基于 BTM 和 K-means 的微博话题检测

李卫疆 王真真 余正涛
(昆明理工大学信息工程与自动化学院 昆明 650500)

摘 要 近年来,微博等社交网络的发展给人们的沟通交流提供了方便。由于每条微博都限定在140字以内,因此产生了大量的短文本信息。从短文本中发现话题日渐成为一项重要的课题。传统的话题模型(如概率潜在语义分析(PLSA)、潜在狄利克雷分配(LDA)等)在处理短文本方面都面临着严重的数据稀疏问题。另外,当数据集比较集中并且话题文档间的差别较明显时,K-means 聚类算法能够聚类出有区分度的话题。引入 BTM 话题模型来处理微博数据这样的短文本,以缓解数据稀疏的问题。同时,整合了 K-means 聚类算法来对 BTM 模型所发现的话题进行聚类。在新浪微博短文本集上进行的实验证明了此方法发现话题的有效性。

关键词 短文本,话题模型,话题发现,K-means 聚类
中图法分类号 TP393.092,TP391.1 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.02.042

Micro-blog Topic Detection Method Integrating BTM Topic Model and K-means Clustering

LI Wei-jiang WANG Zhen-zhen YU Zheng-tao
(Department of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China)

Abstract Recently, the development of micro-blog provides people with convenient communication. Because every piece of micro-blog is limited in 140 words, large scale of short texts appear. In the meantime, discovering topics from short texts genuinely becomes an intractable problem. It is hard for traditional topic model to model short texts, such as probabilistic latent semantic analysis (PLSA) and Latent Dirichlet Allocation (LDA). They suffer from the severe data sparsity when disposing short texts. Moreover, K-means clustering algorithm can make topics discriminative when datasets is intensive and the difference between topic documents is distinct. In order to improve data sparsity, BTM topic model was employed to process short texts-micro-blog data for alleviating the problem of sparsity in this paper. At the same time, we integrated K-means clustering algorithm into BTM(Bi-term Topic Model) for topics discovery further. The results of experiments on Sina micro-blog short text collections demonstrate that our method can discover topics effectively.

Keywords Short text, Topic model, Topic discovery, K-means clustering

1 引言

微博是一个能够共享各种信息、获取热门话题的社交平台,其显著特点是有字数限制。用户发微博时,仅能在140字内表达出要分享的新事物。随着简短微博文的转发,大量的短文本充斥着网络。如果能从大量短文本中发现话题,则能了解到最新、最受关注的话题,这有利于引导公众舆论、控制网络谣言。另外,发现话题对于研究信息安全和话题演变也有重要的作用。总之,对微博话题发现的研究不仅具有理论研究意义,还存在着社会现实意义[1]。

话题检测需要经过预处理、建模、相似度计算等一系列过程。其中,建模是很重要的一部分。最初,人们提出了向量空间模型(VSM)和统计语言模型来实现话题检测中的建模过程。但是,这些模型没有考虑语义部分,影响到了其发现话题的性能。PLSA[2]在 LSA 模型中融入了统计概念,以此改善模型。随后,D. Bei 提出了 LDA 模型[3],其通过文档-词共现来发现话题,并在文本话题检测领域取得了显著的成果。但是,微博长度是短小的,只包含少量内容,以至于存在的文档-词共现[4]比较少。另外,由于微博表达方式的随意性,相关话题词出现的次数就会比较少,且上下文并不丰富。以上两方面原因使得传统话题模型(如 PLSA、LDA 模型)在微博话题发现中面临着数据稀疏的挑战。

在话题检测的最后环节中通常要做出一些决策,例如单遍聚类算法[5],目前该算法普遍用于话题检测,采用增量聚类的方式处理新的报道。新报道和历史话题中心的相似度对话题检测的性能会产生很大的影响,而层次聚类算法[6]恰好适合相似度计算和距离向量。另外,K-means 聚类算法[7]也是一种广泛使用的聚类算法,它的计算速度较快,且聚类效果理

到稿日期:2015-11-25 返修日期:2016-03-29 本文受地区科学基金项目:基于统计机器翻译和自动文摘的查询扩展研究(61363045),云南省自然科学基金重点项目(2013FA130),科技部中青年科技创新领军人才项目(2014HE001)资助。
李卫疆(1969—),男,博士,副教授,主要研究方向为信息检索、自然语言处理,E-mail:hrbrichard@126.com;王真真(1991—),女,硕士,主要研究方向为自然语言处理;余正涛(1973—),男,博士,教授,主要研究方向为自然语言处理、信息检索。

想。虽然每种聚类算法都有其独特的优点，但是当簇是紧凑的并且簇与簇之间明显分离时，K-means 的效果较好[8]。因此当数据集偏于集中且话题间差别明显时，K-means 聚类得出的结果比较理想。

对于话题模型，由晏小辉提出的 BTM(Bi-term Topic Model)能够从短文本中发现更多较突出且语义一致的话题[4]。然后在此基础上，结合 K-means 聚类将会获得更满意的实验结果。受文献[7]的启发，本文引入了基于 BTM 话题模型和 K-means 聚类的微博话题检测的方法。利用 BTM 模型从短文本中获取话题，然后进行 K-means 聚类分析，从而得到区分度较好的话题实验结果。实验结果证明了所提方法的可行性和有效性。

本文第 2 节简单地介绍了相关工作；第 3 节介绍了 BTM 话题模型和 K-means 聚类算法；第 4 节展示了实验结果；最后总结了全文并展望未来工作。

2 相关工作

相关工作分为两部分：1)介绍解决数据稀疏问题的相关研究；2)介绍关于微博话题发现的相关研究。

2.1 数据稀疏问题的解决方法

根据前人的研究，在微博这类短文本集上进行文本挖掘是有难度的。所以当采用传统方法处理微博话题时，会不可避免地产生数据缺乏的问题，从而严重影响了处理短文本的有效性。

LDA 是一种普遍、稳定的非监督话题模型。研究者经常基于 LDA 模型进行一些改进来处理短文本[10]。一方面，他们通过额外的信息加强短文表示。例如：亓晓青和景晓军[11]提出应用于微博的 LDA 模型改进-用户与关联扩展(ULLDA)，将用户信息和文本关联起来，改进了 LDA 话题模型。另一方面，研究者扩展了传统的话题模型来缓解文本限制的问题。例如，谢昊和江红[10]在 LDA 的基础上提出了 RTLDA 微博生成模型，它主要解决字数受限的问题。同时，他们采用 Gibbs 算法进一步挖掘每条微博文的话题。Daniel Ramage 等[12]采用局部监督的可扩展学习模型，先将 Twitter 的内容用低维表示，然后再标记用户和推特内容。

然而，收集额外信息将增加一定的工作量，并且这样的模型依赖于额外的信息。例如，ULLDA 模型在获得用户较关注的话题上非常有效，但是它将对用户信息产生依赖。另外，基于传统模型的扩展模型也具有一定的局限性。例如，RTLDA 模型中存在这样的一个假设：每条微博只包含一个话题，很显然这个假设忽略了微博话题的多样性。

2.2 微博话题发现的研究

早期的研究主要采用聚类算法从短文本中发现话题。Ting Huang 等[13]利用自适应的 K-means 聚类算法在微博数据中发现社区。他们首先提出 CLARANS 算法，然后在此基础上结合自适应策略进行改善。孙胜[14]在向量空间模型的基础上运用了 SP&HA 算法来检测话题，先通过向量空间模型表示文本，后采用 SP&HA 算法进行计算。如果这些方法能够结合话题模型，将会取得更有效的实验结果。

目前，人们倾向于结合话题模型和文本聚类来发现话题。熊组涛[9]提出了基于稀疏特征的中文微博短文本聚类方法研究，该研究用 LDA 表示文本，然后结合 K-means 算法和层次聚类算法对微博数据聚类。米文丽等[15]在概率模型的基础上发现微博热点话题，通过 pLSA(probabilistic Latent Semantic Analysis)对微博话题进行建模，由 K-means 聚类算法产生话题，接着依据话题热度和顺序发现当前的热点话题。路荣等[7]基于隐主题分析和文本聚类来发现微博新闻话题。其中 K-means 和层次聚类算法被用来进一步聚类新闻话题。

鉴于结合话题模型和聚类算法的可行性，本文采用 BTM 话题模型来改善数据稀疏的问题；其次对 K-means 聚类算法和层次聚类算法的效果进行比较；最后采用 K-means 聚类算法以获得区分度较好的话题。

3 BTM 话题模型和 K-means 聚类算法

由于传统的话题模型通过文档-词共现来学习话题，当处理短文本时它们将面临数据稀疏的问题。为了解决这个问题，本文采用 BTM 话题模型从微博中获取不同的话题。然后，采用 K-means 聚类算法将 BTM 发现的话题进行聚类。

3.1 基本思想和方法

本文首先对微博数据集进行分词、去停用词等处理。其次，采用 BTM 进行建模，将微博短文本形成词对(bi-term)，从而获得到相应的微博话题以及话题词分布。然后，将此分布情况的文本数据通过 StringToWordVector 过滤器转换为空间向量模型，并将其作为文本聚类工具 WEKA 的输入。最后，由 K-means 聚类算法对话题词进行聚类得到最终结果。算法流程如图 1 所示。

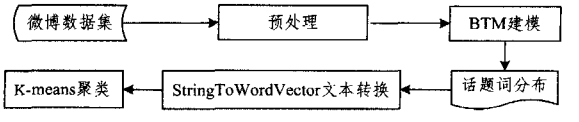

图 1 算法流程

3.2 BTM 建模及参数估计

Bi-term 代表短文本中一对无序的共现词[4]。由于每条微博都包含少量的内容，因此每条微博都可以被看作一个独立的文本单元。任意两个不同的词都将组成一个 bi-term。例如："上海外滩踩踏事件"，从这一句话中可以提取"上海外滩"、"上海踩踏"、"上海事件"、"外滩踩踏"、"外滩事件"、"踩踏事件"这几个 bi-term。本文依次从所有文档中提取双词项，提取出的双词项将组成 BTM 模型的训练数据。

LDA 模型是对带有潜在话题结构的文档的生成过程进行建模，而 BTM 模型则是对带有潜在话题结构的 bi-term 的生成过程进行建模。假定 α 和 β 是狄利克雷先验参数。BTM 的生成过程[4]如图 2 所示。

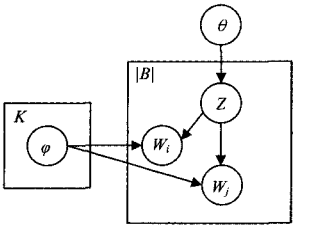

图 2 BTM 的图解模型

图中，θ 是 BTM 语料库中的微博话题 Z 的分布，φ 为微博话题下话题词的分布，话题词为词对(bi-term)w_i，w_j。由

图 2 可以看出，BTM 模型利用整个微博短文本集的信息将微博短文本形成词对，从而以整个语料库的层面来描述微博话题 Z，这样既保持了词之间的相关性，又获得了不同词表达不同话题的独立性。

词对 bi-term(b)的联合概率表示如下：

$$P(b)=\sum_z P(z)P(w_i|z)P(w_j|z)=\sum_z \theta_z \phi_{i|z}\phi_{j|z} \tag{1}$$

那么整个 BTM 语料库的概率表示为：

$$P(B)=\prod_{(i,j)}\sum_z \theta_z \phi_{i|z}\phi_{j|z} \tag{2}$$

类似于 LDA，ϕ 和 θ 是 BTM 话题模型的隐含变量，因此需要从微博中的观察变量词项来估计参数 ϕ 和 θ。BTM 通过 Gibbs 来推断微博话题下话题词的分布和话题分布。

首先，为 bi-term 计算条件概率分布：

$$P(z|z_{-b},B,\alpha,\beta)\propto(C_{bz}^{BT}+\alpha)\frac{(C_{wz}^{WT}+\beta)\ (C_{jz}^{WT}+\beta)}{(\sum_w C_{wz}^{WT}+M\beta)^2} \tag{3}$$

其中，$P(z|z_{-b},B,\alpha,\beta)$ 是每个 bi-term 的条件概率分布，z_{-b} 表示除了第 b 个 bi-term 的所有 bi-terms 的话题分布。B 是 bi-terms的集合。C_{bz}^{BT} 是 bi-term b 分配给微博话题 Z 的次数，C_{wz}^{WT} 是分配给微博话题 Z 的话题词 w，C_{iz}^{WT} 和 C_{jz}^{WT} 是分配给话题 z 的话题词 i 和 j。

其次，确定整个 BTM 语料库的话题分布和微博话题下话题词的分布：

$$\phi_{w|z}=\frac{C_{wz}^{WT}+\beta}{\sum_w C_{wz}^{WT}+M\beta} \tag{4}$$

$$\theta_z=\frac{C_{bz}^{BT}+\alpha}{n_b+T\alpha} \tag{5}$$

其中，$\phi_{w|z}$是话题-词分布，θ_z 是话题分布，n_b 是 bi-terms 的总数。

最后，通过式(6)和式(7)计算每条微博中 bi-term 的分布和每条微博的话题分布，其中，$n_z(b)$ 是微博 d 中 bi-term b 的频率。

$$P(b|d)=\frac{n_d(b)}{\sum_b n_d(b)} \tag{6}$$

$$P(z|d)=\sum_b P(z|b)P(b|d) \tag{7}$$

3.3 数据规格化

通过 BTM 模型得到的微博话题-话题词分布为文本形式，那么利用 WEKA 进行文本聚类时，需要先通过转换器将文本数据转换成 ARFF 文件，即 WEKA 的输入形式；然后使用过滤器将 ARFF 文件中的微博话题数据模型转换为向量空间模型，从而完成 K-means 进行文本聚类的准备工作。

文本格式转换需要先将 BTM 所获得的微博话题-话题词分布存储为文本形式，然后使 BTM 获得的每个话题下的话题词独立为不同的类，并作为聚类的对象。那么，利用转换器可以把这些不同的类转换成 ARFF 文件，其形式如表 1 所列。

表 1　ARFF 文件格式

@relation D_weka_BTM-Text
@attribute text string
@attribute @@class@@ {class1,class2,…,}
@data
'topic1_word1 topic1_word2 …'class1
'topic2_word1 topic2_word2 …'class2
'topic3_word1 topic3_word2 …'class2
…

数据模型转换：将 ARFF 文件中的微博话题词作为过滤器的输入，经过 Tokenizer，Stemming 和 Stopwords 等数据处理以及 TF 和 IDF 的统计计算，最后通过归一化把 ARFF 格式的微博话题词转换为向量空间模型。

3.4 K-means 聚类算法

在经过数据规格化后，本文采用 K-means 聚类算法对话题进行聚类[16]。K-means 是一种基于划分的聚类算法，它不仅简单，而且应用广泛。K-means 在聚类之前需要先指定聚类的簇数，然后数据部分将被划分为所指定的数目。本文中，K-means 首先选择 BTM 获得的 k 个话题中的首个高频话题词作为初始聚类的中心。其次，计算出平均值并重新分配话题词以产生新簇，最后更新簇中心，至达到收敛状态。

基于 BTM 的 K-means 算法的基本思想如下。

算法 1　基于 BTM 话题模型的 K-means 算法

输入：簇数目 k，由 BTM 话题模型发现的话题集 B

输出：簇集 D

步骤：

1)从簇集 D 中随机选择 k 个数据点即高频话题词作为簇的中心；

2)将每个话题词分配到最近的中心；

3)通过计算话题的平均值更新簇的中心：

$$d_{ij}=\sqrt{\sum_{k=1}^{n}(x_{ik}-x_{jk})^2} \tag{8}$$

x_{ik}，x_{jk}为簇 k 的话题词 i 和话题词 j；

4)直至话题达到收敛

$$\lim \Sigma(x_i-x)^2=0 \tag{9}$$

当数据集比较密集，并且簇与簇之间的差别很明显时，K-means 能够取得较好的聚类结果。另外，BTM 模型能够有效地发现不同的话题，在此基础上再运用 K-means 算法来加强话题发现的效果，从而得到区分度较好的话题。

4　实验与结果分析

本文将在新浪微博数据集上进行实验，并证明 BTM 话题模型和 K-means 聚类算法结合的有效性。在实验中，首先将 BTM 和 LDA 两种模型发现话题的质量进行比较；然后，评估 BTM 模型结合 HC 聚类算法的效果，以及 BTM 与 K-means 结合的效果；最后，再将 LDA&K-means 方法与 BTM&K-means 方法进行对比，从而证明 BTM 模型比传统的 LDA 模型在处理微博短文本方面更具实用性。

4.1 数据集

本文的数据集是随机从新浪微博中采集的，然后采集到 10 类话题，即“开学季”、“抗战阅兵”、“读书”、“天津大爆炸”、“上海踩踏事件”、“金融”、“汽车”、“电脑”、“体育”、“招聘”。每个话题包括大约 100 条微博。

为了提高数据集的有效性，本文用中文分词系统 NLPIR (ICTCLAS2013)来进行数据的预处理：1)分词；2)去除停用词；3)去除单字；4)将带有@用户的微博去掉；5)删除副本。

4.2 BTM 模型与 LDA 模型发现话题的不同结果

在预处理每条微博后，使用 BTM 话题模型来获得微博的话题。首先将话题数目设置为 10，然后列出前 3 个(top)话题来对比话题的一致性。如表 2 所列，这 3 个话题分别为“上海踩踏事件”、“天津大爆炸”、“抗战阅兵”。

表 2 BTM 模型-话题中的高频词和非高频词

Top5	Topic 1	Topic 2	Topic 3
1	踩踏	天津	战争
2	践踏	爆炸	阅兵
3	事件	塘沽	周年
4	上海	事故	70
5	外滩	关注	胜利
Non-Top3	Topic 1	Topic 2	Topic 3
1	名单	讨论	中国
2	公告	事件	经历
3	死亡	罕见	高涨

另一方面，话题的非高频(non-top)词汇如表 2 的 Non-Top3 所列。第一个话题中的这些词表达了“死亡名单的公告”，第二个话题中的词说明“天津塘沽大爆炸事件已经引起了人们的热议”，而最后一个话题中的非高概率词“中国”、“经历”、“高涨”陈述了中国人民对抗战大阅兵饱有高涨的情绪。值得注意的是：非高频话题词和高频话题词在含义表达上保持了一致，并且与话题也保持有密切的相关度，这说明 BTM 在微博短文集上能够有效地发现不同的话题。

同样，在预处理原始微博后，采用 LDA 模型来获得微博话题，将其结果与 BTM 模型所得结果进行对比，如表 3 所列。在话题“上海踩踏事件”中高频话题词出现“极好”、“大海”和“生命”，并且非高频话题词中的“孩子”、“参加”这些话题词与上海踩踏事件缺乏密切的关系。在 Topic2 和 Topic3 中依次可以看到高频词“天气”、“说话”以及非高频词“工作”、“会议”并没有与“天津大爆炸”话题保持语义一致，高频词“招聘”、“工作”和非高频词“校园”、“日期”与“抗战阅兵”话题并不相符，而是出现“招聘”等区别于 Topic3 的话题词。

表 3 LDA 模型-话题中的高频词和非高频词

Top5	Topic 1	Topic 2	Topic 3
1	极好	天津	中国
2	踩踏	爆炸	招聘
3	事件	大事	70
4	大海	天气	工作
5	生命	说话	周年
Non-Top3	Topic 1	Topic 2	Topic 3
1	孩子	工作	校园
2	参加	会议	日期
3	发生	事故	名额

由此对比结果可以得出，BTM 在较短微博集上解决数据稀疏问题的性能优于 LDA 模型，能够使得非高频话题词依然与话题保持语义一致性。

4.3 BTM 结合聚类算法的评估

基于 BTM 模型，本文采用聚类算法来进一步发现话题。采用这种方法可以避免话题的冗余和交叉，保持较好的区分度，同时可以保证话题的完整性。

本文中使用 Weka 实现聚类，其中用“TextDirectoryLoader”和“StringToWordVector”进行话题文本的处理。然后，以层次聚类作为理论基线来评估 K-means 聚类算法。众所周知，准确率(P)和召回率(R)能够反映准确的程度和查全程度。如果需要更多的相关话题，则会追求准确率；如果期望话题更为全面，则倾向于召回率。本文采用 F 值(F-measure)来衡量聚类效果，因为 F 值结合了准确率和召回率。

准确率(P)、召回率(R)和 F 值(F-measure)的表达式如式(10)—式(12)所示。

$$P(i,j)=\frac{\text{the number of class } i \text{ in cluster } j}{\text{the number of documents in cluster } j} \tag{10}$$

$$R(i,j)=\frac{\text{the number of class } i \text{ in cluster } j}{\text{the whole number of documents with class } i} \tag{11}$$

$$F(i)=\frac{2PR}{P+R} \tag{12}$$

其中，i 代表类，j 表示簇。另外，对于类 i，F 值高的簇指向类 i。

$$F=\frac{\sum_i[|i|\times F(i)]}{\sum_i|i|} \tag{13}$$

最后，用每类的加权平均得到最后的聚类结果。

下文在 BTM 发现话题的结果的基础上将 K-means 聚类和层次聚类进行了对比。首先，设置话题数目为 7 和 10 两种。然后，在不同簇的情况下测试层次聚类和 K-means 聚类的 F 值。表 4 列出了话题数目为 7 且簇数目为 5,6,7,8 时的两种不同方法的 F 值。同时，从图 3 可以得出 K-means 的 F 值总是高于层次聚类的 F 值；并且，当簇的数目越接近话题数目时所呈现的 F 值越高。

表 4 话题数为 7 时层次聚类和 K-means 聚类在不同簇的 F 值

Cluster	HC-F	K-means-F
5	0.655	0.649
6	0.596	0.798
7	0.770	0.774
8	0.768	0.845

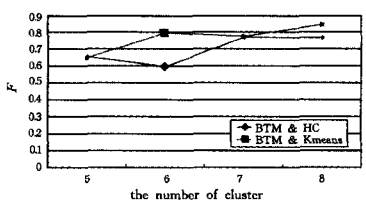

图 3 7 个话题下不同簇数目的 F 值

其次，通过 BTM 模型获得了 10 个话题，并设置簇数为 4,5,6,7,8,9,10,11,12 来分别获得 K-means 聚类和层次聚类的 F 值。表 5 表明，结合了 BTM 话题模型后，K-means 的聚类效果比层次聚类的效果更好。如图 4 所示，F 值在0.3～1.0 间波动且当簇的数目接近话题数目时，F 值趋近于 1.0。因为话题类别为 10，即使设置簇的数目为 11 和 12，聚类结果也仅显示为 10 个簇，所以 F 值依然保持为 1.0。

表 5 话题数为 10 时层次聚类和 K-means 聚类在不同簇的 F 值

Cluster	HC-F	K-means-F
4	0.333	0.515
5	0.490	0.575
6	0.576	0.718
7	0.697	0.743
8	0.837	0.849
9	0.903	0.915

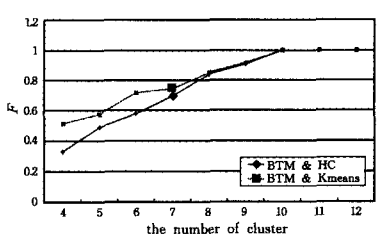

图 4 10 个话题下不同簇数目的 F 值

4.4 BTM&K-means 与 LDA& K-means

本节将 BTM 结合 K-means 的算法与 LDA 结合 K-means 的算法进行比较。在 BTM 建模后,分别获得 7 个和 10 个不同的微博话题,K-means 基于这些微博话题下的话题词聚为不同的簇,并依次计算出相应的 F 值。同样地,在 LDA 建模后,加以 K-means 聚类得到相应的 F 值,与之形成对比。话题数为 7 时,BTM 结合 K-means 与 LDA 结合 K-means 在不同簇的 F 值如表 6 所列。

表 6　话题数为 7 时 LDA+K-means 与 BTM+K-means 在不同簇的 F 值

Cluster	F_LDA+K-means	F_BTM+K-means
5	0.647	0.649
6	0.714	0.798
7	0.774	0.774
8	0.829	0.845

如图 5 所示,BTM& K-means 的走势要高于 LDA & K-means。由于微博短文本中与话题相关的词汇出现的频率较低,话题词较少,这将影响 LDA 处理文本的性能。而 BTM 模型能够有效解决数据稀疏问题,在发现不同话题后结合 K-means聚类可以取得理想的效果。

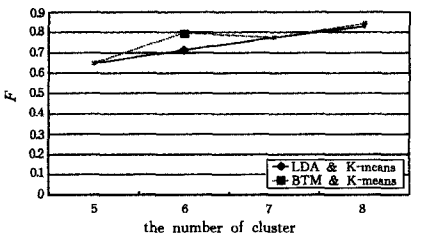

图 5　7 个话题下 BTM 与 LDA 结合 K-means 的 F 值

同时,BTM 和 LDA 模型获分别结合 K-means 聚类算法在 10 个不同话题的情况下,,在不同簇的 F 值如表 7 所列。图 6 示出了 BTM&K-means 和 LDA&K-means 的效果,相比之下,LDA 与 K-means 相结合的效果并没有 BTM 与 K-means 相结合的效果好。

表 7　话题数为 10 时 LDA+K-means 与 BTM+K-means 在不同簇的 F 值

Cluster	HC-F	K-means-F
4	0.509	0.515
5	0.521	0.575
6	0.681	0.718
7	0.740	0.743
8	0.806	0.849
9	0.909	0.915

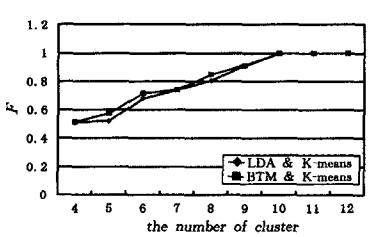

图 6　10 个话题下 BTM 与 LDA 结合 K-means 的 F 值

结束语　话题发现是一项日益重要的工作,这是由于话题传播的影响是不可控的。与普通的文本挖掘相比,从微博数据中发现潜在话题将面临数据稀疏和信息量缺乏的问题。另一方面,聚类方法的选择也是应该进行比较和改善的。针对以上问题和难点,本文提出结合 BTM 模型和 K-means 聚类算法来解决数据稀疏问题和改善聚类的问题。本文实验部分是在新浪微博数据集上实施的,最后取得了 F 值较高的相应话题。

参 考 文 献

[1] HUANG S Q, YANG Y T, LI H K, et al. Topic Detection from Microblog Based on Text Clustering and Topic Model Analysis [C]// 2014 Asia-Pacific Services Computing Conference. IEEE, 2014:88-92.

[2] HOFMANN T. Probabilistic latent semantic indexing [C]// Proc. of the 22nd Annual ACM Conference on Research and Development in Information Retrieval. California, Berkeley, 1999: 50-57.

[3] BLEI D, NG A, JORDAN M. Latent dirichlet allocation[J]. The Journal of Machine Learning Research, 2003(3):993-1022.

[4] YAN X H, GUO J F, LAN Y Y, et al. A Biterm Topic Model for Short Texts [C] // International Conference on World Wide Web. ACM, 2013:1445-1456.

[5] LIU S B, LIU L. Combining Parametric and Nonparametric Topic Model to Discover Microblog Event [C] // Information Science, Electronics and Electrical Engineering (ISEEE). IEEE, 2014:1527-1531.

[6] WANG Y Y, WANG L, QI J, et al. Improved Text Clustering Algorithm and Application in Microblogging Pubic Opinin Analysis[C]// 2013 Fourth World Congress on Software Engineering (WCSE). IEEE, 2013:27-31.

[7] LU R, XIANG L, LIU M R, et al. Discovering News Topics from Micro-blogs based on Hidden Topics Analysis and Text Clustering[J]. Pattern Recognition & Artificial Intelligence, 2012, 25(3):382-387. (in Chinese)
路荣,项亮,刘明荣,等. 基于隐主题分析和文本聚类的微博客中新闻话题的发现[J]. 模式识别与人工智能, 2012, 25(3):382-387.

[8] HAN J W, MICHELINE K. 数据挖掘:概念与技术(第 2 版)[M]. 范明,孟小峰,译. 2007:263-266.

[9] XIONG Z T. Clustering Algorithm Research in Micro-blog Short Text based on Sparse Feature[J]. Software Guide, 2014, 13(1): 133-135. (in Chinese)
熊祖涛. 基于稀疏特征的中文微博短文本聚类方法研究[J]. 软件导刊, 2014, 13(1):133-135.

[10] XIE H, JIANG H. Improved LDA model for micro-blog topic mining[J]. Journal of East China Nornal University, 2013(6): 93-101.

[11] 亓晓青,景晓军. 应用于微博的 LDA 模型改进[EB/OL]. http://www.paper.edu.cn.

[12] RAMAGE D, DUMAIL S T, LIEBLING D J. Characterizing Micro-blogs with Topic Model[C]// 4th International AAAI Conference on Weblogs and Socail Media. 2010:130-137.

[13] HUANG T, PENG D L, CAO L D. Discovering Communities with Self-adaptive k Clustering in Micro-blog Data[C]// 2012 Second International Conference on Cloud and Green Computing (CGC). IEEE, 2012:383-390.

(下转第 274 页)

随机序列。实验分为两个部分:1)比较本文算法与其他算法的表现;2)检测当每步贪心算法公共格中的结点数量不同时本文算法的表现。

由定义 1 与定义 2 可知,格中的 α,β 结点对计算结果无本质影响,在实际计算中可省略。

4.1 3 个算法的实验比较

本节比较本文算法与其他近似算法的表现。其中 N 为原始序列集合中序列的平均长度。数据集中 10(50)代表 10 条长度为 50 的序列集合,以此类推。相关位置算法简称为相位算法。

表 1 本文算法与其他算法的实验结果

数据集	贪心算法	相位算法	本文算法(1N)	
	LCS 长度	LCS 长度	LCS 长度	LCS 数量
10(50)	12	13	14	8
10(100)	25	29	29	2
40(100)	18	20	21	8
100(200)	40	39	40	2
100(300)	62	59	63	2

从表 1 的实验结果中可以归纳出,在设定公共格的结点数量是数据集中的序列平均长度的 1 倍时,对于不同的数据集,本文的算法得到的 LCS 长度均优于传统贪心算法与相关位置算法;且本文算法所求得的 LCS 为多条,相比于传统贪心算法与相关位置算法,可以保留更多的信息。该实验也验证了本文算法最终求得的解为多解的结论的正确性。

4.2 公共格的结点数量不同时本文算法的表现

从表 2 所列的实验结果中可以归纳出,当公共格的结点数量分别为 $1N,3N,5N$(N 为原始序列集合中序列的平均长度)时,若求解出的 LCS 的长度没有改变,那么解集中 LCS 的条数便会增多。若解集中 LCS 的长度增加,则原解集中 LCS 均变为了次长公共子序列,因此解集中 LCS 的数量会减少。

表 2 格期望不同的情况下的算法表现

格期望	1N		3N		5N	
数据集	长度	数量	长度	数量	长度	数量
10(50)	14	8	14	18	15	1
10(100)	29	2	29	20	30	2
40(100)	21	8	21	17	22	98
100(200)	40	2	42	8	42	20
100(300)	63	2	64	2	64	14

可以总结为:若增加公共格的结点数量,公共格中保存的公共子序列的信息也会增加,则公共格中的路径数量就会增加,就会有更大的可能去寻找到更长的公共子序列。即使 LCS 的长度没有改变,格中包含的比原来更多的信息也可以使得最终求解出的 LCS 的数量增加。

结束语 求解最长公共子序列在诸多领域内有着重要的应用,如信息检索、基因序列匹配等。求信息序列的最长公共子序列可以提取信息序列的公共信息,从而进行进一步的检索与分类。求基因序列的最长公共子序列可以获取基因序列间的匹配度。

本文算法引入了代数结构“格”,通过动态规划求解出两条序列的公共格,并结合贪心策略递归求解当前格与当前序列的公共格。公共格的路径保存了多条公共子序列,由于格结构的多路径性质,使得求解出的最长公共子序列有多个。对算法的相关定理给出了理论证明。通过实验验证了本文算法的正确性,归纳讨论了本文算法在公共格结点不同的情况下所求得的解集的变化趋势与原因。

参 考 文 献

[1] MAIER D. The complexity of some problem on subsequences and superseuqences[J]. Journal of the ACM (JACM), 1978, 25(2): 322-366.

[2] BERGROTH L, HAKONEN H, RAITA T. A survey of longest common subsequence algorithms[C]// Processings. Seventh International Symposium on String Processing and Information Retrieval, 2000(SPIRE 2000). IEEE, 2000: 39-48.

[3] ATTWOOD T, FINDLAY J. Fingerprinting g-protein-coupled receptors[J]. Protein engineering, 1994, 7(2): 195-203.

[4] SANKOFF D, BLANCHETTE M. Phylogenetic invariants for genome rearrangements[J]. Journal of Computational Biology, 1999, 6(3/4): 431-445.

[5] SHUKLA A, AGARWAL S. A relative position based algorithm to find out the longest common subsequence from multiple biological sequences[C]// Proceedings of the 2010 International Conference on Computer and Communication Technology (ICCCT). 2010: 496-502.

[6] RIZVI S, AGARWAL P. A new index-based parallel algorithm for finding longest common subsequence in multiple dna sequences[C]// International Conference in Cognitive Systems. Citeseer, 2005.

[7] HAKATA K, IMAI H. Algorithms for the longest common subsequence problem for multiple strings based on geometric maxima[J]. Optimization Methods and Software, 1998, 10(2): 233-260.

[8] BOURQUE G, PEVZNER P A. Genome-scale evolution: reconstructing gene orders in the ancestral species[J]. Genome research, 2002, 12(1): 26-36.

[9] SHERIDAN R P, VENKATARAGHAVAN R. A systematic search for protein signature sequences[J]. Proteins-Structure Function and Bioinformatics, 1992, 14(1): 16-28.

(上接第 261 页)

[14] SUN S P. Research on Chinese Micro-blog Hot Topic Detection and Tracking[D]. Beijing: Beijing Jiaotong University, 2011. (in Chinese)
孙胜平. 中文微博客热点话题检测与跟踪技术研究[D]. 北京: 北京交通大学, 2011.

[15] MI W L, SUN Y X. Microblog Hot Topics Discovery Method based on Probabilistic Topic Model[J]. Computer Systems & Applications, 2014, 23(8): 163-167. (in Chinese)
米文丽, 孙曰昕. 利用概率主题模型的微博热点话题发现方法[J]. 计算机系统应用, 2014, 23(8): 163-167.

[16] ZHENG L. Reserch and Application of Topic Detection on Micro-Blog[D]. Harbin: Harbin Institute of Technology, 2012. (in Chinese)
郑磊. 微博客话题检测的研究与实现[D]. 哈尔滨: 哈尔滨工业大学, 2012.