北京大学学报(自然科学版) 第 52 卷 第 1 期 2016 年 1 月
Acta Scientiarum Naturalium Universitatis Pekinensis, Vol. 52, No. 1 (Jan. 2016)
doi: 10.13209/j.0479-8023.2016.018

一种基于词覆盖的新闻事件脉络链构建方法

付佳兵 董守斌†

华南理工大学广东省计算机网络重点实验室, 广州 510640; †通信作者, E-mail: sbdong@scut.edu.cn

摘要 针对目前构建新闻脉络链只关注新闻脉络链的主题相似性和文档重要性, 而忽略新闻脉络链逻辑连贯性和可解释性的不足, 以及新闻数据集合指数级增长带来的算法复杂度问题, 从词覆盖的角度提出一种新闻脉络链构建方法, 利用新闻的评论信息来定位新闻事件转折点, 用主题相似与稀疏差异的思想以及 RPCA 方法对文档进行逻辑建模, 利用随机游走以及图遍历的方法, 量化并生成可解释且具有很好逻辑连贯性的脉络链。双盲实验表明, 与其他算法相比, 该方法取得较好的效果。

关键词 新闻脉络; 词覆盖; 可解释; 健壮主成分分析; 随机游走

中图分类号 TP391

Constructing a News Story Chain from Word Coverage Perspective

FU Jiabing, DONG Shoubin†

Guangdong key Laboratory of Communications, South China University of Technology, Guangzhou 510640;
† Corresponding author, E-mail: sbdong@scut.edu.cn

Abstract Current studies merely focus on a story chain's similarity of topic relationship and importance of documents, whilst almost ignoring its logical coherency and explainability. Along with algorithm complexity brought about by exponential growth in sets of news data, a story chain from word coverage perspective is constructed, taking advantage of the story comments to position the turning point of each event. The ideas of similarity of topic relationship and sparsity differences as well as RPCA approach are used to conduct logical modeling for the documents. Random walk and graph traversals are adopted to quantify and construct an explainable and logically coherent story chain. The double-blind experiment reveals that proposed method outperforms other algorithms.

Key words story chain; word coverage; explainable; RPCA; random walk

面对大量信息, 读者容易迷失在局部的信息中, 逐渐丧失对信息的全局把控。因此, 构建新闻脉络链成为当今迫切需要解决的问题。构建新闻脉络链是对新闻事件故事发展脉络结构的捕捉, 因此新闻故事发展需要具备很好的逻辑发展特性和可解释性机制。现有脉络链构建研究存在以下三方面的问题: 1)脉络陷入局部, 没有全局观; 2)脉络关注主要集中在文档重要性、相关性以及相似性等一维内容层面, 忽视文档之间的二维逻辑连贯语义层面; 3)脉络构建复杂度高, 多数研究为保证脉络的全局特性而对整个数据集进行多次迭代, 缺乏对数据集大小进行有效降级。

基于以上问题, 本文提出一种新闻脉络链构建方法, 将脉络构建视为词覆盖问题, 在依赖新闻内在逻辑性进行词覆盖的同时, 也完成了结构化逻辑且可解释的脉络构建。本文算法可根据读者感兴趣的新闻热点事件, 自动生成该新闻事件的新闻脉络链, 能够帮助读者把控新闻事件的全局发展脉络。例如, 读者对马航(MH370)事件感兴趣, 那么算法给出的可能输出如图 1 所示。

广东省前沿与关键技术创新专项(2014B010112006)和广东省产学研省部合作专项资金(2013B090500087)资助
收稿日期: 2015–06–19; 修回日期: 2015–08–17; 网络出版日期: 2015–09–30

1) 马来西亚航空称与 1 架载 239 人飞机失去联系

2) 马航载 239 人飞北京航班失联机型为波音 777

3) 马来西亚航空失去联系飞机上有 160 名中国人

4) 马航官员: 约在 120 海里处飞机与地面失去联系

5) 越军高层称不明物或为失联客机的一部分

6) 调查组称马航失联客机空中解体可能性增大

图 1　马航事件新闻脉络例子

Fig. 1　An example story chain of MH370 news event

1　相关工作

信息过载使得研究人员开始寻找各种信息中隐含的故事发展脉络，比如微博[1−4]，新闻[5−7]，论文[8−9]以及邮件[10]。故事生成[11−12]较早开始对故事脉络进行定义和建模，但只关注规则模版的设定及其推演。事件检测[13−14]尝试发现信息中隐含的新闻事件，但并不尝试将其连接起来形成完整的脉络发展。文献[2−3]尝试解决脉络的连接问题，但基于局部贪心的思想缺乏全局观。文献[5]构建的是全局脉络，其代价是需对整个候选新闻集合进行迭代，严重影响算法的可扩展性。事件追踪[15]利用有监督的机器学习算法，将新闻划分到大的新闻子类，但是需要进行人工标注，难度大。事件追踪与 TDT[16](主题检测和追踪)的思想类似，不同之处在于后者将事件追踪抽象为主题追踪。TDT 致力于生成文本的故事链，主要包括五大任务：故事分段、主题追踪、主题检测、起始故事检测以及链接检测。大部分 TDT 的研究主要关注文本相关性或者相似性，在其基础上进行文本分类和聚类，并未考虑文本间的逻辑转换关系[17]。文献[18]通过考虑文本间相互作用构建主题结构图，基于结构图对主题变化趋势进行追踪。类似地，文献[13−19]通过发现新闻事件子类，并利用其相互依赖关系构建图结构(动态主题模型)，但是均未考虑图结构的连贯性问题。

MDS(多文本总结)通过选取代表性的句子，以时序的方式构建时间轴，完成对文档集合的总结，为构建脉络提供了一个文本总结的新思路。句子的选择标准方法有很多种，基本上分为三大类：一类是句子本身的属性，比如文献[20]用句子的信息含量(通过最大化信息含量高的词)，文献[21−22]用句子的相关性、覆盖性、连贯性以及多样性，文献[23]用句子的不确定性，文献[24]用句子的代表性和差异性；另一类是句子的结构属性，文献[19−25]通过构建句子图谱，使用图谱的中心化句子节点作为候选句子；第三类是前两类的综合，文献[26−27]通过矩阵分解对句子进行潜在的语义分析。这三大类方法(包括第三类的潜在语义)以及 TDT 都无法给生成的脉络结果提供可解释性，而缺乏可解释机制会大大增加对脉络链的理解难度。

2　算法总体设计

2.1　词覆盖方法

典型的搜索查询任务流程如下：给定查询词集合 q，搜索引擎在数据库中逐个扫描并返回覆盖查询词 q 的文档集合。简单归纳可知，搜索引擎的工作实质上是基于查询词的文档覆盖。新闻脉络链，其反映的是新闻事件的逻辑发展，与搜索引擎有相同亦有不同：相同之处是都可看成文档覆盖问题；不同之处是文档覆盖的查询词不再是用户输入的新闻事件查询词，取而代之的是能反映该查询词所对应新闻事件新闻脉络的词集合 Q。一旦结果文档集合 D 能覆盖反映该新闻事件脉络的词集合 Q，那么文档集合 D 即是结果新闻脉络。对比异同，新闻脉络链的构建引擎实质上是附加查询词扩展层的搜索引擎。

2.2　设计框架

基于查询词返回结果文档的研究已经很成熟，因此构建新闻脉络图的关键是快速定位词集合 Q，即完成从新闻事件查询词 q 到 Q 的扩展。Q 很难通过 q 的直接扩展得到，因为在未彻底了解新闻事件前，无法预先得知 Q；即便了解，由于理解上的主观性，也无法确切得知 Q。因此，Q 只是概念化的词集合，无从获知。若 Q 已知，则新闻脉络已知。由于无法“正面”得知 Q，本文则通过采用不断缩小候选词集合的方法，不断逼近真实的 Q，从而间接获取 Q。

新闻脉络链由许多形如 $A \rightarrow B$ 的逻辑转移组成，其中新闻文档 A 和 B 是逻辑转移主体，逻辑转移词集合“$\rightarrow$”代表转移依据，因此 $A \rightarrow \mathrm{B}$ 描述新闻文档 A 依据逻辑词集合 W 转移到新闻文档 B。本算法预先锁定包含逻辑转移主体的文档集合 $D_c=\{d_1, d_2, \ldots, d_c\}$，然后进一步筛选逻辑主体 $D_L=\{d_1, d_2, \ldots, d_m\}$，最后具体化转移依据得出最终的 $W_L=\{w_1, w_2, \ldots, w_n\}$，在逼近 Q 的同时也完成覆盖 Q 且结构化 Q 的过程，具体流程如图 2 所示。

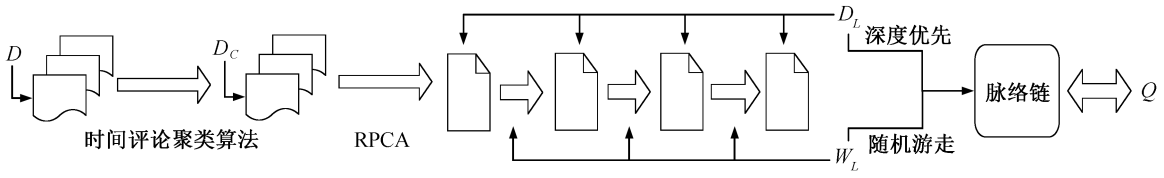


图 2 算法流程
Fig. 2 Procedure of our algorithms

2.3 算法描述

2.3.1 时间评论聚类算法

某段时间特别受关注的新闻很可能是新闻事件的转折点，即所需捕捉候选新闻文档集合，因此可利用新闻报道的用户关注度来定位新闻事件转折点。用户关注行为有强弱两种：强用户评论行为和弱用户浏览行为，一般浏览行为很难准确获取。文献[28]表明用户的评论和浏览行为存在强一致性，即评论行为越多，浏览行为也越多，因此可通过度量评论行为来达到度量评论和浏览行为的目的。

K-means 算法[29]是最简单易行的聚类算法之一，它能够快速有效地处理大规模数据，运用十分广泛。本文用二维元组<评论数量，评论时间>表示样本点 x^i，采用K-means对样本集合$\{x^1, x^2, \ldots, x^n\}$进行聚类，剔除小于10 个样本的小型类别，保留剩余类别所有样本点。

2.3.2 文档建模算法

主题模型 pLSI[30]和 LDA[31]广泛用于文档建模领域。给定一篇文档，形式化描述如下：

$$D = T + N,$$

其中 T 代表低秩主题部分，N 代表高斯噪音部分。

但这并不总符合现实情况。如图 3(a)所示，文档中常出现一些频率异常高的词，因此词频分布误差并不是主题模型所假设的噪音方差小且服从高斯分布，而是高频噪音误差。高频噪音误差并非没有价值，相反地，它恰恰最能反映文档间的差异性。基于此，如图 3(b)所示，对文档进行低秩主题部分–稀疏高频部分建模，形式化表示如下：

$$D = A + B,$$

其中 A 代表低秩主题部分，B 代表稀疏高频部分。为便于后续描述，将 A 表述为主题部分，B 表述为稀疏部分。

文档集合 D 的分离过程是在尽可能用低秩主题模型拟合文档集合的基础上，最小化 S 中的非零项个数。只有在尽可能剥离共有主题部分之后，S 才能准确描述文档集合中文档之间的差异性部分，因此分离定义如下：

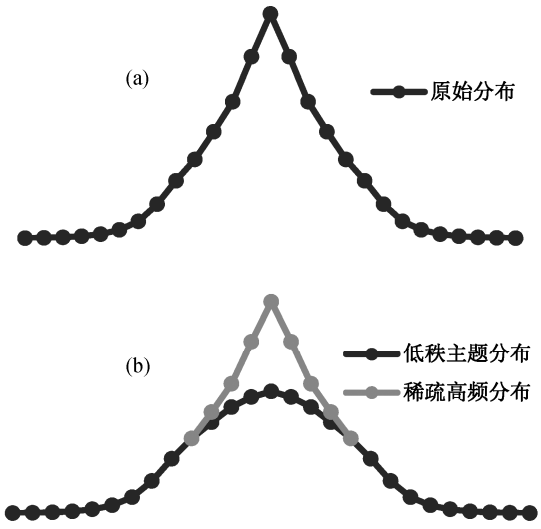


图 3 词频分布
Fig. 3 Distribution of word frequency

$$\min_{A,B} \left\{ \|S\| : \boldsymbol{A} + \boldsymbol{B} = \boldsymbol{D} \text{ 且 } \mathrm{Rank}(\boldsymbol{B}) < K \right\}, \qquad (1)$$

其中 $\boldsymbol{D} \in \boldsymbol{N}^{m\times n}$ 是待分解文档–词频矩阵，m 和 n 分别是文档和词的数量，$\boldsymbol{A}, \boldsymbol{B} \in \boldsymbol{R}^{m\times n}$ 对应于主题和稀疏部分，$\|\cdot\|$ 计算矩阵中非零项的数量。式(1)假设给定文档集合在主题数为 K 的情况下可对原文档集合进行足够好的描述。最后，采用 RPCA[32](健壮主成分分析)求解式(1)，得出$\boldsymbol{A}$ 和 $\boldsymbol{B}$。

2.3.3 随机游走算法

关系作用传递分为显式和隐式：显式关系传递指两篇文档包含相同的词；隐式关系传递指两篇文档中的前一篇包含这个词，而后一篇文档缺失这个词。后者的隐式关系传递是指同一隐含语义，在文档中因作者、文章题材等影响而会采取不同表达。比如一篇文章包含律师，另外一篇包含诉讼或者法庭，即使后一篇文档通篇不包含律师这个词，但两

篇文章本质上仍隐含转移关系。本文采用随机游走模型来对文档的显式和隐式传递关系进行建模，如图 4(a)所示，分析 $d_1 \to d_4$，显示关系传递为 $d_1 \to w_2 \to d_4$，隐式关系传递为 $d_1 \to w_1 \to d_3 \to w_4 \to d_4$ 和 $d_1 \to w_1 \to d_3 \to w_3 \to d_4$。可以看出，随机游走模型能很好地融合显式和隐式文档关系。

文献[5]定义 Influence $(d_i, d_j|w)$，即两篇文档 i 和 j 基于词 w 的跳转概率，通过 Influence 将转移依据在两篇文档转移中的影响进行量化。为了计算 Influence $(d_i, d_j|w)$，文献[5]的定义如下:

$$\prod_i(v)=\varepsilon\cdot 1(v=d_i)+(1-\varepsilon)\sum_{(u,\ v)\in E}\prod_i(u)P(v|u), \quad (2)$$

$$\prod_i{}^{w}(v)=\varepsilon\cdot 1(v=d_i)+(1-\varepsilon)\sum_{(u,v)\in E}\prod_i(u)P^{w}(v|u), \quad (3)$$

其中 $\prod_i(v)$ 为随机游走的驻留分布，ε为控制参数，$P(v|u)$ 为 v 到 u 的概率。如果 v 是文档，u 是词，则 $P(v|u)$是词 u 在文档中的 tf-idf 值; 如果 v 是词，u 是文档，则 $P(v|u)$是归一化 u 在所有文档中的 tf-idf 值后得出的值。$P^w(v|u)$是去除节点 w 的出边(原始 w 的出边变为指向 w 自己，不再传递重要性给其他节点)后的 $P(v|u)$，即不通过 w 节点，节点 v 到 u 的概率，那么通过式(4)即可间接获取 Influence$(d_i, d_j|w)$，即词在文档 i 和 j 转移中的贡献，见图 4 (b)。本文研究如何进行逻辑转移。鉴于逻辑转移更多发生在文档间的差异部分，因此两篇文档间发生转移的词 w 被限定在文档的差异部分包含的词，这样得出的 Influence $(d_i, d_j|w)$ 可更加准确地描述词在文档 i 和 j 转移中的逻辑贡献。由于差异部分的稀疏性质，算法的复杂度可进一步降低，可扩展性得到加强。

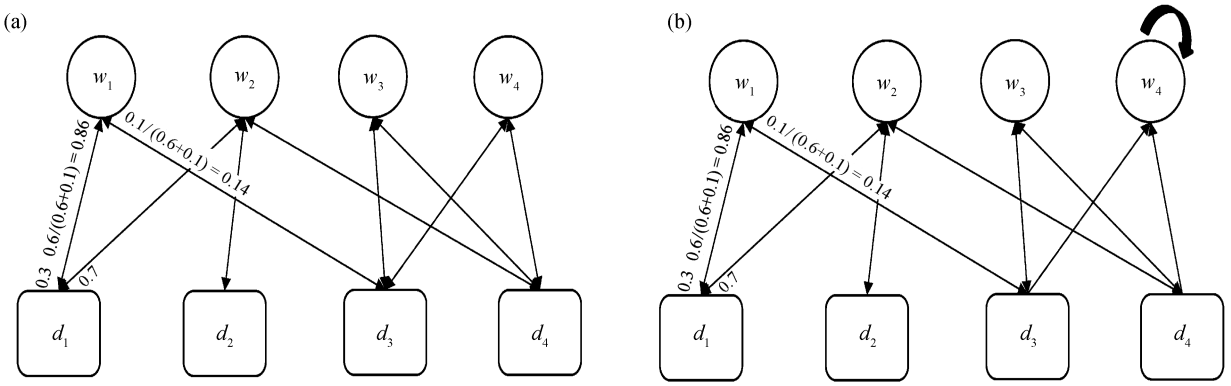
$$\text{Influence}(d_i, d_j \mid w) = \prod_i(d_j) - \prod_i{}^{w}(d_j) \ 。 \quad (4)$$

2.3.4　链生成算法

两篇文档发生逻辑转移，相似性是必要条件，充分条件是文档间必须有差异，过于相似或者过于不相似都将导致文档间相似和差异的比例不均衡，直接影响文档间转移的质量。过于相似文档间发生转移类似文本主题聚类，而聚类并不能反映其逻辑意义。过于不相似文档间发生转移类似随机选取文档进行转移，得出的结果将因为噪音的影响而失真。因此本文将文档的主题相似作为判定转移的条件(降低噪音)，之后通过差异部分具体量化转移。这种策略将大幅度提高转移结果的准确性。新闻脉络链由多个逻辑转移构成，因此本算法通过计算文档间主题相似度，在此基础上建立时序有向图。如图 5(a)所示，节点代表文档，边的粗细是定义在主题空间的节点间正弦距离，定义如下:

$$\text{sim}(\bar{d}_i, \bar{d}_j) = \frac{\sum_{k=1}^{n} d_{ik}\cdot d_{jk}}{\sqrt{\sum_{k=1}^{n} d_{ik}\cdot d_{ik}}\sqrt{\sum_{k=1}^{n} d_{jk}\cdot d_{jk}}}, \quad (5)$$

其中 $\bar{d}_i$ 是文档在主题部分 A 所构成的主题词空间的映射，也即 A_i (矩阵 A 中的第 i 行)，d_{ik} 是 A_{ik}。过强或过弱的边被剔除(灰色)，假定保留 $\text{sim}(\bar{d}_i, \bar{d}_j) \in (\alpha, \beta)$ 的边，其中 $\alpha, \beta \in (0,1)$，得到有向时序图的两条候选链(图 5(b))，最后筛选出具有最高逻辑连贯指标的结果链。根据木桶效应，事物逻辑连贯由组成该事物的多个部分中逻辑连贯性最薄弱一环决定，因此文献[5]将其定义如下:



(a) 随机游走模型; (b) 剔除 W_4 节点出边的游走模型

图 4　随机游走模型

Fig. 4　Random walk model

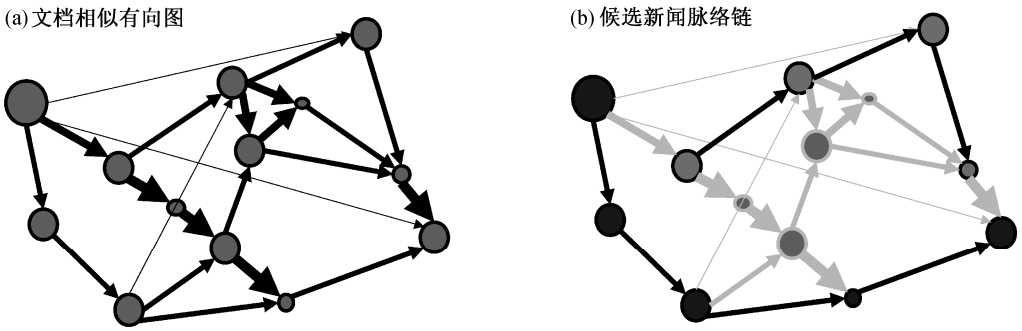


图 5 文档相似有向图
Fig. 5 Document's similarity directed graph

$$\text{Coherence}(d_1, ..., d_n) = \min_{i=1,...,n} \sum_w \text{Influence}(d_i, d_{i+1} \mid w) \cdot 1(w \in d_i \cap d_{i+1}) 。 \quad (6)$$

给定起点文档和重点文档，即可根据逻辑性指标，在时序图中遍历找到最佳脉络图，并附加可解释的转移依据。

3 实验设计

3.1 数据集

本实验使用的数据集来自新浪网新闻专题搜索引擎，通过抓取基于关键字 MH370 搜索返回的结果，得到与马航相关的新闻事件文档集合。对新闻去重后，对文档集合的评论信息进行分析抓取，形成最终的原始文档集合，具体描述见表 1。

3.2 时间评论聚类算法有效性验证

通过对新闻–用户评论数据进行 K-means 聚类分析，将得到的结果与参照的人工编辑脉络链进行对比，对假设“某新闻是新闻事件发展转折点的可能性大小正比于用户对该新闻的关注行为强度”做可行性假设。剔除评论数低于 1500 的样本点，对原始数据进行聚类并得到多个时间簇，时间簇所包含的时间点(以天为单位)即预测的新闻核心事件发生日期。对比人工编辑新闻链中新闻文档发表时间发现，聚类得到的新闻转折点发生时间与人工编辑的基本上吻合，如图 6 所示。

从 3 月 8 日到 12 月 24 日，共有 291 个日期，人工编辑提供 17 个标准日期答案，聚类算法提供 58 个预测日期并命中其中16个，唯一未命中的是 5 月

表 1 原始数据集描述
Table 1 Description of original dataset

时间范围	新闻数	评论信息	
		有	没有
2014–03–08—2014–12–24	2895	2756	139

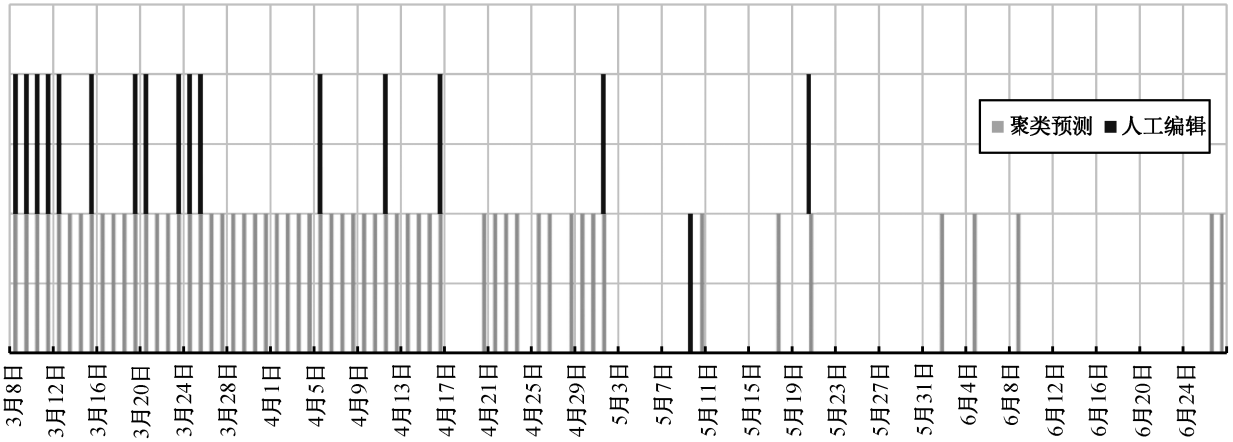


图 6 核心新闻发布时间预测
Fig. 6 Core news publishing date prediction

9 日，但算法提供非常接近的预测日期 5 月 10 日，因此预测错误可能由新闻的延迟导致。为定量描述聚类算法有效性，定义召回率和准确率的指标如下：

$$召回率 = \frac{算法命中的日期个数}{标准答案日期个数}, \tag{7}$$

$$准确率 = \frac{算法命中的日期个数}{算法给出的总共预测日期个数}。 \tag{8}$$

通过计算可知，未做预测之前召回率为1，准确率为 5.84%；经过利用用户评论信息聚类预测后，召回率为 94.12%，基本上接近 1，准确率为 27.59%，数据集合数量由 2895 变为 494，缩小至原来的 1/5。在保证不丢失新闻脉络信息的同时，大大减少了候选新闻数据集合的大小，数据集的数量级也有大幅度降低，提升了算法的可扩展性。

3.3　文档建模效果分析

本节实验数据集合为时间聚类方法得到的候选新闻集合。过滤掉词频出现小于 10 的低频词，将实验数据集转换成为$D \in N^{494\times1010}$，通过RPCA 得到矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{B}$，预处理新闻集合的具体描述见表 2。

图 7 描述的是源文档 d_0：《马航机场员工推搡中国记者 大声骂人竖中指》。利用随机游走模型计算文档 d_0 到另两篇文档{d_1, d_2}转移中转移依据 w(比如推搡)的影响，可以看出 $d_0 \rightarrow d_1$ 基于每一个词转移的概率都接近 0，结果是合理的。$d_0 \rightarrow d_2$ 基于每个词转移的概率之间差别较大，选取其中几个影响较大的词：推搡、员工、中国、道歉、马航，可以看出结果词能较好地解释两篇文章转移依据(显式和隐式)。至于“记者”没有出现在转移依据中，是因为“记者”在新闻文档中虽然出现频次高但意义小，比如“据新华社记者报道”和“记者某某报道”，因此在分词预处理阶段，连同词“报道”同时被过滤掉，不参与后续转移。

表 2　候选数据集描述

Table 2　Description of original dataset

时间范围	新闻数	词数(8934)	
		词频≥10	词频<10
2014–03–08—2014–06–27	495	1010	7924

3.4　链评价

实验设定 α= 0.5，β= 0.8，得出候选结果链条 Coherence 指标最高为 0.01008646。限于篇幅，只列出“本文算法_结果 2”及其序号 5→序号 6 文档的转移依据“疑似 (7.0305495×10^{-4})，残骸 (6.433595×10^{-4})，祈祷 (3.8924068×10^{-4})，并非 (3.8924022×10^{-4})，海面 (3.1231903×10^{-4})”，词后的数字代表该词在文档间转移的量化影响，具体结果如图 8 所示。

新闻事件逻辑脉络链具有很强主观特性，比如链的可读性、逻辑性以及可解释性等，无法类比搜索引擎或者推荐引擎评价的标准和客观。ROUGH 系列评价指标是多文本总结领域常用指标，但它较主观，不能反映真实的效果。鉴于本文构建脉络链的高主观特性(后续实验已证实，不同用户对同一脉络结果打分相差很大)，本文通过用户调查对生成的链进行评价。

脉络链构建的工程性质使得相关算法虽然多，但基本上不公开源码，因此论文方法不可再现。为避免因个人工程实现原因导致对比算法效果降低，

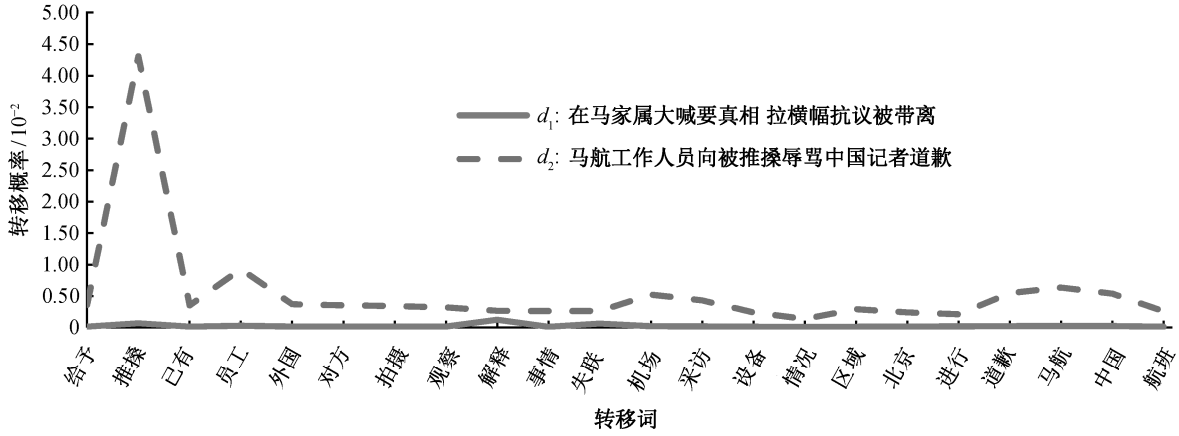


图 7　文档词转移影响

Fig. 7　Word influence of document transferring

1) 越南南部发现马航失联客机信号
2) 越南决定派直升机赴发现漂浮物海域核查
3) 新加坡搜救船发现漂浮物或可捞起
4) 越船发现疑似飞机残骸黄色物体 或为部分舷窗
5) 越南军方发现疑似失联客机残骸的漂浮物
6) 马航: 越南海域发现漂浮物与失联航班无关

图 8 新闻脉络链最终结果
Fig. 8 Final result of news storyline

多角度比较算法效果，本文算法与 3 个经典的算法进行比较，三者分别代表随机性、主题聚类特性以及相关性。

1) 随机选取算法：代表随机性的思想，随机选取固定数目的新闻文档作为脉络链的候选文档。

2) K-means 聚类算法：代表主题聚类特性，在话题追踪领域广泛应用。通过将文档利用主题分量进行描述，并基于这个主题向量空间进行 K-means 聚类，得出聚类簇，然后选择最靠近类簇中心的文档作为脉络链的候选文档。

3) 最短路径算法：通过文档的主题余弦相似度构建一个图，权重为相似度，寻找权重最大的路径，这是一个局部算法。

4) 本文算法：设定链的长度为6，生成候选链。

本文对 20 名大学生进行双盲问卷调查，为其提供 5 个待评价脉络链(图 8 及附录)，受调查者需回答两个问题：一个是知识量，即读完能够对事件脉络的了解程度；另一个是逻辑连贯性，即展示脉络链的逻辑连贯程度。二者分数都是 1~5 之间的整数。为保证问卷调查的公平和客观，本文给出两个待评估的链(未将脉络链的可解释加上，若加上，效果会更好)，其余 3 个算法各给出一条脉络链，并提供人工编辑的脉络链供参考，结果见表 3。

从实验结果可看出，本文算法得出的两个脉络结果均在知识量方面高于或者略高于其他算法，逻辑连贯性的指标远高于其他算法，说明本文算法能在保证较高的知识量的前提下，较好地捕捉脉络发展之间的逻辑性，其他结果见附录。

表 3 基于双盲用户调查的 4 种算法评估结果
Table 3 Evaluated result of four algorithms in double-blind user study

算法	知识量(平均)	逻辑连贯性(平均)
随机选取算法	3.50	2.90
K-means 聚类算法	3.45	3.00
最短路径算法	2.90	3.25
本文算法_结果 1	3.65	4.15
本文算法_结果 2	4.15	4.13

4 结语

基于现有脉络图存在的三方面不足，本文从词覆盖角度考虑逻辑脉络链生成问题。在保证新闻脉络基本无损的情况下，利用新闻评论信息对数据集进行 5 倍压缩。通过对文档进行 RPCA 建模，利用主题相似与稀疏差异的思想对文档进行逻辑建模并量化，形成可解释且具较好逻辑连贯性的脉络链，解决了贪心相似或者主题聚类的脉络局部化问题。本文方法简单，最终结果脉络链取决于用户给定的起始和终点文档，无须每次对整个集合进行迭代。

本文构建脉络链的最终评价标准是逻辑连贯性，而逻辑连贯性取决于具体转移词的累加。选取哪些词以及累加转移如何计算，都可由用户个性化指定。如用户喜欢逻辑跳跃缓慢的链，就返回相邻逻辑转移增长相对平缓的链，反之亦然。与此同时，用户也可对词转移的影响进行人为指定。比如用户喜欢某些特定词，即可人为调高相应词的转移影响，结果链包含用户喜欢的事件的转移几率就会提高。在为用户生成逻辑连贯且可解释的脉络链的同时，利用链可解释性的展示可获取用户的反馈，因此研究如何为用户提供个性化的逻辑脉络链是下一步要做的工作。目前结果链是单链，反映事件的某一个侧面，将来可考虑构建成脉络图，使之包含的信息更加全面。因此，如何融合多条链，也是将来要考虑的工作。

参考文献

[1] Lin Chen, Lin Chun, Li Jingxuan, et al. Generating event storylines from microblogs // Proceedings of the 21st ACM International Conference on Information and Knowledge Management. New York, 2012: 175–184

[2] Sakaki T, Okazaki M, Matsuo Y. Earthquake shakes twitter users: real-time event detection by social sensors // Proceedings of WWW 2010. Raleigh, 2010: 851–860

[3] Shamma D A, Kennedy L, Churchill E F. Peaks and persistence: modeling the shape of microblog conversations // Proceedings of CSCW 2011. Hang-

zhou, 2011: 355–358

[4] Shi J, Malik J. Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(8): 888–905

[5] Shahaf D, Guestrin C. Connecting the dots between news articles // Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, 2010: 623–632

[6] Shahaf D, Guestrin C, Horvitz E. Trains of thought: Generating information maps // Proceedings of the 21st International Conference on World Wide Web. New York, 2012: 899–908

[7] Shahaf D, Yang J, Suen C, et al. Information cartography: creating zoomable, large-scale maps of information // Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, 2013: 1097–1105

[8] Shahaf D, Guestrin C, Horvitz E. Metro maps of science // Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, 2012: 1122–1130

[9] El-Arini K, Guestrin C. Beyond keyword search: discovering relevant scientific literature // Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, 2011: 439–447

[10] Lewis D D, Knowles K A. Threading electronic mail: a preliminary study. Information Processing and Management, 1997, 33(2): 209–217

[11] Turner S R. The creative process: a computer model of storytelling and creativity. Hillsdale: Lawrence Erlbaum Associates Inc, 1994

[12] Niehaus J, Young R M. A computational model of inferencing in narrative // AAAI Spring Symposium'09. Stanford, 2009: 75–82

[13] Kleinberg J. Bursty and hierarchical structure in streams. Data Mining & Knowledge Discovery, 2003, 7(4): 373–397

[14] Yang Y, Ault T, Pierce T, et al. Improving text categorization methods for event tracking // SIGIR 2000. Athens, 2000: 65–72

[15] Masand B, Linoff G, Waltz D. Classifying news stories using memory based reasoning // SIGIR. Copenhagen,1992: 59–65

[16] Allan J. Introduction to topic detection and tracking // Topic Detection and Tracking. Norwell, MA, 2002: 1–16

[17] Lavrenko V, Allan J, DeGuzman E, et al. Relevance models for topic detection and tracking // Proceedings of HLT 2002. San Francisco, 2002: 115–121

[18] Morinaga S, Yamanishi K. Tracking dynamics of topic trends using a finite mixture model // Proceedings of SIGKDD 2004. Seattle, 2004: 811–816

[19] Kumar R, Mahadevan U, Sivakumar D. A graph-theoretic approach to extract storylines from search results // Proceedings of SIGKDD 2004. Seattle, 2004: 216–225

[20] Yih W, Goodman J, Vanderwende L, et al. Multi-document summarization by maximizing informative content-words // The 20th International Joint Conference on Artificial Intelligence. Hyderabad, 2007: 1776–1782

[21] Yan R, Wan X, Otterbacher J, et al. Evolutionary timeline summarization: a balanced optimization framework via iterative substitution // Proceedings of SIGIR. New York, 2011: 745–754

[22] Yan Rui, Jiang Han, Lapata M, et al. i, poet: automatic Chinese poetry composition through a generative summarization framework under constrained optimization // Proceedings of IJCAI 2013. Beijing, 2013: 2197–2203

[23] Wan Xiaojun, Zhang Jianmin. CTSUM: extracting more certain summaries for news articles // Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval. New York, 2014: 787–796

[24] Wei F, Li W, Lu Q, et al. Query-sensitive mutual reinforcement chain and its application in query-oriented multi-document summarization // Proceedings of SIGIR 2008. New York, 2008: 283–290

[25] Li J, Li L, Li T. MSSF: a multi-document summarization framework based on submodularity // Proceedings of SIGIR 2011. Beijing, 2011: 1247–1248

[26] Wang D, Li T, Zhu S, et al. Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization // Proceedings of SIGIR 2008. New York, 2008: 307–314

[27] Lee D, Seung H. Algorithms for non-negative matrix factorization // Advances in neural information processing systems, NIPS 2001. Vancouver, 2001:

556–562

[28] Mishne G, Glance N. Leave a reply: an analysis of weblog comments // Third Annual Workshop on the Weblogging Ecosystem. Edinburgh, 2006: 1–8

[29] Mcqueen J. Some methods for classification and analysis of multivariate observations // Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Berkeley, 1967: 281–297

[30] Bai B, Weston J, Grangier D, et al. Supervised semantic indexing. Lecture Notes in Computer Science, 2009, 5478: 761–765

[31] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation. J Mach Learn Res, 2003, 3: 993–1022

[32] Candès E J, Li X, Ma Y, et al. Robust principal component analysis?. Journal of the Acm, 2011, 58(3): 219–226

附录 脉络链构建结果

随机选取算法	K-means 聚类算法
调查组称马航失联客机空中解体可能性增大。	马来 1 家 18 口人因孙子生日改签未登上失联客机。
马内政部称利用失窃护照登机者为亚裔人士。	我国搜救船舶搜救方案确定。
大马号令近 1800 艘渔船约 2 万渔民加入客机搜救。	马航称失联客机向卫星发送数小时信号信息不实。
马航: 家属提供的可以播通的电话属实。	ABC: 失联客机曾做出“战术躲避动作”。
运-20 总设计师:马航 MH370 航班应是瞬间崩溃。	外交部就澳大利亚发现疑似马航客机物件答问。
中方称在南印度洋发现疑似漂浮物 呈马蹄形。	反恐专家深度解读:飞机缘何飞往澳西南海域。
本文算法_结果 2	**最短路径算法**
马来西亚航空称与 1 架载 239 人飞机失去联系。	马航称请家属做好心理准备 有家属现场失声痛哭。
马航载 239 人飞北京航班失联 机型为波音 777。	马航称请家属做好心理准备被砸水瓶 有家属昏倒。
马来西亚航空失去联系飞机上有 160 名中国人。	马航失联客机 1 名家属与马大使馆交涉后晕倒。
马航官员:约在 120 海里处飞机与地面失去联系。	中国家属在抗议 马来西亚大使在沉默。
越军高层称不明物或为失联客机的一部分。	马来学生在台遭陆生大骂: 满口谎言的畜生。
调查组称马航失联客机空中解体可能性增大。	MH370 家属欲募集 500 万美元悬赏线人曝内。