文章编号:1004-9037(2014)05-0720-10

基于拓扑结构的微博话题摘要生成算法

赵 斌 吉根林 徐 伟 顾彦慧

(南京师范大学计算机科学与技术学院,南京,210023)

摘要:话题摘要是自然语言处理中对文本进行内容归纳和概要生成的技术。传统的话题摘要研究主要针对新闻、Web 网页和博客这样的长文本,本文研究微博短文本的话题摘要问题。本文以微博转发消息为对象,提出具有拓扑结构的微博话题摘要生成算法(Microblog topic summarization,MTS)。首先通过微博转发上下文确定代表性词项;然后识别微博转发中的话题区域,从广度和深度两个方向对话题进行归并操作;最后,基于转发关系生成具有拓扑结构的微博话题摘要。本文实验采用真实的微博事件数据集验证 MTS 算法的有效性和可行性,并采用可视化方式展现微博话题摘要的结果。

关键词:微博;话题摘要;拓扑结构;转发;可视化

中图分类号:TP391 **文献标志码**:A

Microblog Topic Summarization Based on Topology Structures

Zhao Bin, *Ji Genlin*, *Xu Wei*, *Gu Yanhui*

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

Abstract: Topic summarization is a natural language processing for creating summaries of topic information. Previous work focused on summaries of news, web documents and blogs, while seldom on microblog topic summaries. A microblog topic summarization (MTS) method is proposed based on topology structures for microblog retweets. First, representative terms are selected according to structural relationships between retweeting tweets. Second, topic areas are identified after topic nodes are merged by using depth-first and breath-frist methods. Third, topic-oriented summaries with topology structure are generated through measuring adjacent topic nodes on the retweeting graph. Finally, experiments on the real-world event datasets show the effectiveness of the proposed methods. Visual topic summary trees are also produced for remarkably emphasizing the insight behind the evolving topics.

Key words: microblog; topic summarization; topology structure; retweeting; visualization

引 言

微博是当前流行的社交网络应用。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,个人的意见和评论逐渐汇聚融合形成群体观点,这是社会舆论的重要组成。所以,分析微博热点事件的群体观点,重现话题演化的完整过程,是一个具有理论意义和应用价值的研究课题。

自动摘要是自然语言处理中的经典课题。它通过分析文本,对内容进行归纳总结然后自动生成摘要。由于近年来互联网的蓬勃发展,针对用户生成内容(User-generated content, UGC)的自动摘要研究成为了学术界新的热点。文献[1]研究博客的话题摘要技术。从博客评论中获得代表性词项,选择包含代表性词项的代表性语句来表达话题信息。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,采用非监督的话题模型发现

基金项目:江苏省高校自然科学基金(13KJB520014)资助项目。
收稿日期:2014-06-07;**修订日期**:2014-08-07

兼具两种媒体特性的内容,以此构成事件摘要。文献[3]研究新闻评论的摘要,采用最大边缘相关(Maximal marginal relevance, MMR)和 Rating & Length (RL)排序算法发现代表性评论。文献[4]研究微博的上下文摘要,利用用户的交互信息构建用户的影响力模型,生成有效的上下文摘要。在可视化研究方面,相关工作较少。文献[5]设计并实现了一个交互式的文本分析系统(Text in sight via automated responsive analysis, TIARA),采用可视化技术分析大规模文本集中话题信息及其演化过程。

此外,在基于时序的话题摘要研究方面,文献[6]对摘要间的时间关系进行建模,提出了结合全局摘要和局部摘要的排序框架,用以生成新闻事件的时间线。文献[7]提出了一种新的统计模型(Text-based information diffusion and evolution, TIDE),用于在线社区话题传播和演化的联合推断,该模型综合考虑了文档内容、社交影响力和话题演化 3 方面因素。文献[8]提出了一种挖掘新闻演化过程的优化框架,目标函数的设计充分考虑了所生成时间线的相关性、全面性、一致性和多样性。文献[9]关注话题演化的拓扑结构,通过量化话题的演变过程识别话题。文献[10]提出基于主题的概率迁移矩阵,用于研究科技论文中的话题演化问题,并采用 GPU 技术加速计算过程。文献[11]提出了基于图模型的时间线摘要框架,生成同时含有图片和文字的可视化时间线。

本文主要研究微博短文本的话题摘要算法。传统的长文本摘要方法无法直接应用于微博话题摘要研究中。具体理由如下:

(1)长文本的话题摘要方法不适合短文本。传统的话题摘要研究对象主要为新闻、Web 文档和博客这样的长文本。这种类型的文本统计特征明显,通过分析话题内容的相关性可以有效归纳话题的摘要。但是,微博这样的短文本无法采用类似的方法。因为微博文本长度较短,统计特征不足,很难直接度量微博短文本之间的内容相关性。

(2)传统摘要的抽取式方法不适合微博的主题摘要。由于微博消息(尤其是转发消息)文本长度较短,主要依靠上下文表达语义。因而单一的语句无法传递完整的话题信息。所以,微博话题摘要不能采用抽取式的摘要方法。

(3)微博话题摘要中应该具有拓扑结构。传统的话题摘要中仅包含内容信息,但是微博不同于这样的 UGC 文本。微博通过"转发"功能实现信息的快速传播,"转发"将多条消息文本串联形成具有独立话题信息的文本序列。沿着转发序列,旧话题结束,新话题开始。这样,话题间的链接关系反映了热点事件中话题随时间演变的过程,也是话题迁移的方向。因而,为了理解微博话题,既需要消息文本,又需要文本间的拓扑关系。由此可见,拓扑结构不仅有助于微博话题摘要的提取,还是微博话题摘要的固有结构。

依据微博话题摘要的上述特点,本文基于微博转发关系重构话题区域,采用度量话题区域相关性的方法替代微博文本间的相关性度量,以克服短文本统计特征不足的缺点,最后生成具有拓扑结构的微博话题摘要。本文提出的话题摘要算法不仅可以反映话题的内容和强度,还能够通过话题间的拓扑结构表现出话题演化的过程。需要说明的是,本文研究的对象为微博转发消息,属于树型结构。

1　问题描述

话题摘要是自然语言处理研究中的经典问题,大多针对长文本集合,如文献[1～3]。近年来,学术界开始研究新兴社交媒体中的摘要问题,如文献[6～11]。现有的工作大多集中于基于时间线的话题摘要研究。但是,话题在社交媒体传播中具有持续性,单从时间维度上无法准确识别话题的边界。因而,本文在研究微博话题摘要时,充分利用话题间转发的拓扑结构发现话题区域,进而提取话题摘要。

在微博话题摘要问题研究中,设微博转发集合 $G=(V,E)$,其中 V 为微博文本集合,$V=\{v_1,v_2,\cdots,v_n\}$,v_i 经过分词后得到词项集合,记为 $T(v_i)$,所有词项的集合为 $T=\bigcup T(v_1)$;E 为文本间的转发关系集合,$E=\{\langle v_i,v_j\rangle \mid v_i\in V,v_j\in V\}$,$\langle v_i,v_j\rangle$ 为消息 v_i 到 v_j 的转发。不难发现,根据"转发"的基本特性,G 实际上为树型结构。因而,本文也将 G 称之为转发树。

微博的话题摘要问题描述为:给定微博转发集合 G 和最终话题数 s,依照话题在转发中演变的拓扑结构生成话题摘要 $G'=(V',E')$,G'和 G 同为树型结构。其中,V' 为话题集合,$V'=\{v'_1,v'_2,\cdots,v'_s\}$,话题 v'_i 由代表性词项构成,记为 $R(v'_i)$,所有话题的代表性词项集合 $R=\bigcup R(v'_i)$,E'为话题之间的链接关系,如图 1 所示。

在本文中,话题结点采用代表性词项来表示,记为 R,代表性词项 R 由词项集合 T 过滤所得,因而 $R\subseteq T$。

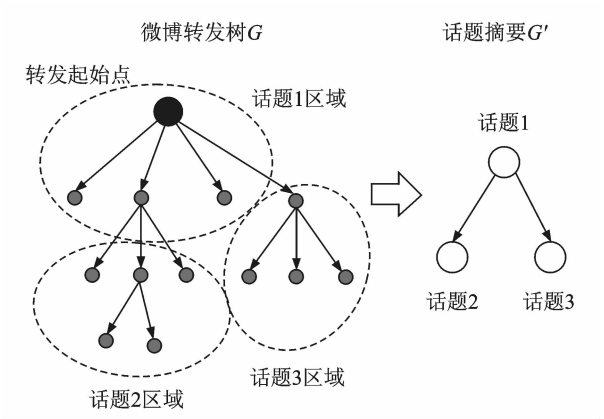

图 1 微博转发消息的话题摘要

Fig. 1 Topic summarization of microblog retweeting graph

2 基于转发的微博话题摘要算法

本文提出的微博话题摘要包含两个要素:话题结点和话题间的拓扑结构。话题结点反映了微博热点事件的内容信息,而话题间的拓扑结构反映了话题随转发扩散的演化过程。微博话题摘要中含有了这两个要素,不仅可以表示微博热点事件的内容概要信息,还可以反映事件演化的过程。

2.1 基于转发上下文的代表性词项选择方法

由于话题由代表性词项表示而成,因而从微博消息中识别出代表性词项对于生成话题摘要十分重要。在大量的词项集合中只有代表性词项才能反映热点事件的话题信息。因此,识别代表性词项不是一个简单的过程。例如,在"天安门广场乱扔垃圾"事件[12]中"个人""集体""道德"就属于代表性词项,它们来自于讨论"个人主义"和"集体主义"之间关系的微博消息。由于对微博消息进行分词后会产生大量词语,除了代表性词项以外,还包括停用词和非代表性词。常见研究表明,无法采用简单过滤的方式将代表性词项单独识别出来。

文献[13]发现代表性词项在时间分布上具有明显的"爆发性",这是它们与非代表性词和停用词的重要区别。但是经由真实数据集的实验发现,无法完全采用此特性准确识别出代表性词项。一些评论性的词语在微博讨论中也可能具有"爆发性",例如,"赞""强"和"给力"等情感词语,但是它们不具有任何话题信息。

鉴于微博话题在转发中具有持续性和更迭性的特点,本文结合此特征提出基于转发上下文的代表性词项识别方法。

2.1.1 微博转发中的代表性词项

代表性词项是话题信息的基本载体,它随着不同话题在转发链中交替出现。通常,表示话题的代表性词项在微博转发集合中具有两种基本特征:爆发性和持续性。

代表性词项的"爆发性"反映在转发树上,表现为词项在某个局部区域中的密集出现,而在时间分布上表现为明显的突起形状。不同于文献[13],本文的词项熵计算是基于"日"而不是"小时",并且选择出现频率高的词项。这样得到的词项才能具有话题的代表性。

而代表性词项的"持续性"依托于话题在转发拓扑结构上的演化特点。新话题在转发开启,然后持续被讨论转发。集中表现为代表性词项在转发的局部区域持续出现。因而,本文除了使用"熵"过滤代表性词项,还利用转发构成的链接关系来进一步识别微博话题中的代表性词项。

本文提出的基于链接关系的代表性词项选取方法描述为

$$\{T(v_i)\cap T(v_{j1})\cap\cdots\cap T(v_{jn})\mid\langle v_i,v_{jk}\rangle\in E,k\in[1,n]\}$$

式中:$T(v_i)$包含候选的代表性词项,$v_{j1},\cdots,v_{jn}$为v_i在转发树上的子结点。不难发现,如果要求出现在v_i中的词项t必须出现在所有的后续转发结点中,未免过于严苛。因此,设定阈值$\theta_t\in(0,1]$,如果v_i后续结点中有超过θ_t的结点都包含词项t,则t为代表性词项。

以"天安门广场乱扔垃圾"事件为例。图 2 中的话题区域 1 包含 5 个词项,"个体""公德""集体""国家"和"社会"。v_1,v_2,v_3,v_4为其中的 4 条微博消息,可以发现此区域主要讨论"个体"和"集体"之间关系。由于它们同属于一个话题区域,因而在v_1中讨论的话题信息,同样也会通过转发关系延续到后续的转发结点v_2,v_3,v_4中。其中"公德"出现在所有 4 个结点中,因而"公德"为代表性词项。而"个体"虽然没有出现在结点v_4中,但是如果阈值$\theta_t=60\%$,"个体"依然是代表性词项,最后可以肯定的是"权利"不属于此话题区域的代表性词项。

词项的"爆发性"和"持续性"源于话题在转发传播中的表现。本文基于这两种特性对词项进行选择。最终得到可以代表话题的代表性词项。

2.1.2 代表性词项选择算法

根据 2.1.1 节中代表性词项选择的基本思想,本文提出基于转发关系的代表性词项选择算法,描

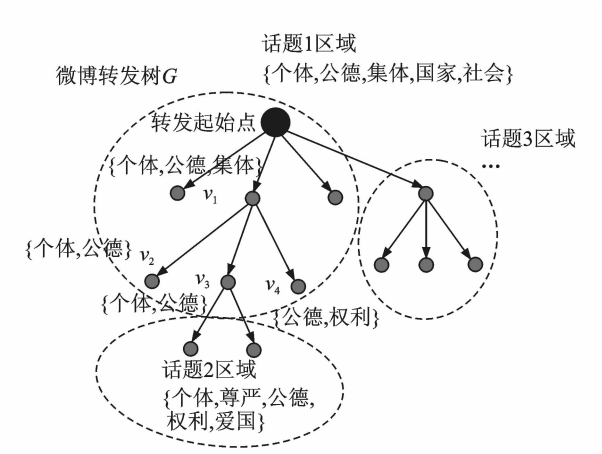

图 2　转发中的话题示例

Fig. 2　Topic example on a retweeting graph

述如下：

输入：微博转发树 $G(V,E)$，阈值 θ_t

输出：代表性词项 R

$R=\varnothing$

将根结点插入 Q；//Q 为待遍历结点的队列

do{

$v=\mathrm{pop}(Q)$；

$A=\{v_j|\langle v,v_j\rangle\in E\}$

for $t\in T(v)\wedge t\notin R$ do {

if $\dfrac{|\{v_j|t\in T(v_j)\wedge v_j\in A\}|}{|A|}>\theta_t$

$R=R\cup\{t\}$；

}

push(Q,A)；

}while ($Q!=\varnothing$)；

return R；

上述算法的时间复杂度和消息文本分词后的词项规模有关，每个结点分词后的最大词项数为常量。所以，代表性词项选择算法的时间复杂度为 $O(|V|)$。

2.2　基于树结构的话题摘要生成算法

话题摘要生成算法的目的是归纳出话题概要以及重构出话题之间的拓扑结构。随着用户在热点事件中关注焦点的改变，传播中话题信息自然地具有持续性和更迭性。具体表现为每个话题对应于转发树中的一个连通子图，用户的话题讨论会在该连通子图中逐渐展开。当话题更迭时，讨论焦点从一个话题连通子图转移到相邻的另一个话题连通子图中。本文利用微博话题在消息转发中的变化规律，探测出话题的区域，并且为每个区域生成对应的话题摘要。

为了探测出话题的区域，需要对转发树中的消息结点进行归并操作。将转发树中相同话题的结点进行合并生成新的结点。相应地，新结点的代表性词项来源于被合并结点的共同话题。在转发树上执行归并操作需要在“深度”和“广度”两个不同方向上进行。“深度”指微博消息间的转发关系，如图 2 中 v_1 和 v_2，v_3，v_4 的关系。“广度”指具有共同转发对象的结点间关系，如图 2 中 v_2，v_3 和 v_4 之间的关系。本文通过文本相似性度量将同属相同话题的结点进行合并，然后以共同的代表性词项作为它们的话题信息。

通过真实数据集的分析发现，转发树中的结点通常具有 3 种类型：话题型结点、评论型结点和转发型结点。话题型结点具有丰富的内容信息，包含代表性词项，较容易提炼出话题信息；评论型结点主要含有表达意见和情绪的情感词，相对于话题型结点能提炼的代表性词项较少；转发型结点的文本长度较短甚至为空，主要用于传递，目的在于帮助话题扩散。由此可见，评论型结点和转发型结点在微博转发中更依赖于所处的上下文环境，用户理解这样的消息，需要参考邻近发布的其他微博消息。也就是说，它们无法脱离上下文环境独立表达话题信息。所以，评论型结点和转发型结点在语义上往往被话题型结点所“覆盖”。

依据上述思想，本文提出具有拓扑结构的微博话题摘要生成算法（Microblog topic summarization，MTS）。该方法从深度和广度两个方向进行话题归并，以话题型结点为中心合并其余两种类型的结点，最终形成连通的话题区域。在合并的同时，邻接话题的边界也会逐渐显现，即话题之间的拓扑结构会自然形成。

2.2.1　深度方向的话题归并算法

对转发树的归并操作应该从根结点开始。本文采用层次遍历的策略依序访问每个结点，沿转发序列合并同属相同话题的子结点，提取共同的代表性词项作为新话题的代表。

深度话题归并算法描述如下：

输入：微博转发树 $G(V,E)$，阈值 θ_d

输出：话题树 $G'=(V',E')$

$R=\varnothing$；

将根结点插入 Q；//Q 为待遍历结点的队列

do{

$v=\mathrm{pop}(Q)$；

$A=\{v_j|\langle v,v_j\rangle\in E\}$

for $v_j\in A$ do {

if $\frac{|\{R(v)\cap R(v_j)|}{|R(v)\cup R(v_j)|}>\theta_d$ {

$R(v)=R(v)\cap R(v_j)$;

$A=A\cup\{v_k|\langle v_j,v_k\rangle\in E\}$;

$A=A-\{v_j\}$;

}

}

push(Q,A);

$V'=V'\cup\{v\}\cup A$;

$E'=E'\cup\{\langle v,v_j\rangle|v_j\in A\}$;

}while ($Q!=\varnothing$);

return G';

上述算法采用层次遍历策略访问结点,因而算法的时间复杂度为 $O(|V|)$。

2.2.2 *广度方向的话题归并算法*

广度方向上的话题归并主要针对具有共同转发对象的结点。本文在转发中采用代表性词项表示微博消息的话题信息。同一结点的转发中往往存在相似甚至相同的微博转发消息。由于它们具有接近的代表性词项,因而应该合并。以图 2 为例,v_2,v_3 和 v_4 具有共同的转发对象 v_1,其中 v_2 和 v_3 应该合并。按照上述思想,广度方向的话题归并算法基于广度优先策略依序遍历转发关系中的中间结点,采用聚类思想对其子结点进行归并操作。

广度话题归并算法描述如下:

输入:转发树中指定结点 v 及其子结点集合 $S=\{v_1,v_2,\cdots,v_m\}$,$v_i\subseteq R$,阈值 θ_b。

输出:聚类结果 $S'=\{v'_1,v'_2,\cdots,v'_{m'}\}$,$v'_i\subseteq R$

$S'=\varnothing$;

for $v_i\in S$ do {

$v'_i=v_i$; $S'=S'+v'_i$;

}

do {

计算 S'中元素的相似度矩阵$[Sim(v'_i,v'_j)]$

if $\mathrm{Max}(Sim(v'_i,v'_j))\geqslant\theta_b\wedge i\neq j$ {

$v'_i=v'_i-v'_j$;

$S'=S'-\{v'_j\}$;

} else break;

}while (true);

return S;

上述算法最主要的运算代价为相似度矩阵的计算,所以其时间复杂度为 $O(m^2)$,m 为待聚类的结点总数。

2.3 话题区域相关性度量及合并

当热点事件数据集规模较大时,MTS 算法生成的摘要规模可能会非常庞大,这样不利于热点事件摘要结果的展现。因而,合理的话题粒度对于结果表现非常重要。不难发现,话题粒度越小,话题信息就越琐碎,话题摘要的规模也就越大;相反,如果话题粒度越大,话题摘要的规模也就越小。当然,话题粒度太大,同样也不利于理解事件信息。

为了解决上述问题,本文通过度量话题相关性合并相似话题结点。具体操作为,首先对话题树中的所有邻接“话题对”按照相似度进行排序。然后根据用户设定的话题规模 s,从相似度最高的“话题对”开始合并,每合并一次,总话题数减少一个,如此重复合并,直到符合用户设定的话题规模 s。最后,选择话题中出现频率最高的词项集表示话题信息。整个算法的时间复杂度为 $O(|E|)$。

3 实验结果与分析

3.1 数据集介绍

为了研究微博热点事件的话题摘要问题,本文采用腾讯微博 API 收集了两个热点事件的微博消息,分别是“天安门广场乱扔垃圾”(TAM)和“2013 年苹果发布会”(ARC)。表 1 列出了两个热点事件微博数据集的基本情况,其中 MRTG 为 TAM 数据集中的最大转发消息集。本实验将采用 MRTG 进行算法性能比较,采用 ARC 数据集展现算法在大数据集上的可视化表现。需要说明的是,ARC 由在转发关系上不连通的转发集合构成。为了处理方便,本文设定了一个虚拟起点连接所有的转发集合。

表 1 微博热点事件数据集介绍

Table 1 Dataset description of microblog hot events

数据集	发布时间	用户数	微博数
TAM	2013 年 10 月	2 653	2 774
MRTG	2013 年 10 月	597	618
ARC	2013 年 9 月	7 934	10 179

为了深入了解微博热点事件转发消息的基本特征,本文对 MRTG 数据集进行了用户参与度的统计和转发树的可视化。如表 2 所示,大多数用户的参与度并不高,仅发布一条微博消息。图 3 也印

表 2　MRTG 数据集的用户参与情况

Table 2　Statistic data of MRTG postings

发布微博数	参与人数
3	1
2	19
1	577

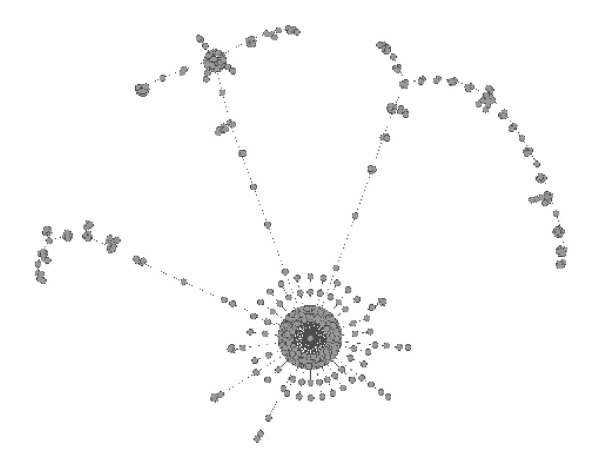

图 3　“天安门广场乱扔垃圾”事件的最大转发图

Fig. 3　Maximal retweeting graph of TAM event

证了此观点,在转发树中大量的微博消息集中在首次转发中,只有 3 条较长的转发路径。主要原因是相关议题引发了特定用户群的积极讨论,因而转发路径比较长。

不难发现,群体性的话题信息在转发树中呈现出“碎片化”,单一用户或者群体发表的内容无法具有广泛的代表性。因而,本文提出的 MTS 算法从转发关系入手跨越用户界限提炼话题信息,并且采用可视化的方式展现话题摘要的结果以及话题演化的拓扑关系。

3.2　基准测试算法

为了验证话题摘要算法 MTS 的有效性,本文设计了 3 个基准测试算法与之比较。它们分别是聚类算法(Text clustering for topic summarization, TC)[14]、结点出度排序法(Topic-node degree ranking for topic summarization, TDR)和 LDA(Latent Dirichlet allocation)算法[15]。TC 算法是采用基于密度的空间聚类算法(Density-based spatial clustering of applications with noise, DBSCAN)聚类实现的抽取式摘要算法,该算法在不考虑链接关系的情况下只关注消息文本的内容,按照文本相似度进行聚类,选取规模最大的 K 个聚类簇作为话题摘要。文本相似度计算采用基于词频-逆向文档频率(Term frequency-inverse document frequency, TF-IDF)向量和余弦相似性度量的方法来实现。而 TDR 算法按照结点出度对消息文本进行排序,选择出度高的前 K 个消息文本作为摘要。由于这些消息具有较高的关注度,因而具有代表性。LDA 算法是目前比较流行的基于概率的主题建模算法,该算法可以生成文本集合的主题信息,但是无法得到主题间的拓扑关系。所以,只有 TC 和 TDR 算法可以利用转发树中的连通关系表示话题间的拓扑关系。

需要说明的是,本文除 LDA 算法以外,其余算法均采用 3.1 节介绍的方法选择代表性词项。但是,由于不同算法生成话题的方法不同,各算法生成的话题所含代表性词项集也不同。所以,不同算法生成的话题质量是有差异的。

3.3　评估方法

话题摘要的评估一直是一个比较困难的问题。2005 年 NIST 组织的 DUC(Document understanding conferences)评测中确定了摘要人工评测的具体指标,但是本文研究的微博文本与新闻数据相差很大,如本文引言所述。因为没有基准测试的数据集可供测评使用,本文采用人工抽取的方式生成微博话题摘要的评测数据(即人工摘要),并且采用信息检索中的准确率、召回率和 F1-measure 三种指标[16]评价算法所生成摘要的全面程度和正确情况。此外,由于话题间的拓扑关系来源于微博“转发”,只要最终生成的话题是全面和正确的,本文算法的归并过程就不会造成微博话题在拓扑结构上出现错误的连接。

为了评估 MTS 算法的性能,本文采用两种不同规模的数据集进行测试,分别为 MRTG 和 ARC 数据集。虽然 MRTG 数据集规模较小,但是适合采用手工方式提取摘要。为了保证人工生成结果的客观性和有效性。本实验选择 5 位不同研究人员对 618 条微博进行逐一阅读(微博顺序随机生成),总结出共同的代表性话题及其词项,如表 3 所示。从此表中可以发现主要为话题摘要,仅含有少量评论摘要。这符合话题演化对于文本性质的基本要求。

除此以外,本文还采用 ARC 数据集展示 MTS 算法在大数据集上生成摘要的可视化效果。

本文实验将从代表性词项集、话题摘要结果和可视化展示这 3 方面,比较和分析上述 4 种算法的实验结果,以验证 MTS 算法的有效性和可行性。

表 3 对 MRTG 数据集经过人工处理生成的话题摘要

Table 3 Manual topic summarization on MRTG dataset

序号	话题摘要(代表性词项集)
0	权利 个体 公德 尊严 集体主义 自发 美国 爱国
1	冷漠 国家 社会 毛主席 提倡 拥护 官员 人民币 服务
2	打倒 万恶 美帝国主义 呲牙 美国 邪路
3	个体 压制 创造 思想 自由 谄媚 封建 专制 行政 服务 机构 衙府 卑劣
4	权利 地方 国家 监狱 演员
5	高喊 誓死 捍卫 打倒 文革 公安 攻击 事件 公民 社会
6	美国人 结婚 中国好女人 移民 投资 条件
7	美国 政府 民事 机构 安全 法律 军事 机构 国会 民主 选举 国家 总统 民政 反制裁
8	保障 民权 爱国主义 集体主义 建立 国家 现代 奴隶制
9	本性 家人 组织 登台 目的 违背 自然规律

表 4 在 MRTG 数据集上 4 种算法的代表性词项集比较

Table 4 Performances of four algorithms on MRTG dataset

算法	词项数	准确率/%	召回率/%	F1/%
TC	73	56.1	46.5	50.9
TDR	53	67.9	40.9	51.0
LDA	28	50.0	15.9	24.1
MTS	81	81.7	75.0	78.1

表 5 在 MRTG 数据集上 4 种算法的运行时间比较

Table 5 Runtimes of four algorithms on MRTG dataset

算法	运行时间/s
LDA	32
TC	900
TDR	5
MTS	28

在代表性词项集方面,采用信息检索中的准确率、召回率和 F1-measure 指标测试摘要结果中代表性词项的准确程度和全面性。在话题摘要结果方面,本文将检测算法生成的摘要结果和人工摘要的匹配程度,还有摘要结果中评论型摘要的比例,如果比例越低表明摘要结果的参考价值越高。在可视化方面,展示话题摘要的结果,包括话题间的拓扑结构,通过话题结点的尺寸表现话题的强度,以展示话题的重要性和差异性。

3.4 实验结果与分析

3.4.1 TAM 数据集实验结果分析

本次实验演化结果中的话题数量设定为 10(包含转发树的根结点)。TC 算法选择最大的前 10 个聚类作为话题摘要;TDR 算法选择出度最高的前 10 个微博消息结点作为话题摘要;LDA 算法直接生成 10 个主题;MTS 算法根据事先设定的话题数生成指定规模的话题摘要结果。本文实验设定 MTS 算法中 θ_t,θ_d 和 θ_b 均为 0.8。

首先比较各种算法生成摘要的代表性词项集。表 4 给出了 4 种算法在准确率、召回率和 F1-measure 指标方面的比较结果,表 5 则是算法执行时间的对比。其中,人工摘要中含有 88 个代表性词项。从实验结果可以得出以下 3 点:

(1)基于聚类的 TC 算法发现的代表性词项并不理想,而且算法执行时间过长。主要原因是,聚类的摘要结果来自于文本聚类的结果,从规模大的聚类簇中提取话题摘要。但是,这样的结点在内容上往往缺乏代表性词项,只是被简单地大量重复。例如,评论型结点"美国邪路"在本文的数据集中重复性很高,却不包含任何实质的话题信息。除此以外,由于聚类中需要进行相似性度量,而此操作的时间复杂度较高,所以耗费了算法大部分的执行时间。

(2)TDR 算法的效果同样也不理想。虽然它的执行时间最短,但是发现的代表性词项并不多。TDR 算法按照转发数挑选转发结点。影响力大的用户所发布的结点更容易被选中,它们的转发数比较高。但是这些结点主要的功能在于传播话题,本身反映的话题信息相对有限。

(3)LDA 算法发现的代表性词项最少,而且运算时间和 MTS 算法相当。由于 LDA 算法比较适合长文本的主题建模,然而微博转发消息大多不具有完整的文本语句,所以效果并不理想。

相比之下,MTS 算法兼顾摘要全面性和准确性,在代表性词项方面优势明显,更接近于人工摘要的结果,算法执行时间也在可接受的范围内。

在摘要结果及其可视化方面,由于 LDA 算法只能生成微博话题摘要的内容信息,而无法生成微博话题摘要的拓扑关系。因而,图 4 展示了除 LDA 算法以外的 3 种算法的话题摘要结果,图中白色结点代表评论型摘要(词项主要为情感词),而灰色结点代表话题型摘要(词项主要为名词)。每个结点的尺寸表示话题的强度,有向边表示话题迁移的方向。TC 算法的话题强度与聚类结果的规模成正比;TDR算法的话题强度与结点的转发数

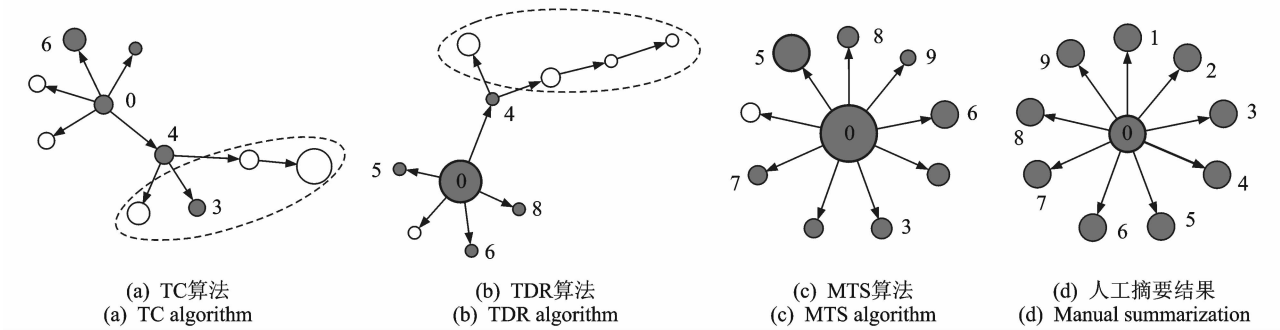

图 4　在 MRTG 数据集上 4 种方法话题摘要的拓扑结构

Fig. 4　Topic summarization drawings of four algorithms on MRTG dataset

成正比;而 MTS 算法的话题强度与话题区域“覆盖”的结点数量成正比。但是,人工摘要只可以提取出话题内容,却无法准确标注话题强度。因而图 4(d)中的结点尺寸相同。

为了展示 3 种算法挖掘出的具体话题内容,将可视化图中话题结点的词项集与表 3 人工选择的词项集进行相似度匹配,选择相似度超过阈值 0.8 且最匹配的摘要序号作为标注的记号。不难发现,TC 和 TDR 算法的可视化图非常相似,图的形态基本相同,并且包含部分相同的话题摘要。但是,两图中话题型摘要的数量较少。图 4(a,b)的被标记区域中的结点主要为评论型摘要,词项集以情感词为主,而缺少反映话题信息的名词。所以,这两种算法在话题演化图中表现出的实际话题数量为 5 个。

而 MTS 算法中评论型摘要仅 1 个,其余都是话题型摘要。从图中标注的摘要序号也可以发现,MTS 算法的话题摘要与人工标注的结果高度匹配,并且演化的拓扑结构也相同。所以,MTS 算法的话题演化结果优于其他算法。

总之,从上述 MRTG 数据集的各项实验结果来看,MTS 算法可以有效地展示微博热点事件的话题内容、话题强度和话题演化的拓扑结构。

3.4.2　ARC 数据集实验结果分析

除了在小规模的 MRTG 数据集上进行了测试,本文还以“2013 年苹果发布会”的 ARC 数据集作为实验对象,展示 MTS 算法在大数据集上的可视化表现。图 5 为该数据集的转发图(需要说明的是,为了方便展示,图中的起始结点为虚拟结点,它为所有转发的共同起点)。

本次实验在 ARC 数据集上采用 MTS 算法生成了完整的话题演化结果,如图 6 所示。由于话题数量较多,表 6 中仅列出了强度最大的 5 个话题摘要内容。从中可以看出在“2013 年苹果发布会”中大众关注的焦点,如“5 s 的配置”“iWork 免费”等。图 6 中还包含了一条话题演化路径。该路径揭示了用户从讨论 iWork 免费问题转而讨论 iMovie 和

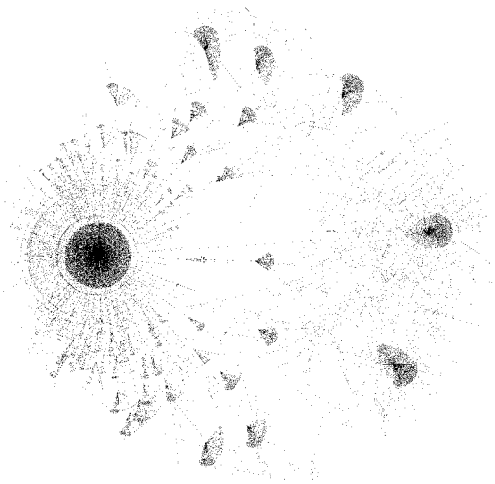

图 5　带虚拟结点的“苹果发布会”转发图

Fig. 5　Retweeting graph with a virtual node of ARC event

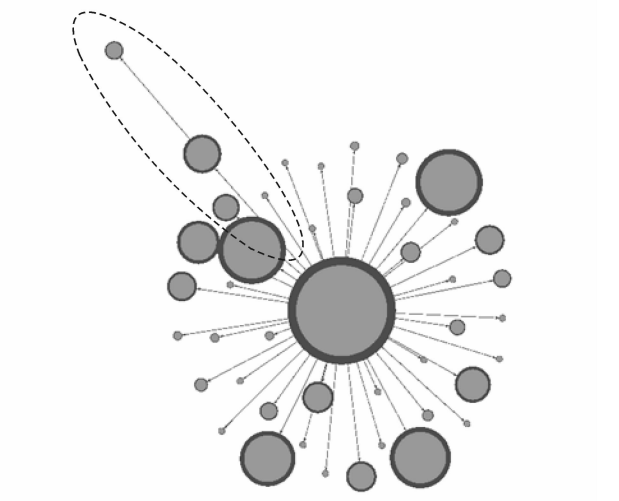

图 6　ARC 数据集话题摘要的拓扑结构

Fig. 6　Topic summarization drawing of ARC dataset

表 6 MTS 算法生成的 ARC 数据集话题摘要

Table 6 Topic summarization from MTS on ARC dataset

序号	话题摘要(代表性词项集)
0	虚拟结点
1	5s 64 位 摄像头 处理器 售价 发布 屏幕
2	Home 键 处理器 摄像头 发布 屏幕
3	收入 免费 系统 Mac 版本 操作系统
4	网络 产品 摄像头 处理器 时间 售价 LTE
5	OS 免费 应用 iWork 升级

iPhoto 编辑软件的免费问题。

需要说明的是,在表 6 中的 0 结点为虚拟结点。由于在 ARC 数据集中存在大量的转发起始结点,它们长度较短且没有明显话题,因而被虚拟结点所“覆盖”。最终导致图 6 中的虚拟结点尺寸较大。

4 结束语

本文以微博转发消息为研究对象,研究微博热点事件的话题摘要问题。本文将微博的话题摘要问题分为两个阶段。首先,利用消息间的链接关系识别代表话题信息的词项集合;然后,提出基于树结构的话题摘要生成算法 MTS。该方法从深度和广度两个方向对转发结点进行归并,汇聚生成话题摘要及其演化的拓扑关系。本文提出的 MTS 算法不仅可以反映话题的内容和强度,还能够表现出话题演化的过程。实验结果表明基于转发树的话题摘要算法 MTS 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 P. Comments-oriented blog summarization by sentence extraction [C]//16th ACM Conference on Information and Knowledge Management (CIKM'07). Lisbon, Portugal: ACM, 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, HI, USA: ACM, 2012:1173-1182.

[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, HI, USA: ACM, 2012:265-274.

[4] 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, Italy: ACM, 2013:527-536.

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

[6] Yan R, Kong L, Huang C, et al. Timeline generation through evolutionary trans-temporal summarization[C]//Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP' 11). Edinburgh, UK: ACL, 2011:433-443.

[7] Lin C X, Mei Q, Han J, et al. The joint inference of topic diffusion and evolution in social communities [C]//11th IEEE International Conference on Data Mining (ICDM' 11). Vancouver, BC, Canada: IEEE, 2011:378-387.

[8] Yan R, Wan X, Otterbacher J, et al. Evolutionary timeline summarization: A balanced optimization framework via iterative substitution[C]//Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'11). Beijing, China: ACM, 2011:745-754.

[9] Jo Y, Hopcroft J E, Lagoze C. The web of topics: Discovering the topology of topic evolution in a corpus[C]//Proceedings of the 20th International Conference on World Wide Web (WWW'11). Hyderabad, India: ACM, 2011:257-266.

[10] Masada T, Takasu A. Extraction of topic evolutions from references in scientific articles and its GPU acceleration[C]//21st ACM International Conference on Information and Knowledge Management (CIKM' 12). Maui, HI, USA: ACM, 2012:1522-1526.

[11] Yan R, Wan X, Lapata M, et al. Visualizing timelines: Evolutionary summarization via iterative reinforcement between text and image streams[C]//21st ACM International Conference on Information and Knowledge Management (CIKM' 12). Maui, HI, USA: ACM, 2012:275-284.

[12] 网易新闻. 历年天安门垃圾吨位[EB/OL]. http://travel. 163. com/13/1008/08/9ALC3DI600063KE8. html,2013-10-08/2014-6-7.

[13] Long R, Wang H, Chen Y, et al. Towards effective event detection, tracking and summarization on mi-

croblog data[C]//Web-Age Information Management 12th International Conference (WAIM'11). Wuhan, China: Springer, 2011:652-663.

[14] Nenkova B A, McKeown K. Automatic summarization[J]. Foundations and Trends in Information Retrieval: Now Publishers, 2011,5(2/3):103-233.

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

[16] Manning C D, Raghavan P, Schütze H. Introduction to information retrieval[M]. New York, USA: Cambridge University Press, 2008:142-143.

作者简介:赵斌(1978-),男,讲师,博士,研究方向:Web 数据挖掘,E-mail:zhaobin@njnu.edu.cn;吉根林(1964-),男,教授,博士生导师,研究方向:数据挖掘技术及应用;徐伟(1990-),男,硕士研究生,研究方向:数据挖掘技术与应用;顾彦慧(1978-),男,博士,研究方向:自然语言处理。