第 i 条微博的转发数, fan_i 表示发布此条微博的作者的粉丝数. 按照公式(6)计算每个类别下每条微博的热度 HT, 按照由大到小的顺序排列, 得到最终结果.

3 实验结果及分析

3.1 数据描述

实验数据来自两部分, 第一部分采用从新浪微博抓取的 2011-11-01 到 2011-11-03 期间发表的 20000 条帖子内容, 经人工处理分为 37 类. 从每个类别中取 33 篇文档共计 1221 篇, 对数据进行如下处理: 1)分词; 2)去停用词; 3)繁体转换成简体处理; 4)去除数据中含有的网址信息; 5)去除数据中的标点符号 6)过滤掉字长小于 10 的微博条目. 另一部分采用从新浪微博抓取的 2013-03-10 的 10000 条微博数据, 用作测试数据, 对该数据也做上述处理.

3.2 参数设定

本节首先讨论抽取主题数目 K 的确定, 在 pLSA 中, K 的取值是按照操作者经验来进行主观选取的, 实验中将 K 设定了不同值, 得到在 K 值不同的情况下, 似然函数取值的不同, 如图 2 所示.

实验结果显示当 K 取 200 的时候得到的似然函数值最大, 但是根据主题区分度和用户浏览方便程度等方面综合考虑本文将 K 设为 80, 另外 EM 算法的收敛

值和最大迭代次数设定为 0.001 和 1000 次.

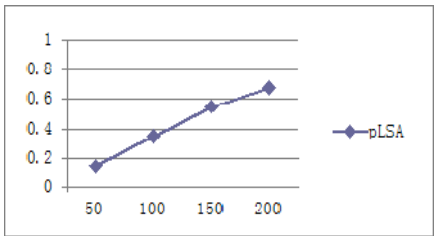

图 2　聚类个数与似然函数值的关系

3.3 实验结果与相关分析

将实验数据的第一部分进行 pLSA 和 Kmeans 聚类, 得到每篇微博属于哪一类的结果文件, 然后从原数据集中取出属于同一类的微博写入文件中, 选取含有最多文档的四个类, 计算准确率[16], 结果如下:

表 1　聚类结果中四个类的准确率

类别	正确数	错误数	准确率
1	25	12	0.68
2	22	13	0.63
3	33	6	0.85
4	26	10	0.72

本文的方法与其他经典方法对比结果如下:

微博热点话题发现方法

表 2　本文方法和传统的 VSM 以及 LSA+Kmeans 方法准确率对比

微博热点话题发现方法	平均准确率
VSM	0.55
LSA+Kmeans	0.65
pLSA+Kmeans	0.72

由实验结果可以看出, 本文的 pLSA+Kmeans 方法在准确率方面优于前两种算法. 这是因为传统的 VSM 方法忽略了中文的同义, 多义及高维向量难于计算的问题, 而传统的 LSA 方法分解出的矩阵中含有负数, 不符合文本特性.

在第二部分数据上运行程序, 描绘出最终的聚类结果如图 3.

图 3 表示每篇文档聚类完成后的结果, 图中每个颜色表示一类, 由于坐标系是平面二维坐标, 选取每个文档向量各维度中最大的两个维度数据来描绘图形得到上图结果. 由图中可以看出文档聚类结果基本符合簇内紧凑, 簇间离散的特点.

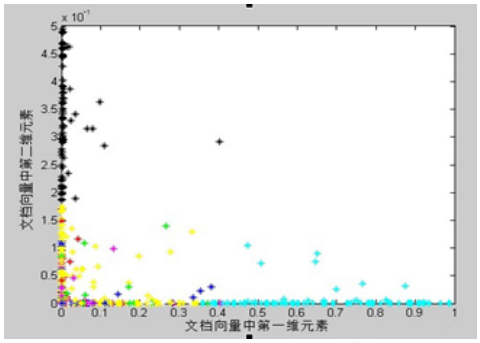

图 3　聚类结果

将聚类测试数据得到的每一个类别, 按照每个类别中含有的文档数量由大到小排序, 取前五类, 选取每一个类中的第一条微博数据作为热点话题, 结果如下:

表 3　微博热点话题

类别	微博条目
1	祝大家 2013 年新年快乐! 欢度元旦节!: 祝大家 2013 年新年快乐! 欢度元旦节!
2	中国互联网应从自身找原因, 很多国家政策是难改变的, 只能适应
3	零利率政策主要借助套利交易, 引导境内资本流向高利率国家, 实现降低本币汇率的目标.
4	继小沈阳被美国国家有线广播电视台 CNN 评为中国最低俗的人之后, 最近罗玉凤凤姐一路打败杨二车娜姆和芙蓉姐姐,被 CNN 评为中国最不要脸的人, 看来美国人在判断不要脸这个问题上非常之公正.
5	昨天国内高调报道中央要以实名制管理三聚氰胺. 虽然不知道蒙牛到底出了什么事, 但估计又是三聚氰胺惹的祸. 反正牛奶是再也不能吃了. 决不能相信说管理得很好的话.

每个类别中 5 个权重较大的词如下:

表 4　每个类别下的关键词

类别	关键词				
1	新年	愿望	许下	调整	公告
2	量化	引导	实施	政策	本币
3	世界	货币	宽松	风险	资本
4	娱乐	凤姐	评为	低俗	不要脸
5	蒙牛	发现	报导	牛奶	查收

表 3 中列出了第二部分实验数据中的 5 个热点话题, 表 4 则列出了每个类下面权重最高的 5 个关键词, 表 4 与表 3 的实验结果一致.

4 结语

随着网络的迅猛发展, 微博的普及程度会越来越大, 对于微博的热点话题发现问题也会备受关注. 本文利用概率潜在语义分析, 有效地克服了传统VSM中高维度和同义多义和LSA中分解出的近似矩阵存在负数的问题. 采用层次聚类和 Kmeans 聚类相结合的方法在一定程度上提高了聚类的准确率, 并完整展示每个主题下的微博内容. 值得注意的是对微博进行预处理时需要考虑粉丝数较少的微博信息, 例如很多广告的微博虽然关注数量多, 但是粉丝数少, 会影响聚类效果. 今后的工作方向主要有下两点: 一是如何处理pLSA 中存在过拟合的问题, 此外, 本文尚未考虑微博用户角色分类, 将角色考虑进去会提高发现热点话题的准确率. 此外, 还应在仿真试验中增加误报率、漏报率等参数, 以增加聚类方法的实用价值.

参考文献

1 Salton G. The SMART retrieval system experiments in automatic document processing. Englewood Cliffs, New Jersey: Prentice Hall Inc. 1971: 337–354.

2 张晨逸,孙建伶.基于 MB-LDA 模型的微博主题挖掘.计算机研究与发展,2011,48(10):1795–1802.

3 郑斐然,苗夺谦,张志飞,等.一种中文微博新闻话题检测的方法.计算机科学,2012,1:138–141.

4 Raghavan VV, Wong MKS. A critical analysis of vector space model for information retrieval. Journal of the American Society for information Science, 1986, 37(5): 279–287.

5 邓一贵,马雯雯.基于隐含语义分析的微博话题发现方法.计算机工程与应用,2012.

6 Hoffmann T. Unsupervised learning by probabilistic latent semantic analysis. Machine Learning, 2001, 42(1): 177–196.

7 Sebastiani F. Machine learning in automated text categorisation. ACM Computing surveys (CSUR), 2001, 34(1): 1–47.

8 Pal A, Counts S. Identifying Topical Authorities in Microblogs. Proc. of Web Search and Data Mining. New York. 2011. 45–54.

9 路荣,项亮,刘明荣,杨青.基于隐主题分析和文本聚类的微博客新闻话题发现研究.第六届全国信息检索学术会议论文集,2010.

10 Ramage D, Dumais S, Liebling D. Characterizing microblogs with topic models. Proc. of the Fourth International Conference on Weblogs and Social Media. MenloPark: AAAI Press, 2010: 130–137.

11 Teevan J, Ramage D, Morris MR. TwitterSearch: a comparison of microblog search and web search. Proc. of the Fourth Association for Computing Machinery International Conference on Web Search and Data Mining. New York, USA. 2011. 35–44.

12 Savage N. Twitter as medium and message. Communications of the Association for Computing Machinery, 2011, 54(3): 18–20.

13 Sakaki T, Okazaki M, Matsuo Y. Earthquake shakes twitter users: real-time event detection by social sensors. Proc. of the 19th International Conference on World Wide Web. 2010, 46(36): 851–860.

14 Zhang Y, Xu G, Zhou X. A latent usage approach for clustering web transaction and building user profile. Proc. of International Conference on Advanced Data Mining and Applications 2005. Wuhan, China. 2005. 231–236.

15 孙胜平.中文微博客热点话题检测与跟踪技术研究[学位论文].北京.北京交通大学,2011.