基于 Lex-PageRank 的微博摘要优化方法

朱明峰 叶施仁 叶仁明
(常州大学信息工程学院 常州 213164)

摘 要 当前,由于全民自媒体兴起而引发了巨大的舆情危机,如何高效快速地从海量的碎片化信息中发现热点并抽取实用信息成为一项重大的挑战。在此背景下,提出一种基于 Lex-PageRank 的微博摘要优化方法,在该方案中,以聚类结果作为实验数据,从微博影响力周期的时间特性和权重属性考虑,提出改进的 Lex-PageRank 算法,从聚类结果中抽取若干文本组织生成摘要。在新浪微博数据基础上进行的对比实验表明,本方案可以有效地从大量文本中提取出关键信息。

关键词 微博,时间特性,权重属性,Lex-PageRank 算法
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.9.052

Extract Summarization Method Based on Lex-PageRank in Chinese Microblog

ZHU Ming-feng YE Shi-ren YE Ren-ming
(School of Information Science & Engineering, Changzhou University, Changzhou 213164, China)

Abstract In recent years, since the rise of personal-media caused a huge public opinion crisis, how to discover hot topics from the fragmentation of the mass microblogging information and extract useful information has become a major challenge. In this background, we proposed an extract summarization method based on improved Lex-PageRank algorithm. In this program, we make a simple clustering and use these clustering results as experimental data. With due consideration for the time characteristics of microblog influence cycle and weight attribute, the improved Lex-PageRank algorithm is combined with MMR algorithm to get many texts from clustering results to generate a summary. The experiment based on Sina Weibo indicates that our method can extract critical information effectively from mass texts.

Keywords Microblog, Time characteristics, Weight attribute, Lex-PageRank algorithm

1 概述

互联网的蓬勃发展大大提升了微博的影响力,近年来微博的发展使其在网民生活的各方面占据着重要位置。微博平台在信息产生和发展的过程中往往面临严重的数据量过载和信息碎片问题[1]。如何使用户轻松阅读到全面、客观的微博信息,是一个非常值得研究的课题[2]。因此,利用计算机组织生成摘要,帮助用户进行阅读是十分必要的。

1958年 Luhn 提出了一种基于高频词打分自动文本摘要方法[3],其依据统计关键词、词频或句子出现频率从文章中提炼最具代表性的若干句子形成摘要。此方法的缺点在于,词频基础上的统计对同义词不敏感,对文本特征的描述说服力不足;此外,这一类的机械摘要内容往往包含大量冗余信息,导致文档覆盖率低下,难以得到理想的阅读效果。美国科罗拉多州立大学的 Jugal Kalita 从语言学角度理解文档集合[4],在大量重复语料的前提下,利用核心词汇权重分布逐渐匹配的方式对文章进行句法分析和语义分析,自动生成摘要。此方法虽然效果优越,但不能适应个人微博信息零散的特性。

本文的基本思想是以文本的特征选择为出发点,以微博词汇的特征权重为基础,首先获取足够数量的文本集,经过预处理、文本向量化、聚类形成若干子主题,并在各个子主题下利用改进的 LexPageRank 选取若干句子生成对应主题摘要[5]。本文研究的目标是使摘要的句子尽可能完整地表达原始文档的信息。其中,文摘句子的优化选择是本方案的核心环节[6]。另一方面,使摘要文本在表述简洁的同时既能保证信息覆盖率,又能最大限度地反映主题内容,是提高文摘质量的一个重要的方面,也是文本研究的重要内容。

2 微博文档子主题生成

微博的子主题由若干人的微博文档集合经过聚类形成,这些子主题分别代表这些人关注的信息和自己的态度等,而摘要的生成则使子主题蕴含的信息更简洁、客观[7]。为了获得相对纯净的聚类文本,首先要对采集的数据进行预处理,主要包括去表情符号、以标点符号(句号、感叹号、问号)为单位切分文档、对文档进行分词、去停用词等。

2.1 文本向量化

本文中文本聚类的首要任务是对博文进行向量化表示。本文使用经典的文本向量化表示方法——向量空间模型(Vector Space Model, VSM)。文本的特征项及其权重表示向量的每一维,特征项的权重按 TF-IDF(Term Frequency-In-

到稿日期:2015-08-26 返修日期:2015-11-02 本文受国家自然科学基金(61272367)资助。
朱明峰(1990-),男,硕士,主要研究方向为数据挖掘;叶施仁(1971-),男,博士,高级工程师,主要研究方向为数据挖掘;叶仁明(1990-),男,硕士,主要研究方向为数据挖掘。

verse Document Frequency)[8]方法来计算：

$$w(t_i,d)=\frac{tf(t_i,d)\times\text{loglog}(\frac{N}{n_i}+0.01)}{\sqrt{\sum_{t_i\in d}[tf(t_i,d)\times\text{loglog}(\frac{N}{n_i}+0.01)]^2}} \quad (1)$$

其中，$w(t_i,d)$为特征项 t_i 在文档 d 中的权重；$tf(t_i,d)$为特征项 t_i 在文档 d 中的词频；N 为文档总数；n_i 为文档中出现特征项 t_i 的文本数，分母为归一化因子，即文档 d_j 的向量化表示为：

$$d_j=(t_{j1}:w_{j1},t_{j2}:w_{j2},\cdots,t_{jm}:w_{jm}) \quad (2)$$

其中，t_{jm}表示第 j 个文档的第 m 个特征项，w_{jm}表示该特征项的权重，m 同时表示向量中特征项的个数。

2.2 文本降维

微博博文作为信息传播和展示的载体，蕴含着丰富多样的内容，在庞大数量的微博用户的驱动下，短时间内往往会产生大量的博文。因此，文本向量化处理之后得到的结果往往是高维数据集，而高维数据稀疏性的特性使得通过计算文本之间的距离来聚类变得异常困难。针对上述问题，本文将进行降维处理。

通常博文句在长度上并不统一，因此在向量化处理后会产生大量补足的零元素，为节省存储空间和提高运算效率，本文采用系数矩阵来进行降维处理。构建的系数矩阵包含以下3个内容：句子序号(SN)、向量位置(VL)、特征词权重(TF-IDF)。微博文档具体表示为：

$$\begin{bmatrix} SN_1 & VL_j & TF\text{-}IDF_j \\ SN_2 & VL_k & TF\text{-}IDF_k \\ \vdots & \vdots & \vdots \\ SN_i & VL_n & TF\text{-}IDF_n \end{bmatrix} \quad (3)$$

其中，SN_i 表示第 i 条句子，VL_n 表示第 n 个特征项的向量位置，$TF\text{-}IDF_n$ 表示第 n 个特征项在语料库中所占特征权重。在实验中，依据此方法存储的博文数据显著降低了计算过程中的复杂度和时间开销。

2.3 划分子主题

在预处理的基础上对微博文档集合进行子主题的划分。Cure算法在识别不规则形状簇的过程中效果出众，在处理异常数据时又表现出健壮性。因此，使用Cure算法对句子进行聚类，以实现微博文档子主题的划分，即聚类数等于子主题数。

本文中Cure算法[9]的基本思想是把每个句子当作一个簇，合并距离最近的簇，直到簇的数目达到指定数目为止。

Cure聚类算法的步骤如下：

Step1 从微博文档数据集 D 中抽取一个随机样本 S，给定簇数目 K(这里的 K 值由文献[8]提供的类内平方误差和准则函数得到)，即下文聚类子主题数；

Step2 计算文档之间的余弦相似度，然后采用对该相似度取对数的方法将其转换成文本之间的距离，其中对数底数在(0,1)范围内取值，并对分割的局部进行聚类，文本间距离表示为：

$$dis(d_i,d_j)=\log_a(sim(d_i,d_j))=\log_a[\frac{d_i*d_j}{|d_i|\times|d_j|}],\quad a\in(0,1) \quad (4)$$

Step3 给定收缩点数量，随机取样并剔除相似度低的孤立点，合并距离最近的簇；

Step4 根据收缩因子，博文句向簇中心移动，直至收缩完成，将数目明显少的簇作为孤立点剔除，更新簇，迭代Step2；

Step5 迭代Step2—Step4，并用相应的类标签来标记数据，直到簇数目不变。

此处聚类生成的子主题相当于为下文摘要起到预先过滤的作用。

2.4 主题生成示例

从第4节实验中随机选取50句博文进行聚类，程序运行结果如图1所示。

主题	数量
中国hadoop技术峰会	9
智能手机+移动互联网改变生活	12
美国Costco超市低毛利如何致胜	10
春节人口迁徙图	8
学堂在线选课人次破百万	13

图1 主题生成示例图

3 句子选择方法的研究

在子主题形成以后，生成摘要还需要解决两个问题：1)在各个子主题中选择摘要句子；2)对摘要句子进行冗余处理。

3.1 基于LexPageRank的优化选择

从子主题中自动抽取文摘是通过选取子主题文本中一组最重要的句子实现的。本文选取摘要句的关键在于定量地表示句子的重要程度。密歇根大学的Gunes Erkan和Dragomir R Radev提出的LexRank算法[10]利用基于随机图形表示的方法作为度量文本集合之间的相关性的标准，该思想的核心为：若一个句子与众多其他句子相似，那么此句话就可能是重要的。

本方案中，首先利用上文余弦距离方法计算子主题下句子之间的相似度，如果两句之间的相似度大于给定的阈值(这里的阈值取决于子主题下句子相似性的均值)，就将这两个句子语义定性为相关并将其连接起来。根据此方法，得到一个关于权重的图 $G=(V,E)$，图中的每个节点代表文本 V 中的一个句子，而 E 中的每条边表示节点之间的相似性。在该方法中，用节点与 $v(v\in V)$相连的边的数目 d 来表示所对应句子包含信息的重要程度，即 d 越大重要性越高，反之亦然。

本文通过计算句子间的相似度构建图 G，对句子相似性进行打分并从高到低排序编号，对句子打分的结果表述为：此句超过相似度阈值的句子数和此主题包含句子总数的比值。最后，针对各句分值进行系数校正，系数 $PR(p_i)$为结合时间衰减和权重因子两方面使用PageRank算法[11]计算出各句的影响力，取值过程见式(7)；校正过程表示为句子分值与校正系数的乘积。据此改进的PageRank算法选取 n 条影响力最大的中心语义句候选入文摘集合。

时间衰减因子：据新浪微博官方调查报告，新浪核心用户最喜欢的功能是“获取最新消息”，占所有功能的23.24%，据此提出时间衰减因子。在博文的生命周期里，博文活跃的时间与其影响力呈现出一定的相关性，即博文发布时间越久，其包含的信息实时性越低，信息量越少，影响力也就越低，反之亦然。本文对每篇博文做时间标记，定义时间衰减因子如下：

$$\delta(t)=\begin{cases} -\frac{\Delta t^{\frac{5}{2}}}{43988}+1, & 0\leqslant\Delta t\leqslant 70 \\ 0, & \Delta t\geqslant 70 \end{cases} \quad (5)$$

其中，$\Delta t=(t-tl)$，t 为当前时间，tl 为博文发布时间，Δt 以小

时为单位。图 1 示出了时间因子的衰减曲线。

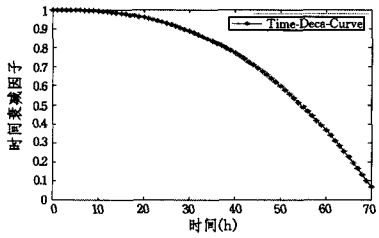

图 2　时间因子的衰减函数曲线图

图 2 中横轴表示时间间隔，单位为小时；纵轴表示与时间相对应的衰减因子值。从图中可以明显观察到，与当前时间间隔越久，时间衰减因子越小：距当前 30h 以内，博文影响力相对充分，30－70h 显著减弱，70h 以后趋向于 0 即影响力失效。

权重因子：一篇博文拥有众多的特征，包括评论数、转发数、粉丝数等。新浪根据其数量规模衡量博主影响力，比如明星、大 V 账号，他们的言论具有突出的公信力和影响力。据此特性，提出决定微博影响力的两个指标：显在影响力和潜在影响力。显在影响力：转发数、评论数、点赞数，具体表示为博文在一定的时间段内被转发、被评论或者被点赞的总次数；潜在影响力：微博博主的粉丝数与该微博所有转发层级转发者的粉丝数之和。本文中定义的权重因子具体描述为：

$$\omega^i=\frac{\sum_{j=1}^{3}x_j^i}{\sum_{i=1}^{n}\sum_{j=1}^{3}x_j^i}\times\frac{y^i}{y}\times n^2 \tag{6}$$

其中，i 表示 LexRank 排序后句子所在网页编号，j 从 1 到 3 分别对应微博的转发数、评论数、点赞数，x_j^i 表示 i 网页博文的转发数、评论数或者点赞数，y 表示包含各层转发粉丝数集合。

综合时间衰减因子，权重因子的 PageRank 算法表示为：

$$PR(p_i)=[\frac{1-d}{N}+d\sum_{p_j\in M(p_i)}\frac{PR(p_j)}{L(p_j)}]\times\delta(t_i)\times\omega^i \tag{7}$$

其中，p_i 表示被研究的页面；d 表示阻尼系数，通常取值区间为[0,1,0.2]；$M(p_i)$表示 p_i 链入页面数量；$L(p_j)$是 p_j 链出页面数量；N 是页面总数。

本文中 PageRank 值是一个特殊矩阵中的特征向量，具体表示为：

$$PR(p_i)=\left|\begin{bmatrix}\frac{(1-d)}{N}\\ \frac{(1-d)}{N}\\ \vdots\\ \frac{(1-d)}{N}\end{bmatrix}+d\begin{bmatrix}\ell(p_1,p_1) & \cdots & \ell(p_i,p_N)\\ \vdots & \ddots & \vdots\\ \ell(p_N,p_1) & \cdots & \ell(p_N,p_N)\end{bmatrix}PR(p_i)\right|\times\delta(t_i)\times\omega^i \tag{8}$$

如果网页 i 有指向网页 j 的一个链接，则 $\ell(p_i,p_j)=1$，否则 $\ell(p_i,p_j)=0$。

考虑微博独有的时间特性和权重特征而设计的 PageRank算法有效弥补了 PageRank 对新页面不敏感的缺陷，优化了候选摘要句对应网页影响力的排名，进而从网页影响力的角度调整了 LexRank 候选摘要句排序，但该方法在信息冗余方面存在瑕疵。

3.1.1　句子选择示例

从 2.4 节中取主题“美国 Costco 超市低毛利如何致胜”进行句子优化选择，程序运行结果如图 3 所示。

主题	美国 Costco 超市低毛利如何致胜
NO. 1	Costco 的成功并非无法复制，以会员费形成资金流盈利的模式，诸如美容、餐饮或其他服务类。
NO. 2	在 Costco 内产品高质量低价格的驱动下，Costco 会员有着超高的忠诚度。
NO. 3	每个小的细分商品品类，在 Costco 只有一到两种选择，Costco 会选择他们认为有“爆款”潜质的商品上架。
NO. 4	Costco 能削减的地方可能就是这里，砍掉一切费用，定位顾客群明确，只挑选质量好价格好的合作商。
NO. 5	面对外部供应商，如果他这家企业在别的地方定的价格比在 Costco 的还低，那么它的商品将永远不会再出现在 Costco 的货架上。
NO. 6	这两条严格地执行下来，才造就了 Costco 商品的低价，平均的毛利率只有 7%，而一般超市的毛利率会在 15%－25%。
NO. 7	目前，Costco 在全球有 671 家仓库，美国作为大本营，占到了 70%，共有 474 家。
NO. 8	其实 Costco 不来中国是非常正确的选择，因为来了肯定是无法维持它超低价格的优势。
NO. 9	雷军对此曾深有感触，三年前自己和金山一帮高管去美国，CEO 张宏江一下飞机就租辆车直奔 Costco。
NO. 10	顺便说一下 Costco 的经营理念等都是我最欣赏的一家了。

图 3　句子选择示例图

3.2　候选摘要句的冗余处理

通常，内容高度相似、意义重合的句子会徒增阅读时间并使读者获取信息的效率低下，因此，本文视此类句子为冗余信息。消除摘要中冗余信息的过程就是在一堆相似信息中提取出差异最大的句子，简而言之，子主题下的摘要句差异越大，其蕴含的信息量越丰富。本文改进卡耐基梅隆大学的 Jade Goldstein 提出的 MMR(Maximal marginal relevance)[12] 方法，在选择摘要内容时，利用余弦距离方法获得的相似度阈值来判别冗余信息，从相似度超过一定阈值的候选文摘句子中随机选择一句加入最终摘要，剩余句视为冗余信息，详细描述为：优先选取上文优化排序得到的首条文摘句作为包含信息量最多的一个句子加入摘要，在计算文摘句子间相似度时加入惩罚因子，对句子长度进行归一化，以避免文档长度造成的偏差，最后迭代选择若干与前者相似性最小的句子构成子主题摘要句。改进的 MMR 算法的步骤为：

1)获取 LexPageRank 处理后的文档集，记为 d_r，其中 r 表示包含的句子数；

2)将 LexPageRank 方法优化选择出的首条文摘句加入最终的摘要，记为 $\max W_j$；

3)选取 $\max W_j$ 作为第一个句子，从子主题中去除相似度超过一定阈值的句子，再选取与 $\max W_j$ 相似度最小的句子作为第一句，这里两个句子之间的相似度计算表示为：

$$Sim(d_i,d_j)=sim(d_i,d_j)\times(1-|\frac{length(d_i)-length(d_j)}{length(d_i)+length(d_j)}|) \tag{9}$$

其中，$Sim(d_i,d_j)$的计算方法与上文 Cure 聚类过程描述相同，$length(d_i)$表示文档 d 中第 i 句的长度，$1-|\frac{length(d_i)-length(d_j)}{length(d_i)+length(d_j)}|$为惩罚因子。

4)迭代执行上述步骤，得到一定数量的 $\max W_j$ 对应句依次加入子主题摘要。

经过该方法筛选的句子组合，不仅将重要信息抽取到摘要中，而且减少了摘要的冗余性。

3.2.1 消除冗余生成的文摘示例

对 3.1.1 节获得的候选文摘进行冗余处理后，程序运行结果如图 4 所示。

主题　　美国 Costco 超市低毛利如何致胜

10%文摘

NO.1　Costco 的成功并非无法复制，以会员费形成资金流盈利的模式，诸如美容、餐饮或其他服务类。

20%文摘

NO.1　Costco 的成功并非无法复制，以会员费形成资金流盈利的模式，诸如美容、餐饮或其他服务类。

NO.2　在 Costco 内产品高质量低价格的驱动下，Costco 会员有着超高的忠诚度。

图 4　消除冗余生成的文摘示例图

4　实验结果及分析

微博摘要生成的实验主要分为两部分：1）获取有效的聚类结果确定子主题；2）根据子主题下的句子生成摘要集合。

4.1　聚类子主题

网络上微博信息量巨大，而各个离散用户之间的博文信息关联度低。为得到理想的子主题聚类结果，首先通过新浪微博接口爬取 NLP、云计算、大数据等领域关联互粉的 1000 个活跃用户最新的 200 条微博，结合 Cure 算法发现聚类子主题。本文使用 F 值作为检测子主题的指标，其用查准率和召回率的几何平均值来表示。在实验过程中，依照类内平方误差和准则，主题数分别取 1 到 10 进行实验，据图 5 所示，当主题数为 5 时，F 值最大。因此本文选择主题数为 5 进行实验。

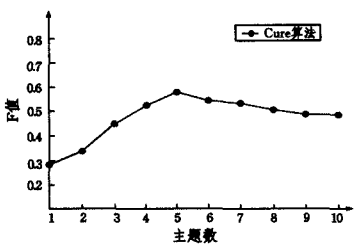

图 5　子主题数聚类结果

确定主题数后，分别取 6 个不同的收缩因子进行实验，得出的实验结果如图 6 所示，当收缩因子为 0.7 时取到的 F 值最佳，因此本文确定 Cure 聚类的收缩因子取 0.7。

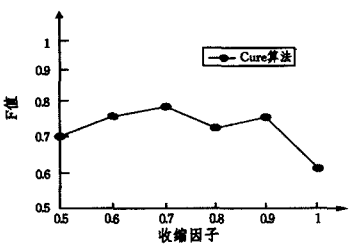

图 6　子主题聚类 F 值比较

4.2　子主题下摘要测评

该阶段使用上文得出的子主题作为语料并对其编号，包括“中国 hadoop 技术峰会(S-A)”10277 句、“智能手机＋移动互联网改变生活(S-B)”4402 句、“美国 Costco 超市低毛利如何致胜((S-C)”9323 句、“春节人口迁徙图(S-D)”6636 句、“学堂在线选课人次破百万(S-E)”5184 句。

下面给出部分“美国 Costco 超市低毛利如何致胜”主题下的摘要实例。

【原文】

全球管理咨询公司麦肯锡刚刚发布《2015 年中国数字消费者调查报告》，其中有一个重要结论就是实体店已死。对于那些会用互联网的人而言，实体店只是展厅，只有少数人会在线下买单，大多都会选择价格更优惠的线上去购物。不可否认互联网对于实体店造成了巨大的冲击，但实体店真的只有坐以待毙？未必，今天的微壳将分享美国 Costco 超市的低毛利的成功案例，以及 Costco 北美实体店无缝零售的新模式。只要懂得变通，实体店和电商共生繁荣的局面是完全能够实现的。

Costco 是美国一家著名商超，以贴近成本的低价格著称。在 Costco 内部，有两条硬性规定帮助了高质量的产品卖得便宜。一个是所有商品的毛利率不超过 14%，一旦高过这个数字，则需要汇报 CEO，再经董事会批准。第二，面对外部供应商，如果获知某家企业在别的地方定的价格比在 Costco 的还低，那么它的商品将永远不会再出现在 Costco 的货架上。

【10%摘要】

今天的微壳将分享美国 Costco 超市的低毛利的成功案例，以及 Costco 北美实体店无缝零售的新模式。

【20%摘要】

今天的微壳将分享美国 Costco 超市的低毛利的成功案例，以及 Costco 北美实体店无缝零售的新模式。

Costco 是美国一家著名商超，以贴近成本的低价格著称。

从上文实例可以观察到，自动生成的摘要与实际微博主题基本吻合。

依照 DUC 2004 的测评方法，由 5 个人分别对原文取压缩比为 10%，20%的句子作为人工摘要，对这些句子进行分词以后作为专家文摘，使用 ROUGE-1，ROUGE-2 对系统生成的文摘与专家文摘加以测评，表 1 列出了实验结果。

表 1　各主题语料测评结果

压缩率	10%的压缩率		20%的压缩率	
子主题	ROUGE-1	ROUGE-2	ROUGE-1	ROUGE-2
S-A	0.472	0.326	0.545	0.502
S-B	0.393	0.297	0.431	0.381
S-C	0.422	0.371	0.539	0.475
S-D	0.497	0.388	0.562	0.442
S-E	0.374	0.301	0.451	0.374

从数据上可以看出，利用本文方法对聚类后的子主题进行文摘抽取表现出较好的效果。

本文以准确率 P、召回率 R 和 F 值这 3 个标准评估自动摘要方法的性能。利用 LexRank 方法和本文提出的改进 Lex-PageRank 方法分别对上述子主题进行文摘处理，得出的对比结果如表 2 所列。

表 2　LexRank 和 Lex-PageRank 的对比结果(%)

方法	LexRank			Lex-PageRank		
子主题	P	R	F	P	R	F
S-A	42.6	40.8	41.7	48.2	41.5	44.6
S-B	49.1	48.7	48.9	52.3	49.8	51.0
S-C	41.1	39.2	40.1	45.7	43.6	44.7
S-D	53.8	49.7	51.6	57.8	52.7	55.1
S-E	40.1	38.3	39.1	42.7	40.9	41.8

通过实验可以观察到，本文方法在各主题下生成的文摘都优于 LexRank 方法的。实验结果表明，聚类算法划分微博文档子主题，并在子主题基础上利用改进的 Lex-PageRank 算法优化生成的摘要表现出较小的冗余性和较大信息覆盖率的特点，证明了本文提出的微博摘要生成算法切实可行。在聚类过程中，所提方法很大程度上过滤了噪声数据并提纯了文摘语料，提高了运算效率和准确性。在子主题句优化选择过

程中，对各子主题下的句子使用基于时间和权重特性的 Lex-PageRank 算法，抽取信息最全面、影响力最充分的句子加入摘要，而对冗余信息的处理又使主题内容表现得清晰明了，大大节省了用户阅读博文的时间，提高了阅读效率。综上所述，提出的基于 Lex-PageRank 算法的微博摘要优化方法可以有效挖掘微博中的热点，为用户提供较为准确和全面的信息。

结束语 由于目前微博上的在线文本日益增多，通过摘要抽取文本内容帮助用户提高阅读效率、缩短阅读时间已经变得很有必要。本文给出的主题模型下基于 Lex-PageRank 算法的微博摘要优化方法，其特点是在微博话题中引入摘要抽取技术，通过聚类在过滤孤立信息的同时将相同主题的博文组织在一起，然后优化选择主题下的句子生成摘要。实验结果表明，所提方法在生成摘要的过程中表现出较理想的覆盖率。但其不足之处在于，以“句子”为知识粒度来表示摘要并不十分贴合用户体验。在以句子为单元的摘要基础上，抽取主旨事件，并排序与润色，形成高可读性的内容摘要，将是下一步的工作。

参 考 文 献

[1] Kwak H, Lee C, Park H, et al. What is Twitter, a social network or a news media[C]// Proceedings of the 19th International Conference on World Wide Web. ACM, 2010: 591-600

[2] Brandow R, Mitze K, Rau L F. Automatic condensation of electronic publication by sentence selection [J]. Information Processing Manage, 1995, 31(5): 575-685

[3] Luhn H P. The Automatic Creation of Literature Abstracts [J]. IBM Journal of Research and Development, 1958, 2(2): 159-165

[4] Cao Yang, Cheng Ying, Pei Lei. A Review on Machine Learning Oriented Automatic Summarization [J]. Library and Information Service, 2014, 58(18): 122-130(in Chinese)
曹洋，成颖，裴雷. 基于机器学习的自动文摘研究综述[J]. 图书情报工作，2014，58(18)：122-130

[5] Han Yong-feng, Xu Xu-yang, Li Bi-cheng, et al. Web News Multi-document Summarization Based on Event Extraction [J]. Journal of Chinese Information Processing, 2012, 26(1): 58-66 (in Chinese)
韩永峰，许旭阳，李弼程，等. 基于事件抽取的网络新闻多文档自动摘要[J]. 中文信息学报，2012，26(1)：58-66

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

[7] Chen Yan-min, Wang Xiao-long, Liu Yuan-chao, et al. Automatic Text Summarization Based on Topic and Content [J]. Computer Engineering and Applications, 2004, 33(5): 11-14(in Chinese)
陈燕敏，王晓龙，刘远超，等. 一种基于文章主题和内容的自动摘要方法[J]. 计算机工程与应用，2004，33(5)：11-14

[8] Xie Hao, Sun Wei. Paragraph-Sentence Mutual Reinforcement Based Automatic Summarization Algorithm [J]. Computer Science, 2013, 40(11A): 246-250(in Chinese)
谢浩，孙伟. 基于段落-句子互增强的自动文摘算法[J]. 计算机科学，2013，40(11A)：246-250

[9] Yang Chang-chun, Zhou Meng. An Improved Hot Topic Detection Method for Microblog Based on CURE Algorithm[J]. Computer Simulcation, 2013, 30(11): 383-387(in Chinese)
杨长春，周猛. 基于改进 CURE 算法的微博热点话题发现[J]. 计算机仿真，2013，30(11)：383-387

[10] Ammar M B, Neji M, Alimi A M. The integration of an emotional system in the Intelligent Tutoring System[C]// The 3rd ACS/IEEE International Conference on Computer Systems andApplications. 2005: 145

[11] Langville A N, Meyer C D. Deeper inside pagerank[J]. Internet Mathematics, 2004, 1(3): 335-380

[12] Jin X, Deng Y F, Zhong Y X. Mixture feature selection strategy applied in cancer classificate from gene expression[D]. Shanghai: IEEE Press, 2005: 4807-4809

(上接第 260 页)

黄武汉，孟祥武，王立才. 移动通信网中基于用户社会化关系挖掘的协同过滤算法[J]. 电子与信息学报，2011，33(12)：3002-3007

[12] Wang Y X, Qiao X Q, Li X F, et al. Research on context-awareness mobile SNS service selection mechanism[J]. Chinese Journal of Computers, 2010, 33(11): 2126-2135(in Chinese)
王玉祥，乔秀全，李晓峰，等. 上下文感知的移动社交网络服务选择机制研究[J]. 计算机学报，2010，33(11)：2126-2135

[13] Groh G, Ehmig C. Recommendations in taste related domains: collaborative filtering vs. social filtering[C]// Proceedings of the 2007 International ACM Conference on Supporting Group Work. ACM, 2007: 127-136

[14] Shangguan Q, Hu L, Cao J, et al. Book Recommendation Based on Joint Multi-relational Model[C]// 2012 Second International Conference on Cloud and Green Computing (CGC). IEEE, 2012: 523-530

[15] Ma H, Yang H, Lyu M R, et al. Sorec: social recommendation using probabilistic matrix factorization[C]// Proceedings of the 17th ACM Conference on Information and Knowledge Management. ACM, 2008: 931-940

[16] Xu W, Cao J, Hu L, et al. A social-aware service recommendation approach for mashup creation[C]// 2013 IEEE 20th International Conference on Web Services (ICWS). IEEE, 2013: 107-114

[17] Golub G, Kahan W. Calculating the singular values and pseudo-inverse of a matrix[J]. Journal of the Society for Industrial & Applied Mathematics, Series B: Numerical Analysis, 1965, 2(2): 205-224

[18] Lee D D, Seung H S. Algorithms for non - negative matrix factorization[C]// Advances in Neural Information Processing Systems. 2001: 556-562

[19] Mnih A, Salakhutdinov R. Probabilistic matrix factorization [C]// Advances in Neural Information Processing Systems. 2007: 1257-1264

[20] Zhou D, Hofmann T, Schölkopf B. Semi-supervised learning on directed graphs[C]// Advances in Neural Information Processing Systems. 2004: 1633-1640

[21] Schafer J B, Frankowski D, Herlocker J, et al. Collaborative filtering recommender systems[M]// The adaptive Web. Springer Berlin Heidelberg, 2007: 291-324

[22] George T, Merugu S. A scalable collaborative filtering framework based on co-clustering[C]// Fifth IEEE International Conference on Data Mining. IEEE, 2005: 625-628