基于隐主题分析和文本聚类的微博客中新闻话题的发现

路 荣 项 亮 刘明荣 杨 青

(中国科学院自动化研究所 模式识别国家重点实验室 北京 100190)

摘 要 提出一种在大规模微博客短文本数据集上发现新闻话题的方法.利用隐主题分析技术,解决短文本相似度度量的问题.在每个时间窗口内,根据新闻的特点选取出最有可能谈论新闻事件的微博客文本,然后用两层的 K 均值和层次聚类的混合聚类方法,对这个时间窗口内的那些最有可能谈论新闻事件的微博文本进行聚类,从而检测出新闻话题.此方法能较好地解决微博客短文本的数据稀疏性及数据量巨大的问题.实验证明该算法的有效性.

关键词 微博客,短文本,隐主题模型,话题发现,混合聚类

中图法分类号 TP 3

Discovering News Topics from Microblogs Based on Hidden Topics Analysis and Text Clustering

LU Rong, XIANG Liang, LIU Ming-Rong, YANG Qing

(*National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing* 100190)

ABSTRACT

A method of news topics extraction from large-scale short posts of microblogging-service is proposed. Through the hidden topic analysis, the similarity measurement of short texts is solved well. In every time window, the short posts which are most likely to talk about news events are selected according to the characteristics of the news. Then, a two-level K-means-hierarchical hybrid clustering method is used to cluster all the selected data into different news topics. The experimental results show the proposed method works well on large-scale microblog dataset.

Key Words Microblog, Short Text, Hidden Topic Model, Topics Extraction, Hybrid Clustering

1 引 言

微博客(微博)是一个基于用户关系的信息分享、传播及获取平台,用户可通过 Web、Wap 及各种客户端组件个人社区,用很短的文字更新信息,并实现即时分享.国内外著名的微博服务包括 Twitter、新浪微博、腾讯微博等.微博内容简单、传播迅速,有利于新闻话题在其中快速扩散.因此从微博中检测出的新闻话题,对舆情监控、信息安全、金融证券、行业调研都有十分重要的意义.

收稿日期:2010-10-13

作者简介 路荣,男,1985 年生,博士研究生,主要研究方向为社会化网络平台上新闻事件的挖掘.E-mail:rlu@nlpr.ia.ac.cn.项亮,男,1985 年生,博士研究生,主要研究方向为推荐系统、社交媒体.刘明荣,男,1982 年生,博士研究生,主要研究方向为社交问答、搜索引擎.杨青,男,1970 年生,研究员,博士生导师,主要研究方向为社会化网络、新媒体.

传统的用特征向量来表示文本的方法,通常以词或短语作为特征,并使用 TF-IDF 的办法来衡量每个特征(即向量每一维)的权重.但是对于微博来说,它的文本内容非常短,同一个词出现在不同短文本中的概率会远小于长文本,这种数据的稀疏性,使得传统文本表示方法很难准确计算文本间的相似度.

有很多方法尝试解决这种短文本的数据稀疏性问题.其中一种办法是通过搜索引擎来扩展短文本的上下文[1-2].这种办法的缺点是非常耗时,尤其对即时系统非常不适合.另一种办法是通过隐主题模型来给短文本建模,将短文本表示成隐主题按一定比例的混合,它的好处是能充分挖掘文本集合的内在信息,从而减少短文本的数据稀疏性的影响.

隐主题建模的研究有很多[3-6].其中,隐语义检索(Latent Semantic Indexing, LSI)[6]是通过构造文本特征向量矩阵,然后对该矩阵进行奇异值分解来实现的.概率隐语义检索(Probabilistic Latent Semantic Indexing, PLSI)[7]则在 LSI 的基础上提出一个有坚实统计理论基础的生成模型.而 Latent Dirichlet Allocation(LDA)[3]则是一个完全的文本生成模型,其基本思想是,每个文本都是由多个隐主题混合而成,而每个隐主题又由多个词混合而成.本文就采用 LDA 建模的方法,它能较好地克服短文本的数据稀疏问题.

由于微博更新的便利性,每时每刻都有大量微博数据产生.目前 Twitter 上,每天更新的微博数已超过 5 千万.如此大规模的数据,如果直接使用聚类算法,效率和质量都难以得到保证.此外,并不是所有的微博都是描述新闻事件的,很多微博只是描述用户心情、状态、工作情况等.因此,本文根据新闻的特点,首先将描述新闻事件的微博选取出来,再使用聚类的方法.但尽管如此,选取出来的微博,数量仍可能非常巨大,常用的层次聚类的方法时间上将无法忍受,而 K 均值聚类的方法,在类中心个数 K 远小于微博总数时,聚类速度将大幅提升,但事先指定类中心个数(新闻话题的个数),这显然并不容易做到.所以,本文采用一种两层的 K 均值和层次聚类的混合聚类方法来克服这两种聚类方法的缺点.

传统的新闻话题的检测与跟踪(Topic Detection and Tracking, TDT)[8-10],其研究对象主要是新闻报道和博客等,一般是较长的文本,且关注点较多包括报道切分、话题跟踪、话题发现、新事件发现和报道关联发现,而研究数据多采用 TREC 会议提供的 TDT 语料[11-12],规模较小,和本文研究背景有较大差别.本文专注的是大规模微博客短文本中新闻话题的检测.实验结果表明,本文提出的整体框架和具体实现,可较好地从大规模微博客数据中检测出新闻话题.

2　基于隐主题挖掘和文本聚类的微博客新闻话题发现

2.1　数据准备

为了本文的研究,我们从国外著名的微博网站 Twitter,选取 21 302 个用户,抓去他们从 2010 年 2 月 24 日~3 月 17 日,共 22 天,发表的所有微博数据.

在 Twitter 上,每篇微博称为一个 tweet.除了普通的 tweet 之外,还有两种常见的 tweet:1)用户之间相互回答,称作 *reply*;2)转发的微博,称作 *retweet*.Twitter 对每篇 *tweet* 的长度都做了不得超过 140 个字符的限制.为了从所有 *tweets* 中,找出那些最有可能描述新闻事件的 *tweets*,在去除停用词后,我们将长度小于 4 的 *tweets* 全部删去,剩下的 *tweets* 作为有效 *tweets*.这样最后我们的数据集中共有3 079 860 个有效 *tweets*,其中最长的有效 *tweet* 含有 28 个单词,涉及用户 14 287 个.

2.2　方法思想和基本框架

对于大规模的微博客短文本,从中发现新闻话题,需要克服两个难点:1)如何表示短文本进行有效地相似度度量;2)如何快速准确的处理大规模的微博数据.

本文首先通过隐主题模型,将微博客短文本集的隐主题充分挖掘出来,从而减少短文本的数据稀疏性对文本相似度度量的影响.然后,根据新闻的特点,从大量数据中选取出最有可能描述新闻话题的

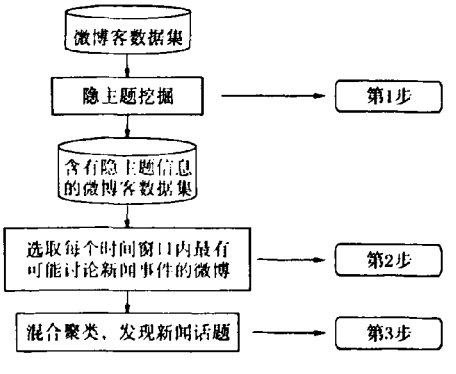

图 1　微博中发现新闻话题的整体框架图

Fig. 1　General framework of discovering news topics from microblog

微博. 再用一种快速准确的混合聚类的方法,将选取出来的微博,通过聚类聚合成不同的新闻话题.

本文提出的方法框架可以用图1来表示.

下面将依次介绍上面提出方法的3个步骤.

2.3 隐主题挖掘

前面提到过,目前常用的隐主题建模的有LSI、PLSI、LDA等方法. LDA模型相对于LSI和PLSI模型具有清晰的层次结构,是一个完全的生成模型. 所以本文选用LDA模型来对收集的Twitter数据集进行隐主题建模.

2.3.1 LDA模型

LDA模型是一种文本生成模型,它可将单个文本表示为所有隐主题的特定比例的混合. 如图2所示,为LDA模型的贝叶斯网络图.

在文档的生成过程中,LDA首先从Dirichlet分布中抽样产生一个文本特定的主体多项式分布. 然后对这些主题反复抽样,产生文本中的每个词. 如图2所示,α和β是文本集合的参数,α反映文本集合中隐主题的相对强弱,β刻画所有隐主题自身的概率分布. 随机变量θ其分量表示目标文档中个隐含主题的比重. z表示目标文档分配在每个词上的隐主题比重,w是目标文档的词向量表示. 更详细的LDA模型的描述请参见文献[3].

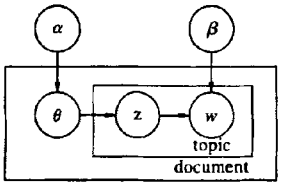

图2 LDA图模型

Fig. 2 Graphical model of LDA

2.3.2 LDA建模结果

对LDA模型的参数进行估计的办法也有很多,常用的有EM算法和吉布斯采样(Gibbs Sampling). 本文选用吉布斯采样方法推断LDA模型的参数,使用GibbsLDA++[13]对收集的Twitter数据集建模. 根据经验,隐主题数目取200,初始超参数α取0.25,β取0.1.

结果如下:

1) $\boldsymbol{\Phi}$,一个$K \times V$的矩阵,K表示隐主题的个数,V文本集合中所有不同词的个数. 它表示每个隐主题生成每个词的概率;

2) $\boldsymbol{\Theta}$,一个$M \times K$的矩阵,M是数据集中文本总数,K是隐主题个数. 它表示数据集中每个文本生成隐主题的概率;

3) Z,它的每一行都是原数据集中的一个文本,但和原数据集中不同的是,所有文本中的每个词都被标记到某一个隐主题中.

2.3.3 单义词单元

如果一个词w在文本集中两个不同的位置(不同文本中,或同一文本的不同位置)出现,并被分配到两个不同的隐主题t_i和t_j下,因为LDA假设各个隐主题之间是相互独立的,所以$w:t_i$和$w:t_j$可看作w的两个不同词义,即w为多义词. 将$w:t$的形式定义为一个单义词单元.

例如"jobs"这个词. 在隐主题建模之后,有两个隐主题产生该词的概率较大,表1列出这两个隐主题及其中发生概率较大的词.

在一些位置上,"jobs"表示和"工作"相关的意义被分配到隐主题"Topic 24"上,另一些位置上"jobs"表示美国苹果公司总裁的姓名,被分配到隐主题"Topic 56"上. 本文把"jobs: Topic 24"和"jobs: Topic 56"称为一个单义词单元. 这样每篇*tweet*都可表示成一些单义词单元组成的向量,如

$$d = (w_1 : t_1, w_2 : t_2, \cdots, w_n : t_n).$$

2.3.2节的结果3就是这种形式的文本向量集合.

2.4 选取新闻微博

在获取每个微博文本的隐主题信息后,可利用新闻的一些特点,从大规模的微博客数据中选取出那些最有可能讨论新闻话题的微博. 本文首先给数据集中出现的所有单义词单元$w:t$评分,然后每个*tweet*的得分就是其包含的所有单义词单元得分之和. 得分最高的*tweets*被认为最有可能谈论新闻话题.

此外,之所以选择单义词单元,而不只仅仅是给词评分,是因为单义词单元可有效解决多义词问题,对准确选取新闻微博有帮助. 后面的实验证明这样做确实可取得更好的效果.

表1 两个产生"jobs"概率较大的隐主题

Table 1 Two hidden topics which are most likely to generate "jobs"

隐主题序号	隐主题生成概率最大的若干个词
Topic 24	job pay work works bills customers jobs stuff creative tools workers
Topic 56	iphone app ipad apple touch store ipod software artist itunes jobs

2.4.1　给单义词单元评分

一般新闻有两个重要特性:1)内容非常新颖,之前很少出现过相似内容,而在某个时段忽然出现;2)内容十分重要,有重大影响力或是极具争议性,因而在出现后的短期内,会引起大量关注,并出现大量相关内容的讨论.

根据这两个特性,先将前面得到所有数据按时间顺序分配到若干个时间窗口中,可有以下结论:如果一个单义词单元在某个时间窗口内相比前一个时间窗口内出现的次数明显增多,可认为它和一些新的话题相联系;如果一个单义词单元在某个时间窗口内,出现的次数比该时间窗口内其它单义词单元明显多,那么认为它和一些重大、热门话题关联.

为此作如下定义. 单义词单元 $w:t$ 的文档频率 df 和用户频率 uf:

$$df(w:t) = |tweets:(w:t) \in tweet|,$$

$$uf(w:t) = |users:(w:t) \in tweet \in users|.$$

df 描述讨论某个单义词单元 $w:t$ 的 *tweets* 数量,uf 描述讨论某个单义词单元 $w:t$ 的用户数量. 那么,在当前时间窗口内单义词单元 $w:t$ 的文档频率可表示为 $df_c(w:t)$,用户频率为 $uf_c(w:t)$. 前一时间窗口中则为 $df_h(w:t)$ 和 $uf_h(w:t)$.

那么单义词单元文档频率和用户频率在两个时间窗口间的变化:

$$\Delta(df) = df_c(w:t) - df_h(w:t),$$

$$\Delta(uf) = uf_c(w:t) - uf_h(w:t).$$

因为

$$变化率 = \frac{变化}{时间窗口长度},$$

而为了便于计算,本工作中时间窗口的大小是固定的. 所以变化率可用变化来代替. 最后给出评分规则如下.

如果 $df(w:t) > \lambda_1$,并且 $uf(w:t) > \lambda_2$,那么单义词单元 $w:t$ 的得分可使用如下公式:

$$score(w:t) = \frac{\Delta(df(w:t))}{1 + df_h(w:t)} + \frac{\Delta(uf(w:t))}{1 + uf_h(w:t)}; \tag{1}$$

否则,$score(w:t) = 0$.

$df(w:t) > \lambda_1$,并且 $uf(w:t) > \lambda_2$ 表明该单义词单元 $w:t$ 在当前时间窗口内,讨论的 *tweets* 和用户都比较多,反映其在当前时间内的重要性;式(1)表示该单义词单元 $w:t$ 的文档频率和用户频率在当前时间窗口和相比前一时间窗口中的增加率,同时反映其"新"和重要的程度.

2.4.2　选取最有可能谈论新闻话题的微博

最后给所有当前时间窗口内的 *tweet* 评分,每个 *tweet* 的得分:

$$score(tweet) = \sum_{n=1}^{N} tfidf(w_i:t_i) \cdot core(w_i:t_i).$$

给评完分的 *tweet*,按照分由高到低排序,选取前面若干个,视为最有可能讨论新闻事件的微博,作为下一步聚类的数据.

2.5　对选取的新闻微博聚类

2.5.1　微博短文本的隐主题空间向量表示和相似度度量

在 2.3.2 节提到的 LDA 建模结果中,结果 2 是一个 $M \times K$ 的矩阵,M 是数据集中文本总数,K 是隐主题个数. 它表示数据集中每个文本生成隐主题的概率,也可看作每个文本在 K 维隐主题空间上的分量值.

本文中隐主题个数 K 选取 200,那么每个 *tweet* 都可表示成一个在 200 维隐主题空间中的向量. 这样有效避免短文本数据稀疏性,导致文本见相似度无法准确度量的问题.

因为每篇文本在 K 个隐主题上的分布都是从同一分布中采样出来的[3],所以,可用 KL 散度来度量两个文本之间的距离.

对于随机变量 X 和 Y,其 KL 散度定义为

$$D_{kl}(X \| Y) = \sum_{n=1}^{N} p(x = n) \log \frac{p(x = n)}{p(y = n)},$$

由于 KL 散度并不具有对称性,即

$$D_{kl}(X \| Y) \neq D_{kl}(Y \| X),$$

所以它并不是很好的距离度量,本文使用它的一个更平滑的具有对称性的变形形式,即 Jensen-Shannon 距离:

$$D_{js}(X \| Y) = \frac{1}{2}[D_{kl}(X \| M) + D_{kl}(Y \| M)],$$

其中 $M = \frac{1}{2}(X + Y)$.

2.5.2　两层的 K 均值和层次聚类的混合聚类

如前所述,微博客如 Twitter 的数据量非常大,即使我们选出最有可能谈论新闻事件的 *tweets*,数据仍非常多. 传统的聚类方法,如层次聚类速度非常慢,根本不适用于大规模数据集. K 均值聚类,则难以提前指定类的数目. K 值太大,讨论同一新闻话题的 *tweets* 可能无法聚入同一个类;K 值太小,讨论不同新闻话题的 *tweets* 势必聚入同一个类中.

所以,本文采用一种两层的混合聚类方法. 首先用 K 均值聚类方法做第一层聚类,选取一个适当大的类数目 K(K 仍应远小于选取出来的微博数量),

这样可充分发挥 K 均值聚类速度快的优点. 然后,对 K 均值聚类的结果,给定合适阈值,再使用层次聚类,直到所有类之间的距离大于该阈值.

3 实验与结果分析

本文针对大规模的微博客数据集,提出基于隐主题挖掘和文本聚类的方法,来实现新闻话题的发现. 主要工作有3步:第1步,使用LDA模型挖掘数据集隐主题信息;第2步,选取最有可能谈论新闻事件的微博客;第3步,用一种混合聚类的方法,将不同微博聚合成一个新闻话题. 实验设定时间窗口的长度为一天.

3.1 评测新闻微博的选择

在Twitter中,转发的 *tweet* 称为 *retweet*. 经验告诉我们,通常无论是普通网页还是博客,有关重要新闻话题往往转载或转发概率较高. 所以,如果本文方法确实能将那些更有可能讨论新闻话题的微博客选取出来,那么在Twitter数据集上选取出来的 *tweets* 中,*retweet* 的比率应该高于平均值. 实验证明了这一点,图3是新闻信息过滤方法的验证,从图3可看出,本文方法能选出那些更有可能谈论新闻话题的微博. 此外图3也说明本文统计单义词单元的频率和变化率的方法是有效的.

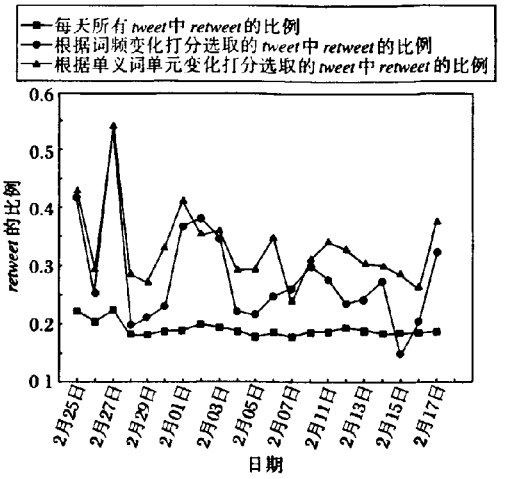

图3 不同方法的 *retweet* 比例

Fig. 3 Retweet proportion of different methods

表2 在3月8日聚类结果中最大的一个类中得分最高的5条 *tweets*

Table 2 5 top scoring *tweets* in biggest cluster on March 8th

作者	*tweet* 文本
SilkCharm	Trivia: Kathryn Bigelow won Best Movie/Best Director for Hurt Locker. Her ex Husband James Cameron, director of Avatar, did not #oscars
LAmovieexaminer	`Precious´star Mo´Nique wins Best Supporting Actress Oscar at the 82nd Academy Awards: http://bit.ly/bc5joW #Oscars
Cocacy	My predictions in the six big categories: Jeff Bridges, Meryl Streep, Mo´Nique, Christoph Waltz, Avatar; kathryn bigelow #theoscars
skynewsbreak	Kathryn Bigelow wins Best Director Oscar for The Hurt Locker – the first woman in history of the Academy Awards to win the accolade
prayoonko	RT @TwitBreakinNews: Christoph Waltz and Mo´Nique have taken the best supporting actor and actress Oscars at the 82nd Academy Awards in Hollywood.

表3 若干天的聚类结果中最大的类中得分最高的 *tweet*

Table 3 Top scoring *tweet* of biggest cluster in several days

日期	作者	*tweet* 文本
Feb. 25	avivao	Dog fight. Lamar claims Obamacare will raise premium prices. Obama claims prices will go down 14-20%, per CBO. #HCSummit #HCS #HCR
Feb. 27	Infidel007	RT @BreakingNews: 54th major aftershock, magnitude 5.0, centered off coast of Bio-Bio, #Chile-U.S. Geological Survey
Mar. 03	NewTechBooks	FAA puts two air traffic control employees on admin leave after teen directs aircraft over JFK: CBS News has repor... http://bit.ly/b8pUwH
Mar. 07	vhernandezcnn	Major development RT @CNNworldgirl: SrPakistan officials tell CNN they have arrested Adam Gadahn, the American-born spokesman for al Qaeda.
Mar. 09	BreakingNews	Update: Ohio State University employee killed, two wounded in campus shooting; suspect in custody-AP
Mar. 11	vivekmadan	Suicide blasts kill 45 in Pakistan's Lahore: Two suicide bombers targeting the Pakistani military killed at least... http://bit.ly/bbofZA
Mar. 15	troyjensen	Plane kills jogger in SC beach emergency landing;http://j.mp/dzTwWb-Damn, that is one series of terrible events!

3.2　混合聚类的结果

首先给出一个已知新闻事件的实例,表明本文方法确能有效发现该新闻话题.在我们的数据集收集期间,第82届奥斯卡颁奖典礼于3月7日晚举行.对于这样一个重要的新闻事件,本文方法确实能有效检测出该新闻话题.

因为本文的时间窗口是按天设置的.所以在3月8日的聚类结果中,最大的一个类就是谈论奥斯卡颁奖典礼这一新闻话题的.其中得分最高的5条tweets如表2所示.

最后列出聚类的一些实验结果.选取每天聚类得到的最大的一个类,并把这个类中得分最高的*tweet*作为该类的新闻话题代表列出.表3是其中若干天的结果,可看出,这些*tweets*谈论的都是重要的新闻话题.

4　结束语

本文在从大规模的微博客短文本数据集中,检测出新闻话题方面首先做了尝试.虽然数据集从Twitter获得,但本文方法同样可应用到其它微博客系统中.文中利用隐主题建模的方法,有效解决短文本集数据稀疏性的问题.选取最有可能讨论新闻话题的微博客的方法,将关注的数据对象规模大大缩小,并在一定程度上排除非新闻博客的干扰.最后,使用一个两层的K均值和层次聚类的混合聚类方法,可快速准确的将所有微博聚集到不同的新闻话题之下.

但是对于本文的工作,仍有可以改进之处.一方面,本文方法并不是实时的.但可通过缩短时间窗口,引入大规模外部数据集,在后台做隐主题挖掘的办法,来实现一个真正实时的微博客新闻发现系统.另一方面,微博由于自身长度的原因,很难全面的描述一个新闻事件,如何选择几篇微博将一个新闻事件完整的描述出来,也是将来工作的方向.

参　考　文　献

[1] Bollegala D, Matsuo Y, Ishizuka M. Measuring Semantic Similarity between Words Using Web Search Engines // Proc of the 16th International Conference on World Wide Web. Banff, Canada, 2007: 757 - 766

[2] Sahami M, Heilman T D. A Web-Based Kernel Function for Measuring the Similarity of Short Text Snippets // Proc of the 15th International Conference on World Wide Web. Edinburgh, UK, 2006: 377 - 386

[3] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation. Journal of Machine Learning Research, 2003, 3: 993 - 1022

[4] Heinrich G. Parameter Estimation for Text Analysis [EB/OL]. [2010-8-10]. http://www.arbylon.net/publications/text-est.pdf

[5] Griffiths T L, Steyvers M. Finding Scientific Topics. Proc of the National Academy of Sciences of the United States of America, 2004, 101(z1): 5228 - 5235

[6] Deerwester S, Dumais S T, Furnas G W, *et al*. Indexing by Latent Semantic Analysis. Journal of the American Society of Information Science, 1990, 41(6): 391 - 407

[7] Hofmann T. Probabilistic Latent Semantic Analysis // Proc of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Berkeley, USA, 1999: 289 - 296

[8] Luo Weihua, Liu Qun, Cheng Xueqi. Development and Analysis of Technology of Topic Detection and Tracking // Proc of the 7th Joint Symposium on Computational Linguistics. Harbin, China, 2003: 560 - 566(in Chinese)
(骆卫华,刘 群,程学旗. 话题检测与跟踪技术的发展与研究// 全国第七届计算语言学联合学术会议论文集. 哈尔滨, 2003: 560 - 566)

[9] Luo Weihua, Yu Manquan, Xu Hongbo, *et al*. The Study of Topic Detection Based on Algorithm of Division and Multi-Level Clustering with Multi-Strategy Optimization. Journal of Chinese Information Processing, 2006, 20(1):29 - 36(in Chinese)
(骆卫华,于满泉,许洪波,等. 基于多策略优化的分治多层聚类算法的话题发现研究. 中文信息学报, 2006, 20(1): 29 - 36)

[10] Hong Yu, Zhang Yu, Liu Ting, *et al*. Topic Detection and Tracking Review. Journal of Chinese Information Processing, 2007, 21(6): 71 - 87(in Chinese)
(洪 宇,张 宇,刘 挺,等. 话题检测与跟踪的评测及研究综述. 中文信息学报, 2007, 21(6): 71 - 87)

[11] Allan J, Papka R, Lavrenko V. On-Line New Event Detection and Tracking // Proc of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Melbourne, Australia, 1998: 37 - 45

[12] Allan J, Carbonell J, Doddington G, *et al*. Topic Detection and Tracking Pilot Study Final Report // Proc of the DARPA Broadcast News Transcription and Understanding Workshop. Landsdowne, USA, 1998: 194 - 218

[13] Phan X H, Nguyen L M, Horiguchi S. Learning to Classify Short and Sparse Text & Web with Hidden Topics from Large-Scale Data Collections // Proc of the 17th International Conference on World Wide Web. Beijing, China, 2008: 91 - 100

基于隐主题分析和文本聚类的微博客中新闻话题的发现

作者： 路荣， 项亮， 刘明荣， 杨青， LU Rong， XIANG Liang， LIU Ming-Rong， YANG Qing
作者单位： 中国科学院自动化研究所模式识别国家重点实验室 北京100190
刊名： 模式识别与人工智能 ISTIC EI PKU
英文刊名： Pattern Recognition and Artificial Intelligence
年，卷(期)： 2012,25(3)

参考文献(13条)

1. Bollegala D;Matsuo Y;Ishizuka M Measuring Semantic Similarity between Words Using Web Search Engines 2007

2. Sahami M;Heilman T D A Web-Based Kernel Function for Measuring the Similarity of Short Text Snippets 2006

3. Blei D M;Ng A Y;Jordan M I Latent Dirichlet Allocation 2003

4. Heinrich G Parameter Estimation for Text Analysis 2010

5. Griffiths T L;Steyvers M Finding Scientific Topics 2004(z1)

6. Deerwester S;Dumais S T;Furnas G W Indexing by Latent Semantic Analysis 1990(06)

7. Hofmann T Probabilistic Latent Semantic Analysis 1999

8. 骆卫华;刘群;程学旗 话题检测与跟踪技术的发展与研究 2003

9. 骆卫华;于满泉;许洪波 基于多策略优化的分治多层聚类算法的话题发现研究[期刊论文]-中文信息学报 2006(01)

10. 洪宇;张宇;刘挺 话题检测与跟踪的评测及研究综述[期刊论文]-中文信息学报 2007(06)

11. AUan J;Papka R;Lavrenko V On-Line New Event Detection and Tracking 1998

12. Allan J;Carbonell J;Doddington G Topic Detection and Tracking Pilot Study Final Report 1998

13. Phan X H;Nguyen L M;Horiguchi S Leaming to Clarify Short and Sparse Text & Web with Hidden Topics from Large-Scale Data Collections 2008