软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2012,23(10):2586−2599 [doi: 10.3724/SP.J.1001.2012.04191]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

短信息的会话检测及组织*

田 野[1], 王文东[1+], 饶京海[2], 王 冠[1], 郭 亮[1], 陈灿峰[2], 马 建[1,3]

[1](网络与交换国家重点实验室(北京邮电大学),北京 100876)

[2](诺基亚研究院,北京 100176)

[3](无锡物联网产业研究院,江苏 无锡 214135)

Conversation Detection and Organization of Mobile Text Messages

TIAN Ye[1], WANG Wen-Dong[1+], RAO Jing-Hai[2], WANG Guan[1], GUO Liang[1], CHEN Can-Feng[2], MA Jian[1,3]

[1](State Key Laboratory of Networking and Switching, Beijing University of Posts and Telecommunications, Beijing 100876, China)

[2](Nokia Research Center, Beijing 100176, China)

[3](Wuxi SensingNet Industrialization Research Institute, Wuxi 214135, China)

+ Corresponding author: E-mail: wdwang@bupt.edu.cn, http://www.bupt.edu.cn

Tian Y, Wang WD, Rao JH, Wang G, Guo L, Chen CF, Ma J. Conversation detection and organization of mobile text messages. *Journal of Software,* 2012,23(10):2586−2599 (in Chinese). http://www.jos.org.cn/1000-9825/4191.htm

Abstract: Mining the latent conversations which are implied in the big amount of text messages stored on one's mobile phone, is a challenging problem. They can hardly be organized by threads, due to lack of necessary metadata such as "subject" and "reply-to". This paper proposes an innovative conversation recognition model based on temporal clustering algorithms and topic detection methods. The study first clusters the text messages into candidate conversations based on their temporal attributes, and then does further analysis using a semantic model based on latent Dirichlet allocation (LDA). In the end, the text messages are organized as conversations based on their integrated correlation of temporal relevancy and topic relevancy. This approach is evaluated with a real dataset, which contain 122 359 text messages collected from 50 university students during 6 months.

Key words: text message; temporal clustering; topic; latent Dirichlet allocation

摘 要: 如何挖掘存储在手机上的大量短信息背后所隐含的会话信息,是一个非常具有挑战性的问题,因为它们并不具备"主题"、"回复"等经常被用于邮件线索分析的元数据.基于此,提出了一种基于时间聚类算法和话题检测的短信息会话识别模型.首先,根据短信息流的时间分布特性,将会话双方的所有短信息划分到一个一个的候选会话中,进而运用基于 latent Dirichlet allocation(LDA)训练出来的语义话题模型,对候选会话进行更深层次的分析;利用该话题模型度量了各个候选会话在话题上的相关度.最后,在综合时间和话题相关度的基础上,通过对候选会话的合并识别出隐含的会话信息.通过对包含了 50 名大学生在 6 个月中产生的 122 359 条短信进行实验验证,证明了该算法的有效性.

关键词: 短信息;时间聚类;话题;latent Dirichlet allocation

中图法分类号: TP391 **文献标识码**: A

* 基金项目: 国家重点基础研究发展计划(973)(2009CB320504); 国家高技术研究发展计划(863)(2011AA01A101)
收稿时间: 2011-05-17; 修改时间: 2011-07-21, 2011-09-22; 定稿时间: 2012-01-16

由于其方便、快捷,短信息成为了手机终端上最受欢迎的应用之一,特别是在年轻人中短信息的使用率非常高.根据 Nielsen(http://blog.nielsen.com/nielsenwire/online_mobile/under-aged-texting-usage-and-actual-cost/)的统计显示,美国的年轻人一个月平均发送 3 146 条短信息.也就是说,不包括睡觉的时间在内,一个人一个小时平均发送超过 10 条短信息.考虑到手机处理能力和存储空间的飞速增长,这些数量庞大的短信息有可能在手机上存储很长一段时间,这就给如何在手机上高效、合理地组织这些短信息带来了新的挑战.

挖掘隐含在海量交互式短文本对象(如短信息、微博、网络聊天室会话信息)背后的知识不仅具有较高的理论研究价值,同时也蕴含了巨大的社会价值与经济价值.政府部门能够借助相关研究成果及早发现社会热点话题,对社会聚焦的热点事件与突发事件进行合理预防和疏导;网络运营商和服务提供商通过挖掘会话信息能够发现用户的潜在兴趣点,定向提供增值业务.本文从用户个人短信息数据的会话检测角度出发展开研究.普通的短信息管理工具提供了基本的排序功能,允许按照收发人、收发时间或者内容来组织.一些功能稍复杂的软件可以将短信息按照会话视图来组织,这样,由同一个联系人发送或接收的短信息被显示到一起.然而即使这样,用户在查看一条短信息时仍然很难获得关于它的上下文信息(如会话主题、前因后果等).对于这样的情况,如果仅仅根据联系人来组织短信息就不能很好地找到相应的会话线索.在本文中,我们尝试根据时间上下文及话题相关性将用户的短信息按照会话主题的形式进行组织.这里,我们对短信息中的会话作了如下定义:

会话. 在特定的参与者之间连续产生的一系列短信息流.它们往往具有一定的时间突发性,并且在某个时间段内集中于某一个确定的主题.

短信息与会话是紧密相关的.首先,大多数短信息都不是孤立的,它们往往是和属于同一次会话中的其他短信息相互关联,隐含因果关系的,属于同一次会话中的所有短信息共同表达某一个主题;其次,每一条短信息都是和某一个联系人相关联的,该联系人或者是短信的发送方或者是接收方,这是短信息的一个重要上下文信息;第三,在一个会话中,大多数短信都是以问答形式存在的,除了首条短信息外,其他短信息大都是对前一条短信息的回复,各条短信息在时间分布上表现出明显的时间序列特性;第四,大多数短信息都是只包含了较少词汇的短文本,同时,表达缺乏规范性,仅仅从单独一条短信息中很难判断其关注的主题[1].由于以上提到的因素,要从组织无序的短信息中发现隐含的会话并提取出相应的主题,是一项非常困难的任务.在自然语言处理领域,话题检测与跟踪技术(topic detection and tracking,简称 TDT)要解决的问题与我们的目标比较类似,它旨在从象 Web 页面以及新闻报道这样的文本对象中发现并跟踪某个特定的主题.然而,这项较为成熟的自然语言处理技术却并不能用于短信息的处理.这是因为,TDT 处理的对象多是结构清晰、主题集中、表达规范、词汇较多的长文本,而短信息并不具备这样的特点.也有相关的研究关注于检测邮件的线索,并将它们以会话的形式组织起来.但是,这种方法同样不适用于短信息,因为在邮件线索检测中常常会用到的“主题”和“回复”等元数据并不存在于短信息中.

针对短信息的基本特征,本文提出了一种基于时间序列分析和文本内容特征分析的方法.仅仅考虑时间上下文的因素,将发生在特定收发双方之间的所有短信息的时间戳看作是一个离散时间序列,根据同一会话中相邻短信间时间间隔小于不同会话间的时间间隔的假设,我们首先将发生在特定收发双方之间的所有短信按照时间分布的疏密程度进行聚类,将每一个聚得的类看作是一个候选会话.然而在某些情况下,一次会话有可能在时间分布上被分割为不连续的部分.因此,仅仅根据时间分布的疏密特性来确定会话并不完全合理.鉴于此,本文在进行候选会话检测的时候尽可能地使得得到的每一个候选会话的时间窗口稍小于或接近真实会话的持续时间,然后,结合候选会话之间在话题内容以及时间分布上的综合相关度对各个候选会话进行平滑处理,使得话题相关并且时间分布集中的候选会话能够被聚合到一起.

本文第 1 节简要介绍相关研究成果.第 2 节详述候选会话检测中用到的自适应聚类算法.第 3 节提出一种基于 LDA(latent Dirichlet allocation)的候选会话平滑算法.第 4 节通过采集到的真实短信数据对本文提出的算法进行验证.第 5 节总结全文并对下一步工作进行展望.

1 相关工作

我们从以下 3 个领域来论述与本文相关的工作:新闻流的话题检测和跟踪、个人及社会化流媒体对象的事件检测以及短文本话题相关度比较.

话题检测与跟踪(TDT)作为自然语言处理领域的一项技术,旨在发现并线索化话题相关的报道.作为一种基于事件的信息组织和检索方式,它包含 5 个基本的部分:报道分段、首篇报道检测、话题跟踪、话题聚类以及链接检测.它所处理的对象基本上都是新闻报道类的文本流集合,集合内的文本通常都是关于同一类新闻报导的.因此,可以采用自然语言处理的方法和技巧从这些文本对象集中提炼出共同的话题.与短信息不同,TDT 处理的对象多是较长和完整的文本,因此,仅仅利用自然语言处理的方法就能很好地达到目的.

有一些研究组织也关注于从个人照片集或者类似 Facebook,YouTube,Flickr 这样的社会化媒体中挖掘出隐含的事件信息.Cooper[2]提出了一套基于时间及图像内容相似度聚类的算法来组织个人图片,其目的是使得在同一次事件中产生的图片被组织到一起.其方法是:首先,根据照片之间时间距离的高斯变换构建一个时间相似度矩阵;然后,利用 Learning Vector Quantization(LVQ)度量时间相关度矩阵的新颖指数.他指出,可以利用时间相似度矩阵的新颖指数来区分两次独立事件的边界.实验证明,该方法能够较好地识别出照片集中隐含的独立事件,然而结果受高斯变换参数的影响较大.

Zhao[3]同样也提出了一套从社会化媒体对象中识别和检测隐含事件的方法.在他的方法中,综合考虑了隐含在邮件、博客等媒体对象中的时间属性和社会关系属性等.然而在他们的方法中用到的某些属性,如社交网络结构,在短信息中是无法获得的.此外,在分析文本对象的内容特征时,他们采用 TF-IDF 来度量不同对象之间的内容相似度,然而实践证明,该方法仅仅对于长文本有效,对包含词汇较少的短信息却不太适合.

在过去十几年中,有不少的学者提出了度量短文本语义相似度的各种方法.Bollegala[4]利用搜索引擎来度量词汇之间的相似度;而 Metzeler[5]也采用了相似的方法,他根据 Web 搜索引擎得到的查询结果来间接度量短文本之间的语义相关性.上述两种方案都依赖于外部搜索引擎的查询结果.Quan[6]提出了一种基于话题模型的短文本相似度计算方法,这种方法与本文采用的方法比较类似,不同点在于,该方法是利用话题模型来修正短文本的特征向量,最终仍然通过计算向量之间的夹角余弦来表示文本之间的语义相关性.而本文则采用了一种求最小最大值的方法,相比较而言,本文采用的方法在不降低结果准确性的同时,能够相对减少计算开销.

2 候选会话识别模型

针对流数据的分析已经成为研究者关注的热点话题,其研究成果被广泛运用于电信、证券、生物等多个领域[7,8].特定两个会话者之间的短信息集合可以看作是一个短文本数据流,其在时间维度上的分布是非均匀的,具有一定的突发性[9].针对这个特点,本文提出了一种候选会话识别模型,以短信息流的时间分布密度为依据进行自适应的层次型聚类,然后,通过比较每个聚类组合的综合聚类质量选择出最优者作为候选会话组合.

假定短信息流可以表示为数据集 $Sms=\{s_1,\ldots,s_i,\ldots\}$,其中,$s_i=(person_i,time_i,content_i)$为顺序到达的第 i 条短信息,$person_i$ 表示第 i 条短信息所对应的会话者(不区分发送方和接收方),$time_i$ 表示第 i 条短信息所对应的时间戳,$content_i$ 表示第 i 条短信息所包含的文字内容.本节所述的候选会话识别模型旨在从顺序到达的短信息流中识别出特定会话者之间存在的潜在会话组合,算法基于以下假设:

在确定的两个会话者之间所产生的短信息流中,由于会话的突发性和连续性,同一会话内的相邻短信息之间的时间间隔分布较为均匀,并且远小于属于不同会话的相邻短信息之间的时间间隔.根据该假设,候选会话识别的关键在于分析短信流的时间分布特性,找出合理的参考时间间隔.若相邻两条短信息之间的时间间隔小于该参考时间间隔,则被视为属于同一会话;反之,则视作不同会话间的分界点.然而,由于用户在交流习惯及会话内容上的不同,很难确定一个普遍适用的参考时间间隔值.

针对上述约束,本文将候选会话识别转化为针对时间序列的层次聚类问题.算法的基本思想是:首先,将所有短信息按照不同会话对象划分到不同的会话空间 $\{\mathbb{C}_1,\ldots,\mathbb{C}_i,\ldots\},\left(\bigcup_i \mathbb{C}_i = Sms,\bigcap_i \mathbb{C}_i = \varnothing\right)$,其中,$\mathbb{C}$表示第 i 个会

话空间,它包含了与第 i 个会话者之间产生的全部短信息.进行划分的目的是因为本文认为在短信息这种通信方式下,一个会话通常只会发生在两个特定的会话者之间,对于多人共同参与某一个会话的情况本文暂不考虑.为了避免不同会话者之间产生的短信息在时间分布上相互重叠影响候选会话识别,需要针对不同的会话空间分别进行处理;第 2 步,针对某一个特定的会话空间$\mathcal{C}$,提取出每条短信息对应的时间戳,并按照产生的时间先后进行排序得到时间序列 $Time=\{time_1,\ldots,time_i,\ldots\}$.对该时间序列进行层次化聚类,最终得到不同的聚类组合$\{Round_1,\ldots,Round_i,\ldots\}$,其中,$Round_i=\{Z_i^1,\ldots,Z_i^j,\ldots,Z_i^{last}\}$ 表示经过第 i 轮聚类后得到的聚类组合,Z_i^j 表示在本轮聚类后得到的第 j 个类,也就是第 j 个候选会话,$Round_i$ 中一共包含了$|last|$个这样的候选会话;第 3 步,计算聚类组合中各个 $Round$ 所对应的聚类质量,最优者即为本节所期望的候选会话组合.本算法包含了两个核心步骤:时间维度上的层次化聚类以及聚类质量评估.

2.1 层次化聚类

层次化聚类算法又称为树聚类算法,它旨在依据数据的关联规则,以分层聚合的方式反复对样本进行合并,以形成一个层次数据序列.就本文所需要解决的问题而言,对不同的会话空间处理方式是一致的,因此,在下文中,我们仅仅以某一个特定的会话空间为例加以阐述.

假定某个会话空间$\mathcal{C}$包含的各条短信息所对应的时间序列为 $Time=\{time_1,\ldots,time_i,\ldots,time_n\}$.首先,对该序列进行初始化处理:计算时间序列 $Time$ 中任意相邻样本之间的时间间隔,并对所得的全部时间间隔进行排序和去重处理,得到一个按升序排列的时间间隔序列 $Gap=\{gap_1,\ldots,gap_i,\ldots,gap_m\}$,$1\leqslant i\leqslant m$(假设经过去重后,序列包含了 m 个互不重复的时间间隔值).算法由树状结构的底部开始逐层向上进行聚合,算法描述如图 1 所示.

Algorithm. 层次聚类算法.

输入:某个会话空间$\mathcal{C}$包含的各条短信息所对应的时间序列:$Time=\{time_1,\ldots,time_i,\ldots,time_n\}$;

输出:针对时间序列 $Time$ 的 m+1 个不同聚类组合:$\{Round_0,\ldots,Round_i,\ldots,\ldots,Round_m\}(1\leqslant i\leqslant m)$.

符号说明:

m:不重复的时间间隔个数;

$reference$:参考时间间隔值;

Z_i^j:第j轮聚类后生成的第i个类;

$|Round_j|$:第j轮聚类后生成的聚类组合中包含的聚类个数;

$d(Z_k^{j-1},Z_i^j)$:类 Z_k^{j-1} 与类 Z_i^j 之间的类间距;

1. 初始化:
   (1) 计算时间序列$Time$中任意相邻样本之间的时间间隔,并对所得时间间隔值进行排序和去重,得到一个按升序排列的时间间隔序列:$Gap=\{gap_1,\ldots,gap_i,\ldots,gap_m\}$
   (2) 构造初始聚类组合:$Round_0=\{Z_1^0,\ldots,Z_n^0\}$,其中,$Z_i^0=\{time_i\}, i=1,2,3,\ldots,n$,
   $Round_1,\ldots,Round_m=\varnothing$
2. 聚类:

```
(1)  for (i=1,j=1;1≤j≤m;j++)
(2)    reference→Gap_j                    //从时间间隔序列中取出第j个值作为参考时间间隔
(3)      Z_1^j ← Z_1^{j-1}                //将第j-1轮聚类结束后得到的第1个类作为第j轮聚类过程的初始值
(4)      for (k=1;1≤k≤|Round_{j-1}|;k++)
(5)        calculate d(Z_k^{j-1},Z_i^j)   //计算相邻两个类之间的类间距
(6)          if ( d(Z_k^{j-1},Z_i^j) ≤reference)  //若类间距小于或等于参考时间距离,则合并生成新类
(7)            Z_i^j = Z_i^j ∪ Z_k^{j-1}
(8)            Round_j = Round_j ∪ {Z_i^j}
(9)          else
(10)           Z_{i++}^j ← Z_k^{j-1}      //若类间距大于参考时间距离,则作为新类的开始
(11)     end for
(12) end for
```

Fig.1 Algorithm of hierarchical clustering

图 1 层次化聚类算法

其中,类间距 $d(Z_k^j, Z_i^{j-1}) = \min_{time_m \in Z_k^j, time_n \in Z_i^{j-1}} \| time_m - time_n \|$.

2.2 候选会话质量评估

经过层次化聚类后,得到 m+1 个不同的聚类组合,其中,每一个聚类组合都对应于针对会话空间$\mathcal{C}$所包含的时间序列 *Time* 的一个划分.要发现最接近用户真实会话模式所对应的时间序列划分,可以等价于从聚类过程完成后得到的不同的聚类组合中发现聚类质量最优者.本文首先构造出通过层次聚类得到的所有划分所对应的聚类指标函数曲线,文献[10,11]证明了聚类质量最佳的组合出现在聚类指标函数的极值点处,因此,本文通过探测曲线极值点来探测最佳聚类组合.

聚类指标由类内紧凑度和类间离散度两个分量决定.给定某一个聚类组合 $Round_j$,其类内紧凑度 $Compactness(Round_j)$和类间离散度 $Separation(Round_j)$分别表示为

$$\begin{cases} Compactness(Round_j) = \sum_{i=1}^{|Round_j|} \sum_{time \in Z_i^j} |time - \overline{Z_i^j}| \\ Separation(Round_j) = \sum_{i=1}^{|Round_j|} |\overline{Z_i^j} - \overline{Time}| \end{cases} \tag{1}$$

$\overline{Z_i^j}$ 和 $\overline{Time}$ 分别表示第 j 轮聚类组合中的第 i 个类以及整个时间序列 *Time* 中所包含的所有数据点的算术平均值.第 j 轮聚类组合的类内紧凑度 $Compactness(Round_j)$的物理含义是每个类的所有数据点与各个类的算术平均点的距离之和,而类间离散度 $Separation(Round_j)$的物理含义是各个类的算术平均点与整个时间序列的算术平均点的距离之和.考虑最极端的情况:$Round_0$ 和 $Round_m$,即当每个数据点单独为一个类和所有数据点被划分到一个类的情况下,很明显有公式(2)成立:

$$\begin{cases} Compactness(Round_0) = Separation(Round_m) = 0 \\ Compactness(Round_m) = Separation(Round_0) = \text{Max} \end{cases} \tag{2}$$

其中,Max 为一待确定的常数值.因此,类内紧凑度和类间离散度具有相同的值域区间.设在第 j 轮聚类中 Z_{i-1}^{j-1} 与 Z_i^{j-1} 合并:

$$\left.\begin{aligned} &Compactness(Round_j) - Compactness(Round_{j-1}) \\ &= \sum_{i=1}^{|Round_j|} \sum_{time \in Z_i^j} |time - \overline{Z_i^j}| - \sum_{i=1}^{|Round_{j-1}|} \sum_{time \in Z_i^{j-1}} |time - \overline{Z_i^{j-1}}| \\ &= \sum_{time \in Z_{i-1}^{j-1} \cup Z_i^{j-1}} |time - \overline{Z_{i-1}^{j-1} \cup Z_i^{j-1}}| - \left(\sum_{time \in Z_{i-1}^{j-1}} |time - \overline{Z_{i-1}^{j-1}}| + \sum_{time \in Z_i^{j-1}} |time - \overline{Z_i^{j-1}}| \right) \\ &\geqslant 0 \end{aligned}\right\} \tag{3}$$

$$\left.\begin{aligned} &Separation(Round_k) - Separation(Round_{j-1}) \\ &= \sum_{i=1}^{|Round_j|} |\overline{Z_i^j} - \overline{Time}| - \sum_{i=1}^{|Round_{j-1}|} |\overline{Z_i^{j-1}} - \overline{Time}| \\ &= |\overline{Z_{i-1}^{j-1} \cup Z_i^{j-1}} - \overline{Time}| - (|\overline{Z_{i-1}^{j-1}} - \overline{Time}| + |\overline{Z_i^{j-1}} - \overline{Time}|) \\ &\leqslant 0 \end{aligned}\right\} \tag{4}$$

因此,$Compactness(Round_j)$是关于自变量 j 的增函数,而 $Separation(Round_j)$是关于自变量 j 的减函数,并且二者的值域区间均为[0,Max].分别对 $Compactness(Round_j)$以及 $Separation(Round_j)$作归一化处理,同时,为了平衡 $Compactness(Round_j)$与 $Separation(Round_j)$在一阶导数上的差异,对类内紧凑度 $Compactness(Round_j)$作参数为α(小于 1)的指数变换,并且进行归一化处理,得到聚类指标函数表达式:

$$Quality(Round_j) = \left(\frac{Compactness(Round_j)}{\text{Max}} \right)^{\alpha} + \frac{Separation(Round_j)}{\text{Max}} \tag{5}$$

由文献[10]可知,最佳聚类对应于类内紧凑度和类间离散度的最佳平衡点,这个点所对应的聚类组合聚类质量达到最优,在数值上反映为聚类指标函数 $Quality(Round_j)$取得极小值.即,最佳聚类:

$$J^*=\text{argmin}_{j\in\{0,1,2,\ldots,m\}}Quality(Round_j).$$

由于对每一轮聚类的结果进行聚类质量评估是一笔较为耗时的开销,可以考虑一种逼近算法,以便能够只通过较少轮聚合就找出近似最优的候选会话组合.由于聚类指标函数在达到最小值之前单调递减,而在过了最小值之后单调递增,因此,可以把评估的范围限定在极值点附近的一个较小的区间内.

3 候选会话平滑模型

经过时间维度的聚类处理之后,会话空间所包含的短信息被划分为多个候选会话.然而,仅仅依据时间维度的分布特性,很难足够准确地将短信息划分到相应的会话中去.这种不精准性体现在两个方面:时间跨度长、相邻短信息间时间间隔大的会话有可能会在候选会话组合检测中被错误地划分为多个较小的会话单元;会话主题无关,但由于相邻短信息之间时间间隔小,本应属于不同会话的短信息有可能会被错误地划分到同一个会话中去.为了降低仅仅依靠时间聚类挖掘短信息流中所隐含的会话模式的不足,需要对候选会话组合进行平滑处理.

本文所提出的候选会话平滑模型同时考虑了候选会话之间的话题关联度和时间关联度,该模型基于这样的考虑:属于同一会话的短信息必然聚焦于相同的主题,相邻候选会话之间如果话题关联较高,而且时间分布较为连续,则有可能是属于同一次会话中产生的短信息;对于非相邻的候选会话,即使话题关联度较高,也被认为属于不同的会话.鉴于此,候选会话平滑模型将各个候选会话看作是包含了数条短信息的短文本.通过计算各个短文本之间的话题关联度以及各个候选会话之间的时间关联度得到综合的关联系数,依据该系数进行相邻候选会话之间的聚合.

3.1 潜在狄利克雷分布(latent Dirichlet allocation,简称LDA)

相邻候选会话之间的综合关联系数由话题关联度和时间关联度所决定,本文采用潜在狄利克雷分布(LDA)训练了一个话题模型,基于该模型,可以比较候选会话在可能的会话主题上的关联度.LDA 是一种生成概率模型,它由 Griffiths 和 Steyvers[12]于 2004 年提出.LDA 基于这样的假设,即一篇文档由不同的话题以一定的概率组合而成,而每个话题本身也是关于一系列词汇的概率分布.设语料集 D 包含 N 个唯一的单词:$W=\{w_1,w_2,\ldots,w_N\}$ 以及 K 个潜在的话题,d_i 代表单词 w_i 所出现的文档,z_i 代表单词 w_i 所要表达的潜在话题.同时,$\theta_j^{(d)}=p(z=j\mid d)$ 代表在指定文档为 d 的前提下,其主题为 j 的后验概率;而 $\varphi_w^{(z_j)}=p(w\mid z=j)$ 代表在指定话题为 j 的前提下,单词 w 出现的后验概率.生成模型 LDA 的形式化描述可以由公式(6)给出:

$$\begin{cases}\theta^{(d)}\sim Dirichlet(\alpha)\\ z_i\mid\theta^{(d_i)}\sim Multinomial(\theta^{(d_i)})\\ \varphi^{(z)}\sim Dirichlet(\beta)\\ w_i\mid z_i,\varphi^{(z_i)}\sim Multinomial(\varphi^{(z_i)})\end{cases}\tag{6}$$

α和β作为 Dirichlet 分布的超参数,主要用于控制分布的稀疏性.根据这个模型,每个单词 $w_i\in W$ 都会被分配一个潜在的话题变量 z_i.

给定某一个观察语料集,LDA 的任务即是通过计算后验概率 $p(z\mid w)=p(w,z)\Big/\sum_z p(w,z)$ 提取出隐含在语料集中的话题 z.然而,因为分母上的和式涉及到数量庞大的项并且很难因式分解,因此直接通过该式计算后验概率非常困难.Griffiths 和 Steyvers 在文献[5]中提出了采用吉布斯抽样(Gibbs sampling)来提取隐含话题的方法,本文也采用了该方法.

Gibbs 抽样根据文档中其他单词的话题分布情况来估值某一个单词属于某个话题的概率.该条件概率分布

可以表示为 $p(z_i=v\mid z_{-i},w,\alpha,\beta)\propto\hat{\varphi}_{w_i}^{(z_i)}\cdot\hat{\theta}^{(d_i)}$, 其中, $\hat{\varphi}_{w_i}^{(z_i)}=\frac{n_{z_i}^{w_i}+\beta}{\sum_{w_i}n_{z_i}^{w_i}+N\times\beta}$, $\hat{\theta}^{(d_i)}=\frac{n_{d_i}^{z_i}+\alpha}{\sum_{z_i}n_{d_i}^{z_i}+K\times\alpha}$. $\hat{\varphi}_{w_i}^{(z_i)}$ 是在给定话题 z_i 的前提下,单词 w_i 的后验概率;而 $\hat{\theta}^{(d_i)}$ 是在给定文档为 d_i 的前提下,话题组合的后验概率.这里,z_{-I} 是指除了单词 w_i 之外,其他单词的话题分配情况;而 $n_{z_i}^{w_i}$ 代表除了当前实例之外,单词 w_i 由话题 z_i 所生成的次数; $n_{d_i}^{z_i}$ 代表除了当前实例之外,话题 z_i 出现在文档 d_i 中的次数.在进行 Gibbs 抽样之初,每个单词都被随机地分配一个话题,作为 Markov 链的初始状态.此后,Markov 链不断迭代,而在每一轮的迭代中,新的状态值由根据上式抽样而得的 z_i 值来确定.经过数轮迭代之后,Markov 链收敛于一个稳定状态,其最终的状态值即是期望的概率分布.

3.2 基于LDA的话题模型

为了训练基于 LDA 的话题模型,需要采用真实的短信息数据.为此,我们收集到了 50 名大学生志愿者近 6 个月中产生的 122 359 条真实短信息.将属于其中 40 名大学生的 92 872 条短信息作为训练数据集,而将属于另外 10 名大学生的 29 487 条短信息作为测试数据集.采用开源的 GibbsLDA++(http://gibbslda.sourceforge.net)来训练话题模型,其中,狄利克雷分布中的超参数α和β分别设为 0.5 和 0.1,话题数 N 设为 50,Gibbs 抽样迭代次数设为 1 000.最终,话题模型中的每个话题被表示为一个向量,向量中的每个元素为一个 Key-Value 对,其中包含了话题中可能出现的词汇以及词汇出现在该话题中的概率.该话题模型的其中一部分如图 2 所示.

Topic 1: (工作,0.06485),(实验,0.05425),(实验室,0.05321),(导师,0.03257),(程序,0.02645),(考勤,0.00266),(项目,0.00134),(讨论,0.00107),(例会,0.00142),...

Topic 3: (上课,0.07631),(学校,0.01261),(同学,0.00396),(老师,0.00316),(考试,0.00308),(教室,0.00297),(作业,0.00184),(宿舍,0.00145),(食堂,0.00124),...

Topic 8: (看电影,0.07963),(电影院,0.04513),(爆米花,0.03154),(阿凡达,0.03207),(朋友,0.01393),(晚上,0.01173),(约会,0.00501),(推荐,0.0023),...

Topic 9: (春节,0.07153),(寒假,0.04122),(新年,0.03081),(火车票,0.02967),(回家,0.02419),(机票,0.0142),(下雪,0.0117),(快乐,0.00912),...

Topic 13: (看病,0.05224),(医院,0.04007),(医生,0.01326),(请假,0.02286),(健康,0.01472),(打针,0.01081),(中医,0.01034),(输液,0.00736),...

Topic 17: (上网,0.084341),(无聊,0.00499),(电脑,0.003004),(朋友,0.002846),(睡觉,0.002688),(开心,0.002214),(游戏,0.002056),(手机,0.001804),...

Topic 23: (打球,0.06413),(比赛,0.03105),(球迷,0.03020),(篮球,0.02669),(决赛,0.02219),(球场,0.01843),(朋友,0.01731),(球队,0.01468),(裁判,0.01468),...

Fig.2 LDA based topic model

图 2 LDA 话题模型

3.3 候选会话话题关联度计算

每个候选会话是一个包含了若干条短信息的短文本,经过分词和去停用词处理后,对应一个含有若干单词的词汇集,其话题特征值由一个 50 维向量 $d_i=\{r_i^1,r_i^2,...,r_i^{50}\}$ 来表征.我们把每个短文本看作是一个 50 维话题空间中的一个点,该点在某个维度上的投影 $r_i^j(1\leqslant j\leqslant 50)$ 即是它关于该维度所对应话题的相关程度,由公式(7)计算得到:

$$r_i^j=\sum_{word\in d_i\cap topic_j}p(word) \tag{7}$$

公式(7)表示短文本 i 关于第 j 个话题的相关度等于话题向量 $topic_j$ 中所包含的在短文本 d_i 中出现过的词汇所对应的概率之和.而两个短文本 d_i 和 d_j 之间的话题相关度 $relevancy_{i,j}$ 则由它们在各个话题维度上的相关度分量决定:

$$relevancy_{i,j}=\max(\min(r_i^1,r_j^1),\min(r_i^2,r_j^2),...,\min(r_i^{50},r_j^{50})) \tag{8}$$

公式(8)隐含地表明,两个文本之间的话题关联度由二者在各个话题维度上的关联度的最大值来决定,而它

们在各个话题维度上的相关度则取决于它们各自与该话题的相关度的较小值.

3.4 候选会话合并

判断相邻候选会话是否应该合并的依据包含两个方面:一是话题关联度,二是时间关联度.只有在话题空间以及时间分布上均具有连续分布特性的候选会话,才具有聚合的必要.因此,本节将综合话题空间以及时间两个方面的因素给出相邻候选会话之间的相关度衡量方法,进而依据该综合相关度进行候选会话的聚合,实现短信息按照真实会话的有效划分.

相邻候选会话之间的综合关联度不仅取决于二者之间的话题相关度,同时还受时间分布特性的影响.相邻候选会话之间即使话题相关度很高,但是如果时间间隔太大也不能视为同一个会话.因此,在决定是否合并相邻两个候选会话时,还需要考虑二者的时间关联度.如果将每个候选会话看作是一个簇,则相邻两个簇 Z_i^* 和 Z_{i+1}^* 之间的时间关联度可由公式(9)计算得到:

$$temporal_{i,i+1}=\exp\left(-\frac{d(Z_i^*,Z_{i+1}^*)}{L}\right) \tag{9}$$

公式(9)中,$d(Z_i^*,Z_{i+1}^*)=\min\limits_{time_i\in Z_i^*,time_{i+1}\in Z_{i+1}^*}\|time_i-time_{i+1}\|$ 表示 Z_i^* 和 Z_{i+1}^* 之间的最短距离,其值等于 Z_i^* 中最后一个样本与 Z_{i+1}^* 中第 1 个样本之间差值的绝对值.参数 L 用于调节相关度系数,在本文中设置为 10 000.

综合话题相关度和时间相关度,我们给出相邻候选会话之间的综合相关度表达式,如公式(10)所示:

$$correlation_{i,i+1}=temporal_{i,i+1}\times relevancy_{i,i+1} \tag{10}$$

在得到相邻会话之间的综合相关度之后,可以进行候选会话之间的合并.将各个候选会话看作是离散分布在一维坐标轴上的一组数据点,利用第 2.1 节中所使用的层次聚类算法对该数列进行聚类处理,选取出聚类质量最优者即为最终的会话组合.需要考虑的一点是,相邻候选会话之间的综合相关度越高,说明其被合并的可能性越高,故二者之间的距离应该越大.因此,本文定义相邻候选会话之间的距离为综合相关度的函数,其表达式如公式(11)所示:

$$dis_{i,i+1}=\exp(-Correlation_{i,i+1}) \tag{11}$$

4 实验验证

为了验证算法的有效性,测试数据集中所包含的 29 487 条短信息的拥有者已经按照真实的会话场景对各自的短信息进行了手工划分.实验从以下几个方面展开.

4.1 参数调节

实验 1 共分为 3 个步骤:第 1 步是参数调节.为了选择出层次聚类过程中公式(5)中参数α的最佳取值,我们随机挑选了 5 名志愿者,用他们之间产生的短信息构造了包含 5 个短信息会话空间的测试子集.通过对比使用不同α值时所获得的性能指标,选择出最佳的α值.表 1 给出了 5 个测试子集的统计特征.实验中,我们采用了 Precision,Recall 以及 F-Score 这 3 项指标来衡量算法的有效性,它们的定义分别为

$$Precision=\frac{\text{正确检测到的会话数}}{\text{检测到的会话总数}},$$

$$Recall=\frac{\text{正确检测到的会话数}}{\text{真实的会话总数}},$$

$$F\text{-}Score=\frac{2\times Precision\times Recall}{Precision+Recall}.$$

Table 1 Testing data for parameter learning of α

表 1 参数α学习所使用的测试数据集

志愿者/联系人	短信息条数
A/*A*1	523
B/*B*3	576
C/*C*6	475
D/*D*4	492
E/*E*8	506

分别为α赋值 0.2,0.4,0.5,0.8 以及 1.0 后发现,当α取值为 0.4 时,F-Score 达到最大值,如图 3 所示.因此在后续实验中,我们对α赋值为 0.4.

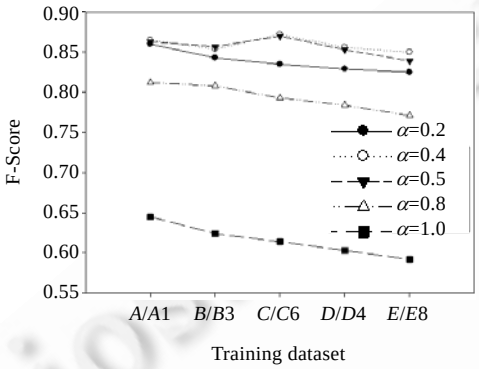
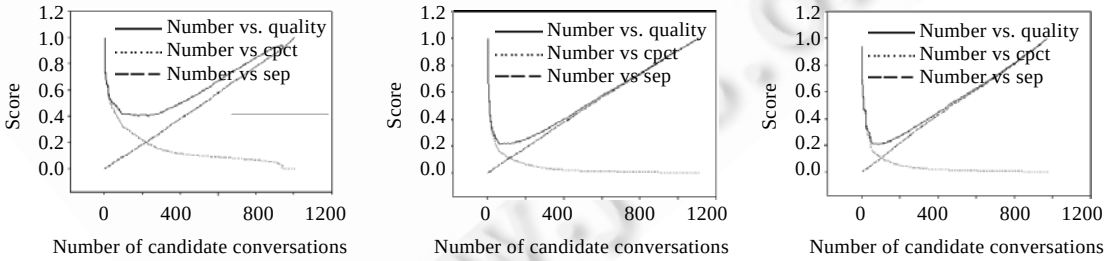


Fig.3 F-Score vs. α

图 3 F-Score vs. α

4.2 候选会话识别

实验的第 2 步,我们随机选取了志愿者 *A*,*B*,*E* 与其联系人 *A*2,*B*4,*E*3 之间的短信息作为测试子集,以验证用时间聚类算法检测候选会话的有效性.图 4 分别给出了这 3 个测试子集在层次聚类过程中,类内凝聚度、类间离散度以及聚类指标函数随着聚类个数增长的变化关系.可以看出,类内凝聚度随着聚类个数的增长呈现逐渐减小的趋势,而类间离散度则呈现出相反的变化趋势,这两个指标的值域区间均在[0,1]之间.作为反映聚类综合效果的指标,“聚类指标函数”呈“V”字型曲线分布.由第 2.2 节的分析可知,聚类指标函数在最小值处所对应的聚类结果即为最佳候选会话组合.



(a) *A*/*A*2 聚类指标 vs.聚类总数 (b) *B*/*B*4 聚类指标 vs.聚类总数 (c) *E*/*E*3 聚类指标 vs.聚类总数

Fig.4 Changing relationship between compactness, separation, quality with cluster number

图 4 类内凝聚度、类间离散度以及聚类指标函数随着聚类个数增长的变化关系

表 2 给出了在 3 个测试子集上进行候选会话识别的统计信息.

Table 2 Statistical results of candidate conversation detection

表 2 候选会话检测统计结果

测试子集	短信息总数	手工标注的会话总数	算法检测到的候选会话总数	相邻候选会话间的最佳时间距离(小时)
A/*A*2	1 001	202	230	0.903 4
B/*B*4	1 136	60	65	23.542 1
E/*E*3	974	46	53	31.641 5

可以看出,采用层次型时间聚类算法检测到的最佳候选会话个数与志愿者手工标注出的会话总数基本吻合.从统计结果来看,由于会话者之间交流习惯方面的差异,会造成检测到的属于不同会话者的最佳候选会话组合中包含的各个相邻候选会话之间的时间距离值偏差较大.*A*/*A*2 之间交流较为频繁,但平均每次会话仅仅包含 4.95 条短信息;*B*/*B*4 以及 *E*/*E*3 之间的交流不如 *A*/*A*2 频繁,但每次会话持续时间较长,且包含的短信息数较多,其中,*B*/*B*4 之间平均每次会话包含 18.93 条短信息,而 *E*/*E*3 之间平均每次会话包含了 27.17 条短信息.这一结果恰恰从另一个侧面反映了人们在日常生活中使用短信息进行交流和沟通的目的以及方式上存在着较大的差异性,不能以相同的阈值来对时间序列进行分段.

4.3 候选会话合并

实验的第 3 步,本文验证了候选会话平滑模型的有效性.该环节包含 3 个测试项:话题模型正交性测试、短文本相似度算法性能测试以及基于话题关联度和时间关联度的候选会话合并有效性测试.

话题模型的正交性对候选会话间话题相关度的计算有着直接的影响.所谓话题模型的正交性,指的是构成话题模型的各个话题维度之间应该满足语义的正交关系,也即是话题与话题之间不存在语义上的重叠.话题模型的正交性越强,候选话题的话题特征表示越合理,越能有效地区分各个候选会话.本文所使用的话题模型是基于 LDA 算法由 Gibbs 抽样而得,每个话题维度是一个包含了多个键值对的向量,其中,每个键值对包含了在该话题中出现的词汇以及词汇所对应的权重.

为了验证该模型的正交性,本文与采用传统的 TF-IDF(term frequency-inverse document frequency)模型构建的话题模型进行了比较.所谓的 TF-IDF 是一种统计方法,用以评估词汇对于一个文件集或一个语料库中的某一个文档的重要程度.词汇的重要性随着它在文档中出现的次数成正比增长,同时也会随着语料库中包含它的文档个数的增多而成反比下降.各个话题维度之间的正交性由夹角余弦表示,两个话题维度正交性越强,则相似度越低,其夹角余弦值也越小.

图 5 和图 6 分别表示由 LDA 和 TF-IDF 生成的两个话题模型的话题向量之间的正交性色图矩阵.其中,位于第 *i* 行第 *j* 列的色块表示第 *i* 个与第 *j* 个话题向量之间的正交度,颜色越深,正交性越明显;反之,正交性越弱.

显然,从整体视觉效果上比较,图 5 比图 6 颜色更深(由于各维度与其自身的夹角余弦为 1,该值与不同维度之间的夹角余弦取值差异太大.反映在图示上,会降低不同维度之间夹角余弦的视觉反差.因此,为了更容易观察,我们将各维度与其自身的夹角余弦值设为 0,在对比两个模型正交性时可以忽略对角线上的取值).因此可以判定,由 LDA 所生成的话题模型较 TF-IDF 所生成的话题模型而言,其各个维度之间的语义差异更为显著.

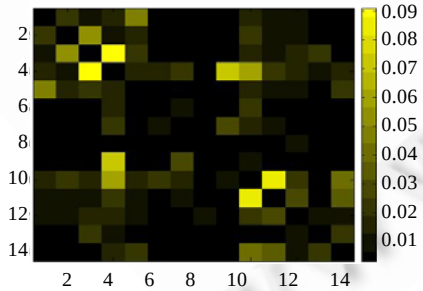


Fig.5 Color map of orthogonality of LDA based topic model

图 5 采用 LDA 所训练话题模型的正交性色图

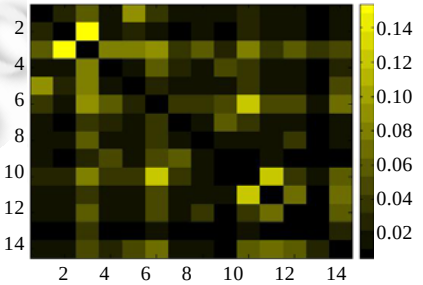


Fig.6 Color map of orthogonality of TF-IDF based topic model

图 6 采用 TF-IDF 所训练话题模型的正交性色图

在第 2 个测试项中,我们对 Quan 等人[6]提出的短文本相似度算法与本文在第 3.3 节所提出的短文本相似度算法进行了比较.与本文所提算法类似,Quan 等人所提出的短文本相似度算法也基于 LDA 话题模型,其具体过程可以描述如下:

- 第 1 步,利用相关语料集训练得到 LDA 话题模型,该话题模型包含了若干个话题向量,每个话题向量与本文在第 3.2 节中所述的话题向量结构相同,向量中的每个元素为一个 Key-Value 对,用以指示每个单词出现在该话题中的概率;
- 第 2 步,用向量表示待比较相似度的两条短文本,其中,向量中的每个元素是短文本中每个词汇的 TF*ID 权值;
- 第 3 步,分别构建两个待比较短文本 d_1 和 d_2 所特有的词汇集 $Dist(d_1)$和 $Dist(d_2)$,$Dist(d_1)$和 $Dist(d_2)$分别包含了只出现在 d_1 中而没有出现在 d_2 中以及仅出现在 d_2 中而没有出现在 d_1 中的所有词汇.算法认为,如果这些来自不同短文本的词汇出现在同一个话题的概率均高于某一个阈值,则可以推断出 d_1 与 d_2 在该话题上有一定的相关性,因此可以用 $Dist(d_1)$与 $Dist(d_2)$中词汇出现在各个话题中的概率值来修正其 tf*idf 权值;
- 第 4 步,针对每个话题向量,分别挑选出 $Dist(d_1)$与 $Dist(d_2)$中在该话题下出现概率最大的词汇,假定这两个词汇在文本 d_1 和 d_2 特征向量中的下标值分别为 m 和 n,如果两个词汇的概率值均大于某个阈值,则用 d_2 中第 m 个词汇的 TF*IDF 权值乘以它在该话题下出现的概率值加上这个词在 d_1 中的 TF*IDF 权值来修正 d_1 所对应的特征向量.类似地,对 d_2 所对应的特征向量也进行相同的处理;
- 第 5 步,计算修正后得到的两个短文本特征向量之间的余弦夹角值,该值即为两个短文本之间的相似度.

可以把上述算法看作是 TF*IDF 模型的一个变种,其相对于传统 TF*IDF 模型的改进之处在于,通过引入话题模型修正了非共现词在特征向量中的 TF*IDF 权重值,这样,一方面可以在一定程度上解决 TF*IDF 模型在计算短文本相似度时面临的特征值稀疏问题;另一方面,使得文本相似度体现了一定的语义相关性.由于引入了“话题”这一概念,话题与词汇之间的相关度会在较大程度上影响最终的文本相似度的大小.与该算法不同,本文在第 3.3 节所提出的短文本话题相关度算法并没有采用 TF*IDF 模型,而是通过比较各个短文本在由话题模型中若干话题向量所张成的话题空间中所投射的坐标值大小来计算其相似度.

从计算复杂度上考虑,文献[6]所提出的算法的时间开销主要在于构建短文本所对应的特征向量,以及利用话题模型对特征向量进行修正.在一个短文本个数为 K 的语料集中,假定包含了 N 个词汇以及 T 个话题,则计算一个短文本所对应的特征向量的时间复杂度是 $O(NK)$,构造所有短文本特征向量的时间复杂度是 $O(NKK)$,生成 $Dist(d_1)$与 $Dist(d_2)$的时间复杂度为 $O(N)$,根据话题模型对待比较的两个短文本所对应的特征向量进行修正的时间复杂度是 $O(2NT)$.因此,该算法总的时间复杂度为 $O(NKK)+O(N)+O(2NT)$.而本文所提算法的时间开销主要在于计算待比较的两个短文本中所有词汇在每个话题分量中出现的概率总和,其时间复杂度是 $O(N_1T)+O(N_2T)$,其中,N_1 和 N_2 分别指的是待比较相似度的两个短文本中所包含的词汇数.通常情况下,N_1 和 N_2 都远小于整个语料集中的词汇数 N.由此可见,本文所采用的短文本相似度算法在时间复杂度上要小于文献[5]所提出的算法.

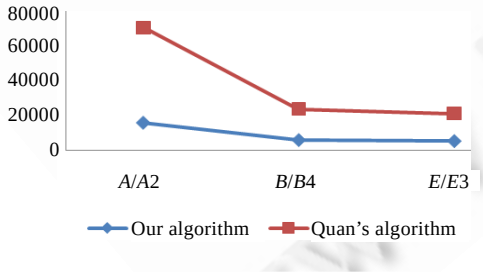


Fig.7 Time performance comparison between our algorithm and Quan's algorithm

图 7 本文所提算法与 Quan 所提算法时间性能对比

为了对比两种算法的时间性能,我们分别采用两种算法计算了在第 1 步实验中得到的 $A/A2$,$B/B4$,$E/E3$ 之间产生的各个相邻候选会话之间的相似度,时间开销如图 7 所示(纵轴以 ms 为单位).

在第 3 个测试项中,我们采用本文提出的候选会话平滑模型对在第 2 步中得到的候选会话组合进

行了合并.为了验证该模型的有效性,本文与其他几种算法进行了比较.

(1) Tao-Hsing Chang[13]提出了一种针对短文本流的分段算法,该算法旨在找出隐含在文本对象中的话题边界.其方法可以分为 4 个步骤:

- 首先,抽取每一条短文本中的核心词汇并对其加以扩展,以避免短文本中关键词较少带来的矩阵稀疏性问题;
- 其次,根据词汇之间在句子和段落中的距离构造词汇距离矩阵;
- 第三,根据词汇距离矩阵计算句子之间的相似度,并依据相似度大小对短文本流进行话题块(topic block)的划分,使得相似度较高的短文本被划分到同一个话题块中;
- 第四,对第 3 步中产生的多个候选划分进行排序,以全局最优者作为最佳划分.

(2) Wang[14]提出了一种针对即时通信短文本的聚类算法 WR-Kmeans,该算法首先依据相邻短信息间时间戳的间隔大小对短信息流进行分段,间隔值小于预先设定阈值的相邻短信息被划分到同一个候选会话中去,进而使用 HowNet 对各个候选会话进行关键词扩展,并利用改进的 VSM 算法计算候选会话之间的相似度.最后,使用聚类算法对相似度较高的候选会话加以合并.

(3) Cooper[2]针对个人照片集提出了一种基于 LVQ 的事件检测方法,该方法在进行事件检测的过程中用到了照片中所包含的时间戳信息,是一种针对时间流的聚类方法.因此,在不考虑短信息文本内容特征的情况下,该方法同样可以用于对短信息的会话检测.Cooper 首先把照片集中的照片按照时间先后排序,然后计算任意两张照片的时间相似度,基于该相似度构造了一个相似度矩阵,然后通过该相似度矩阵计算得到该矩阵的新颖指数,并利用 LVQ 和 Nearest-neighbor 算法将每张照片聚合到各自的事件集中去.

(4) Graham[15]针对照片集也提出了一种名叫 Temporal Similarity 的事件检测算法,其核心在于通过探测照片集所包含的时间戳序列的新颖度峰值来对照片集进行分段.

(5) 最后,本文也对比了利用 Quan 等人[6]所提的文本相似度算法计算得到的相邻会话之间的相似度进行候选会话合并的性能.

图 8 给出了本文所提出的算法与上述 5 种算法在 Precision,Recall 以及 F-Score 这 3 项指标上的对比结果.

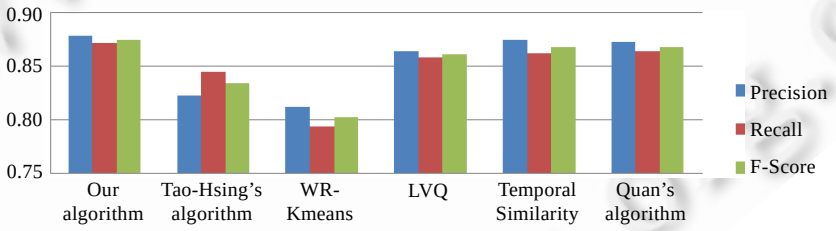


Fig.8 Performance comparison of different algorithms

图 8 不同算法性能指标对比结果

从图 8 可以看出,本文所提出的算法在 Precision,Recall 以及 F-Score 这 3 项指标上都优于另外 5 种算法.究其原因,是因为本文所采用的算法不仅从时间分布的层面对短信息流进行探测得到了最优的划分组合;同时,利用话题模型从语义的层面对中间结果进行了量化,并根据量化指标对中间结果进行了平滑处理.

5 总 结

短信息的使用已经渗透到人们日常生活的方方面面,与此同时,手机的存储能力和处理能力也越来越强大,驻留在手机上的短信息也日益增多.如何挖掘在看似杂乱的短信息背后所隐含的会话信息,是一项具有挑战性的课题.本文首先根据短信息流的时间分布特性,利用层次聚类方法将特定两个会话对象之间产生的所有短信

息划分到潜在的后续会话组合中去,同时,通过比较聚类质量的优劣,遴选出最接近自然分布规律的划分作为最佳后续会话组合.为了对候选会话组合进行优化,进一步地,本文提出了一种新颖的候选会话话题关联度度量算法.综合考虑话题关联度以及时间关联度,在候选会话检测中被分割为不同候选会话单元的同一个会话最终被合并到一起.在真实数据集上的实验结果证明了该算法在 Precision,Recall 以及 F-Score 这 3 项指标上优于其他事件检测算法.

在下一步的工作中,我们会针对多人共同参与同一个会话的模式研究短信息的会话检测,同时考虑针对微博客等新型社会化媒体进行话题分析和事件检测.

致谢 在此,我们向对本文的工作给予支持和建议的诸位老师和同学由衷地表示感谢.

References:

[1] Phan X-H, Nguyen LM, Horiguchi S. Learning to classify short and sparse text & Web with hidden topics from large-scale data collections. In: Proc. of the 17th Int'l Conf. on World Wide Web (WWW 2008). New York: ACM Press, 2008. 91−100. [doi: 10.1145/1367497.1367510]

[2] Cooper M, Foote J, Girgensohn A, Wilcox L. Temporal event clustering for digital photo collections. ACM Trans. on Multimedia Computing, Communications, and Applications (TOMCCAP), 2005,1(3):269−288. [doi: 10.1145/1083314.1083317]

[3] Zhao QK, Mitra P. Event detection and visualization for social text streams. In: Proc. of the Int'l Conf. on Weblogs and Social Media (ICWSM 2007). Colorado, 2007. 26−28. http://www.icwsm.org/papers/3--Zhao-Mitra.pdf

[4] Bollegala D, Matsuo Y, Ishizuka M. Measuring semantic similarity between words using Web search engines. In: Proc. of the 16th Int'l Conf. on World Wide Web (WWW 2007). New York: ACM Press, 2007. 757−766. [doi: 10.1145/1242572.1242675]

[5] Metzler D, Dumais S, Meek C. Similarity measures for short segments of text. In: Amati G, Carpineto C, Romano G, eds. Proc. of the 29th European Conf. on IR Research (ECIR 2007). Berlin, Heidelberg: Springer-Verlag, 2007. 16−27.

[6] Quan XJ, Liu G, Lu Z, Ni XL, Liu WY. Short text similarity based on probabilistic topics. Knowledge and Information Systems, 2010,25(3):473−491. [doi: 10.1007/s10115-009-0250-y]

[7] Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society, Series B (Statistical Methodology), 2000,63(2):411−423. [doi: 10.1111/1467-9868.00293]

[8] Tong HH, Sakurai Y, Eliassi-Rad T, Faloutsos C. Fast mining of complex time-stamped events. In: Proc. of the 17th ACM Conf. on Information and Knowledge Management (CIKM 2008). New York: ACM Press, 2008. 759−768. [doi: 10.1145/1458082.1458184]

[9] Kleinberg J. Bursty and hierarchical structure in streams. Journal of Data Mining and Knowledge Discovery, 2003,7(4):373−397. [doi: 10.1023/A:1024940629314]

[10] Sun HJ, Wang SR, Jiang QS. FCM-Based model selection algorithms for determining the number of cluster. Pattern Recognition, 2004,37(10):2027−2037. [doi: 10.1016/j.patcog.2004.03.012]

[11] Chen LF, Jiang QS, Wang SR. A hierarchical method for determining the number of clusters. Journal of Software, 2008,19(1):62−72 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/19/62.htm [doi: 10.3724/SP.J.1001.2008.00062]

[12] Griffiths TL, Steyvers M. Finding scientific topics. Proc. of the National Academy of Science of the United States of America, 2004,101(Suppl. 1):5228−5235. [doi: 10.1073/pnas.0307752101]

[13] Chang TH, Lee CH. Topic segmentation for short texts. In: Proc. of the 17th Pacific Asia Conf. on Language, Information and Computation. Singapore, 2003. 159−165. http://aclweb.org/anthology-new/Y/Y03/Y03-1018.pdf

[14] Wang L, Jia Y, Han WH. Instant message clustering based on extended vector space model. In: Proc. of the 2nd Int'l Symp. on Advances in Computation and Intelligence (ISICA 2007). LNCS 4683, Berlin, Heidelberg: Springer-Verlag, 2007. 435−443. [doi: 10.1007/978-3-540-74581-5_48]

[15] Graham A, Garcia-Molina H, Paepcke A, Winograd T. Time as the essence for photo browsing through personal digital libraries. In: Proc. of the 2nd ACM/IEEE-CS Joint Conf. on Digital Libraries (JCDL 2002). New York: ACM Press, 2002. 326−335. [doi: 10.1145/544220.544301]

附中文参考文献:

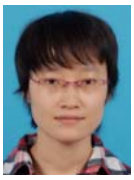
[11] 陈黎飞,姜青山,王声瑞.基于层次划分的最佳聚类数确定方法.软件学报,2008,19(1):62−72. http://www.jos.org.cn/1000-9825/19/62.htm [doi: 10.3724/SP.J.1001.2008.00062]

田野(1981−),男,重庆人,博士生,CCF 学生会员,主要研究领域为文本数据挖掘,社交网络.

王文东(1963−),男,教授,博士生导师,CCF 高级会员,主要研究领域为下一代互联网体系结构,数据挖掘.

饶京海(1975−),男,博士,研究员,主要研究领域为分布式系统,网络服务,语义网络.

王冠(1986−),男,硕士生,主要研究领域为移动互联网.



郭亮(1987−),女,博士生,CCF 学生会员,主要研究领域为社会网络分析,社会搜索.

陈灿峰(1978−),男,博士,研究员,CCF 会员,主要研究领域为短距离无线通信,嵌入式系统,移动互联网.

马建(1959−),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为移动互联网,移动物联网,情景计算.