基于转发图的微博事件主题摘要方法

赵 斌,吉根林,曲维光,顾彦慧

(南京师范大学计算机科学与技术学院,江苏 南京 210023)

[**摘要**] 自动摘要是自然语言处理中研究文本主题提取的重要课题. 传统的摘要研究侧重于新闻、Web 网页和博客等长文本的主题提取. 本文关注以微博为代表的短文本的主题摘要,提出基于图结构的微博主题区域划分方法,并采用 LDA 方法提取微博热点事件的主题信息. 最后,通过可视化方式展现主题内容在微博转发中的变化.

[**关键词**] 主题摘要,微博,可视化

[**中图分类号**]TP391.1 [**文献标志码**]A [**文章编号**]1001-4616(2014)01-0066-05

Topic Summarization of Microblog Events Based on Retweeting Graph

Zhao Bin, Ji Genlin, Qu Weiguang, Gu Yanhui

(School of Computer Science and Technology, Nanjing Normal University, Nanjing 210023, China)

Abstract: Auto summarization is an important research field towards topic extraction in natural language processing. Traditional researches are proposed to generate summaries of long-text corpus, such as news, web documents and blogs, but seldom focus on topic summarization of a short-text corpus like microblogs. In this paper, we propose a topic-oriented partitioning method towards microblog retweeting graphs, and utilize LDA to generate topic summaries of hot events. Finally, we present the visualization graphs of topic summaries remarkably emphasizing the insight behind the evolving events.

Key words: topic summarization, microblog, visualization

微博是当前流行的社交网络平台,它既可以为个人用户记录生活点滴、分享兴趣爱好,也可以为商业用户进行品牌推广、活动策划和形象宣传等营销活动. 不同于传统的互联网应用,其独特的媒体特性赋予了用户更多的话语权,用户既是信息的接收者,也是信息的发布者和传播者. 每当热点事件发生,众多网络用户借助微博平台参与讨论,发表个人观点和表达自身关切. 伴随热点事件持续发展,个人表达的意见和评论逐渐汇聚融合形成群体观点,它代表公众在微博平台上的群体性意见,是社会舆论的重要组成. 因而,提炼微博热点事件中的观点信息,对于监测网络舆论,预测舆情导向和回应社会关切等多方面都可有积极作用.

微博在提供丰富而多样的信息的同时,也使得用户面临海量数据的选择问题. 如何在规模庞大的微博数据集中提炼出有价值的重要信息,是微博主题摘要研究需要解决的重要问题.

随着社交媒体的快速发展,学术界开始研究社会化媒体的摘要方法. 文献[1]研究博客的主题摘要(Topic Summarization)技术,从博客的评论中获得代表性词项,进而选择包含代表性词的代表性语句用来表达主题信息. 文献[2]分析传统的媒体和新兴社会化媒体的特点,结合两者优势,采用基于非监督的主题模型方法发现具有代表性的且在两种媒体间具有互补性的内容信息,以此生成事件摘要. 文献[3]研究新闻评论的摘要问题,发现现有的排序机制无法有效提供评论摘要,因而提出基于评论上下文的主题模型对评论进行聚类计算,而后采用 Maximal Marginal Relevance(MMR)和 Rating & Length(RL)排序算法发现代表性评论. 文献[4]研究 Twitter 中的情感摘要问题,提出以实体(Entity)为中心面向主题的摘要方法. 文

收稿日期:2013-08-10.

基金项目:江苏省高校自然科学基金(13KJB520014)、江苏省自然科学基金(BK2011005)、国家自然科学基金(61272221)、江苏省社科基金(12YYA002).

通讯联系人:赵斌,博士,讲师,研究方向:数据挖掘技术及应用. E-mail:zhaobin@njnu.edu.cn

献[5]提出研究 Twitter 上下文摘要的新问题,为此利用用户的交互信息构建用户的影响力模型,提供有效的上下文摘要信息. 主题摘要的可视化研究方面工作比较少,文献[6]设计并实现了一个交互式、可视化的文本分析工具 TIARA,采用可视化技术分析大规模文本集合中的主题信息.

已有的研究主要以文本为主要目标,采用自然语言处理的方法提炼主题信息. 然而,微博是一种具有传播性和社交性的媒体,微博用户通过转发消息的方式参与热点事件讨论,表达自我观点. 因而,转发构成的链接关系对计算微博消息间的主题相关性有明显的影响作用. 本文假设相邻的转发消息同属于相同的会话区域,消息的主题应该接近. 因此,微博中主题摘要的形成应该考虑用户转发行为的影响.

本文首先收集了真实微博事件数据开展研究,经过预处理构建由消息文本和转发关系形成的转发图,结合 LDA 方法和连通分支算法形成转发图的话题划分,采用可视化方法展现主题摘要信息在用户交互中传播变化的情况.

1　数据集与统计分析

为了研究热点事件的主题摘要方法,本文以“郭美美炫富”事件为研究对象,使用新浪微博 API,通过搜索引擎和意见领袖收集了相关微博,构成了本文后续的实验数据集(GMM),该数据集包含 253 980 条微博,99 452 位参与用户和 23 156 个转发簇.

为了对热点事件数据有一个全面了解,本文对微博数据在消息文本的用户行为两方面做了统计分析.

图 1 展示了 GMM 数据集中不同长度微博的分布,据统计 36% 的微博字数不超过 20 字,数量最多. 19% 的微博字数大于 120 字. 由于微博长度过短无法从中提炼出高质量的主题信息,因而在微博预处理中过短微博将被过滤,但是转发路径仍然被保留. 图 2 展示了用户转发热点微博消息的情况,其中 60% 的用户在热点事件讨论中仅转发一次,此类用户虽然参与讨论积极程度有限,但却是热点事件在微博中传播的主体. 图 3 展示了转发簇(即源自同一起始微博的转发微博集合)规模方面的统计情况,17% 的转发簇包含至少 10 条微博,转发次数与讨论强度成正比. 图 4 展示了转发簇时间跨度的统计情况,其中 91% 的转发簇时间跨度为 1 天.

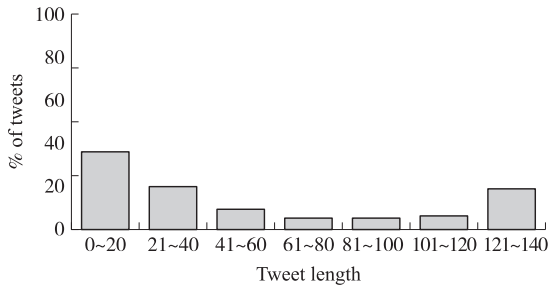

图 1　微博长度分布图

Fig. 1　Tweet length distribution

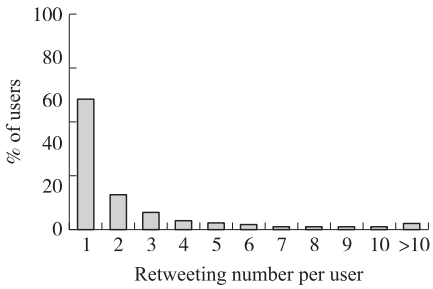

图 2　微博用户分布图

Fig. 2　Weibo user distribution

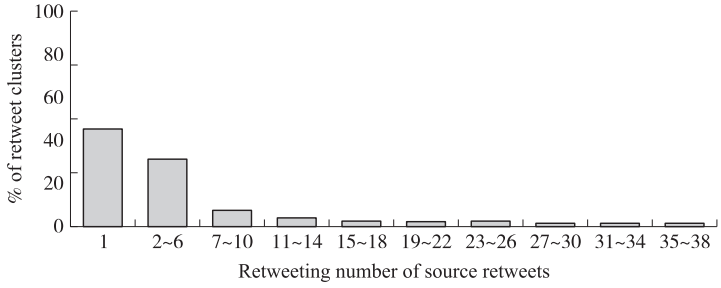

图 3　微博转发情况分布图

Fig. 3　Retweeting distribution

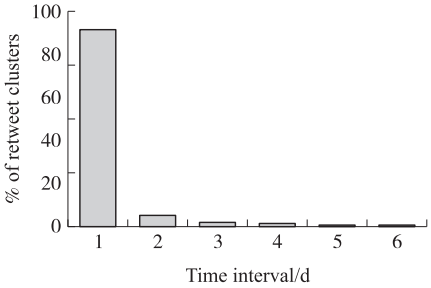

图 4　转发微博簇的时间跨度分布图

Fig. 4　Duration distribution of retweeting clusters

通过对 GMM 数据集的统计可以发现,热点事件在微博平台上传播迅速,用户参与规模巨大,积极参与的用户数虽然在总用户数中比例不高,但是贡献的评论意见较高. 所以,选择用户参与度高的微博转发簇作为主题摘要研究的对象.

2　提取热点事件主题摘要的基本框架

微博热点事件中主题摘要的挖掘过程,包括以下 4 个阶段:

(1)数据收集:根据热点事件关键字或意见领袖账号,收集热点事件微博消息(包括转发消息).

(2)预处理:首先过滤掉无评论微博,对于有评论微博采用自然语言处理工具进行处理,并根据转发消息中残存的转发关系恢复出消息传播路径.

(3)主题提取:采用主题摘要提取方法(如 LDA)得到主题信息.

(4)摘要可视化:采用可视化的方法,基于转发图展示主题摘要结果在转发互动中的变化.

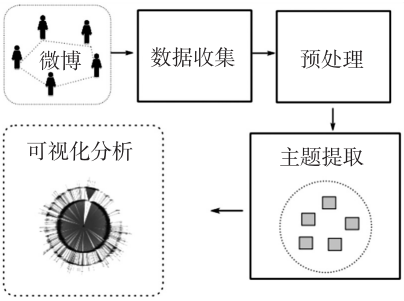

图 5　微博主题摘要系统结构图

Fig. 5　Framework of microblog topic summarization system

3　基于转发关系的主题摘要方法

3.1　问题描述

设微博转发有向图 $\boldsymbol{G}(V,E)$,V 是微博转发消息的集合,$V=\{v_1,v_2,\cdots,v_n\}$,E 是转发关系的集合,$E=\{\boldsymbol{e}_{ij}\mid v_i\in V,v_j\in V\}$,$\boldsymbol{e}_{ij}$为有向边,表明微博 v_j 转发了 v_i. H 是基于微博转发图 $\boldsymbol{G}$ 的话题划分,$H=\{H_i\mid H_i\subseteq \boldsymbol{G}\wedge \bigcup_i V(H_i)=V\}$,$H$ 由转发微博连通子图 H_i 构成.

基于转发图的主题摘要需要解决的问题是,如何按照事件主题合理划分转发微博图 $\boldsymbol{G}$,即得到符合事件主题的划分 H,最后通过处理 H_i 生成对应的主题摘要.

3.2　转发路径恢复算法

本文提出的主题摘要方法是基于微博转发关系的,但是采用微博的应用程序编程接口(API)收集的消息没有保留转发关系,仅在 140 字的微博文本中通过“//”标记表达了部分转发关系. 所以,需要根据转发文本内容尽可能地恢复原有的转发路径.

转发路径恢复方法描述如下:

输入:微博消息集合 V

输出:微博转发关系集合 E

(1)对于指定微博 v_i,将其与其他微博 v_j 进行右完全匹配. 如果匹配成功,则表明 v_j 转发了 v_i,即添加 $\boldsymbol{e}_{ij}$到 E 中.

(2)如果右完全匹配不成功,则可以考虑近似匹配的方式. 取微博 v_j 中“//”标记之后的部分,与其他微博的起始部分匹配,找出匹配文本长度最长的微博 v_i,则表明 v_j 转发了 v_i,即添加 $\boldsymbol{e}_{ij}$到 E 中.

需要说明的是,由于获取数据的方式决定了无法收集到全部的转发数据,因而,为了保证所有的微博都包含在转发路径中,如果一条微博无法匹配则将该微博默认为转发自初始转发微博.

3.3　基于转发关系的主题提取

LDA 是主题建模的常用工具,于 2002 年被首次提出[7]. 本文采用该方法提取微博消息的主题摘要. 通常,微博事件的主题可以表示为词项的概率分布. 主题中的词项概率越大,表示该词项代表该主题的可能性越高. 同样,微博消息本身也可以表示为主题的概率分布,相关主题按照概率由大到小排列,概率越大表明该主题成为所在微博主题的可能性越大. 采用 LDA 的方法为每条微博生成的主题向量为 $\boldsymbol{t}_i=(t_{i1},t_{i2},\cdots,t_{in})$,$t_{ij}(1\leqslant j\leqslant n)$是微博消息 v_i 符合的主题编号,且按照主题概率从大到小排列,得到主题向量是度量微博消息主题相关性的前提.

本文假设转发图中相邻的转发微博消息同属于相同话题区域的可能性较大. 通过微博间主题相似度计算,设定阈值过滤转发关系集合 E,最终得到按照主题划分的连通分支,即微博转发图 $\boldsymbol{G}$ 的 H 划分.

由于微博数据长度短、微博语言不规范,因而微博消息的相似度计算难度较大. 本文并不直接针对文本内容进行相似度计算,而是对 LDA 方法生成的主题向量进行相似性度量.

主题提取算法描述如下:

输入:微博转发图 $\boldsymbol{G}$,参与相似度计算的分量数 k,相似度阈值 θ;

输出:主题划分 H 及其主题摘要

(1) $E'=\varnothing$;

(2) for $v_i \in V$ do {

(3)　采用 LDA 计算 v_i 的主题向量 $\boldsymbol{t}_i=(t_{i1},t_{i2},\cdots,t_{in})$;

(4)　截取 $\boldsymbol{t}_i$ 中最大的 k 个分量排序后生成 $\boldsymbol{t}'_i=(t'_{i1},t'_{i2},\cdots,t'_{ik})$;

(5)}

(6) for $\boldsymbol{e}_{ij} \in E$ do {

(7)　if $\mathrm{Sim}(\boldsymbol{t}'_i,\boldsymbol{t}'_j)>\theta$ then　//相似度计算;

(8)　　$E'=E' \cup \{\boldsymbol{e}_{ij}\}$

(9)}

(10) 根据 E' 生成连通图,得到 $\boldsymbol{G}$ 的主题划分 H;

(11) 采用 LDA 针对每个 H_i 生成主题摘要.

4　实验

为了研究“郭美美”热点事件(http://www.baike.com/wiki/郭美美)的主题摘要,针对 GMM 数据集采用 LDA 提取了典型的主题摘要信息,基本与该事件的描述情况一致,如表 1 所示. 采用已有的方法可以展示整个事件的主题摘要信息. 然而微博是一种传播性媒体,用户在交互讨论中逐渐形成群体意见. 因此,上述方式无法展示出事件主题演化的过程. 为了验证本文所提出方法,从 GMM 数据集中挑选出一个含有 95 条微博的转发簇,如图 6 所示.

实验设定微博主题向量分量数为 10,对于每条主题向量取主题概率最大的前 5 条参与相似度计算. 比较余弦相似度和 Jaccard 相似度生成微博主题摘要的效果. 图 7 为余弦阈值为 0.16 的转发图划分效果,图 8 为余弦阈值为 0.15 的转发图划分效果. 可以发现,虽然划分效果在转发图上明显出现,但由于余弦值太低,已经没有实际相似的意义了. 所以采用余弦度量微博主题相似性是不可行的.

表 1　“郭美美炫富”事件典型主题摘要信息

Table 1　A snapshot of topic summaries on GMM event

编号	主题摘要
1	炫富、名车、拜金、玛莎拉蒂
2	红十字会、捐款、慈善、不能信
3	后爸、潜规则、干女儿、贪污、腐败
4	政府、建党伟业、信任、党、腐败

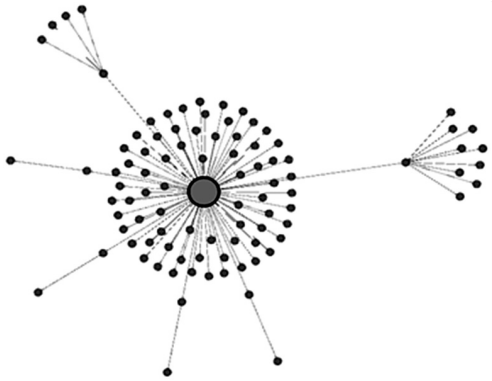

图 6　微博转发簇的转发关系图

Fig. 6　A microblog retweeting graph of GMM event

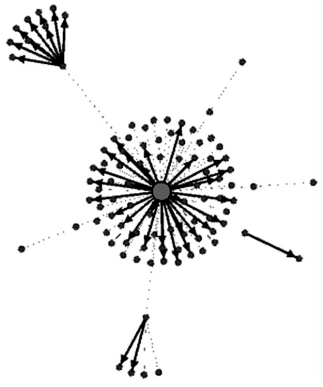

图 7　余弦阈值为 0.157 的转发路径图

Fig. 7　Retweeting graph where the cosine threshold is 0.157

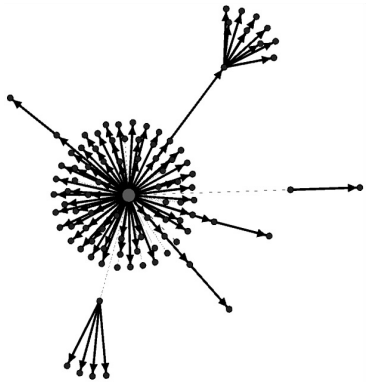

图 8　余弦阈值为 0.15 的转发路径图

Fig. 8　Retweeting graph where the cosine threshold is 0.15

采用度相似生成微博主题摘要的效果如图 9 ~ 图 11 所示.

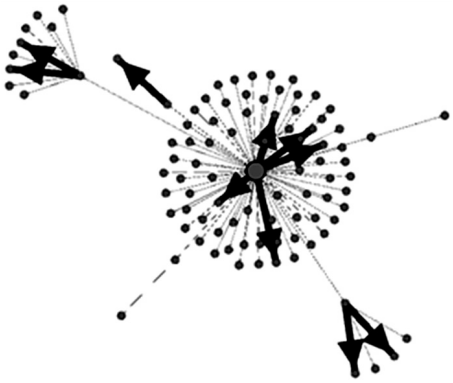

图 9　Jaccard 相似度阈值为 0.8 的转发路径图

Fig. 9　Retweeting graph where the Jaccard coefficient threshold is 0.8

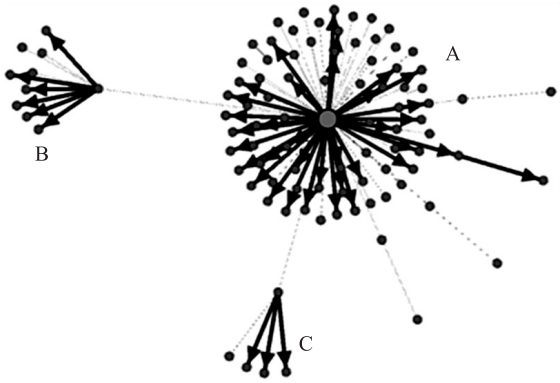

图 10　Jaccard 相似度阈值为 0.7 的转发路径图

Fig. 10　Retweeting graph where the Jaccard coefficient threshold is 0.7

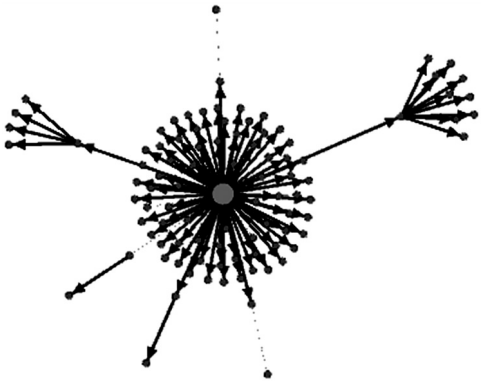

图 11　Jaccard 相似度阈值为 0.6 的转发路径图

Fig. 11　Retweeting graph where the Jaccard coefficient threshold is 0.6

图 9 ~ 11 中,Jaccard 相似度阈值取为 0.7 话题划分效果最佳,分为 ABC 三个话题区域,表 2 中列出了每个话题区域代表主题摘要的主题关键字.

表 2　图 7 中话题子图的主题摘要

Table 2　Topic summaries of connected components in Fig. 7

分区	主题关键词
A	红十字会 郭长江 回家 吃饭
B	扑朔迷离 糗 玩 大
C	孩子 胡戈

5　结语

本文主要研究微博热点事件的主题摘要方法. 和现有的研究不同,本文关注基于链接关系图生成微博文本的主题摘要方法. 以“郭美美炫富”事件微博数据为研究对象,采用 LDA 方法生成微博消息的主题向量,最后采用可视化的方式对不同相似性计算的效果进行比较. 实验结果证明了本文提出的基于转发图的主题摘要方法在微博热点事件分析中的有效性和可行性.

[参考文献]

[1] Hu M, Sun A, Lim E. Comments-oriented blog summarization by sentence extraction[C]//Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management(CIKM'07). Lisbon, 2007: 901-904.

[2] Gao W, Li P, Darwish K. Joint topic modeling for event summarization across news and social media streams[C]//21st ACM International Conference on Information and Knowledge Management(CIKM'12). Maui, 2012: 1 173-1 182.

[3] Ma Z, Sun A, Yuan Q, et al. Topic-driven reader comments summarization[C]//21st ACM International Conference on Information and Knowledge Management(CIKM'12). Maui, 2012: 265-274.

[4] Meng X, Wei F, Liu X, et al. Entity-centric topic-oriented opinion summarization in Twitter[C]//The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD'12). Beijing, 2012: 379-387.

[5] Chang Y, Wang X, Mei Q, et al. Towards Twitter context summarization with user influence models[C]//6th ACM International Conference on Web Search and Data Mining(WSDM'13). Rome, 2013: 527-536.

[6] Liu S, Zhou M X, Pan S, et al. Interactive, topic-based visual text summarization and analysis[C]//Proceedings of the 18th ACM Conference on Information and Knowledge Management(CIKM'09). Hong Kong, 2009: 543-552.

[7] Blei D, Ng A, Jordan M. Latent dirichlet allocation[J]. Journal of Machine Learning Res, 2003, 3(5): 993-1022.

[责任编辑:严海琳]