# 新型的面向新闻评论摘要采集算法①

师 昕, 赵雪青

(西安工程大学 计算机学院, 西安 710048)

**摘 要**: 为了让读者可以更快地获取所有新闻评论中最有代表性的观点, 提出一种新的新闻评论摘要采集算法, 并依此设计出评论摘要采集系统. 该算法将有效地结合聚类算法和排序算法, 首先, 使用改进的 Borderflow 算法对所有评论聚类; 其次, 采用类 PageRank 算法对聚类中的评论进行排序, 选出排名最前的几条评论; 最后, 利用 MMR 算法对 PageRank 算法选出的所有评论进行再次排序, 并选取名次最高的 K 条评论作为评论摘要. 通过仿真实验得到的 NDCG 和 MAP 数据表明, 使用本文算法得到的评论摘要具有更好的有效性和准确性, 更符合读者直观感觉.

**关键词**: 评论摘要; BorderFlow 算法; PageRank 算法; MMR 算法

## Novel News Article Comments Summarization Algorithm of Computer Engineering and Applications

SHI Xin, ZHAO Xue-Qing

(Department of Computer Science, Xi'an Polytechnic University, Xi'an 710032, China)

**Abstract**: In order to make the readers get the most informative and representative opinions efficiently among the news comments, this paper proposes a novel news article comments summarization algorithm and then designs an article summarization system, which combines the clustering algorithm with the ranking algorithm. First, it groups comments using the modified BorderFlow clustering algorithm. Second, for each group, it uses the similar PageRank algorithm to score and rank comments, and selects top comments in each cluster as representation. At last, it ranks the selected comments by MMR algorithm and displays the top-K comments as the comments summarization. According to the experimental statics of NDCG and MAP data, the proposed method meets the intuitive sense of readers more. Meanwhile, it shows the better effectiveness and accuracy theoretically.

**Key words**: comments summarization; BorderFlow algorithm; PageRank algorithm; MMR algorithm

随着 web2.0 技术的发展, 在阅读新闻后, 读者可以随意地在网页新闻文章后留下自己的评论并阅读他人留下的评论, 以此获得不同的观点和信息. 这也成为读者表达自己观点的途径之一, 这些评论中有一些是非常有见地的, 而且不仅仅表达了读者的观点, 有时也会包含一些时下热门话题的信息和讨论, 因此读者十分热衷于阅读带有评论的新闻文章. 根据文献[1]中所做的用户调查, 读者阅读的新闻评论, 有可能影响甚至改变他们的观点或看法.

但是, 在线新闻文章的评论数量, 特别是在那些比较大的网站中, 是非常庞大且实时增长的. 根据调查, 雅虎新闻页面每篇新闻的平均评论数量为 1059 条, 而这些评论中有很大一部分是没有意义或者重复的, 这就导致读者要想阅读完所有的评论会花费很多时间在没有意义的内容上. 因此, 如何帮助读者在短时间内更有效地从大量评论中获取有用的信息就成为一项研究热点.

近年来, 提出了一些针对减轻读者阅读负担的评论向导方法. 一类是依赖网站读者的评分系统, 比如豆瓣网允许读者对所有评论评分并将得分最高的评论显示在最前面. 这种系统需要大量的读者参与, 而且由于每个人的观点都不尽相同, 因此最终的结论可能

① 收稿时间:2016-04-12;收到修改稿时间:2016-05-19 [doi:10.15888/j.cnki.csa.005530]

并不具代表性; 另一类方法是系统自动生成最热门的关键字, 比如 CSDN 网站使用一系列标签来引导读者, 尽管这种方法可以生成比较具有代表性的关键字, 但是由于关键字本身缺少足够的文本信息, 读者很难从抽象的关键字中获得对整个评论的详尽理解.

目前, 将聚集算法和排序算法结合的评论摘要采集系统具有明显优势, 该领域内有许多学者在进行研究, 如文献[2]和文献[3], 且该方法被证明是非常有效的. 因此, 本文旨在寻找一种可以帮助读者快速得到所有评论中具有代表性信息的方法, 提出了一个新的针对新闻评论的摘要采集系统.

为了从大量的评论中很快得到最有代表性的评论, 以便访客可以在较短时间内得到最有效的信息, 考虑到访客评论观点种类的多样性, 使用聚集算法可以将最相似的评论聚集成类, 这样就可以认为同一个类的评论代表一个主题. 文中提出的针对新闻评论的摘要采集算法, 选取一篇文章及其所有评论并最终显示出最有代表性的 K 条评论作为评论摘要. 首先, 该系统对所有评论使用 BorderFlow 聚集算法[4], 根据余弦相似度把它们分别聚集为类; 其次, 使用 PageRank 算法[5]对每一类中的所有评论进行排名, 选出最有代表性的 N 条评论; 最后, 使用 MMR 算法[6]对每一类中选出的代表性评论再进行排名, 并最终选出若干条最有代表性的 K 条评论, 并将这些评论作为这篇文章的评论摘要. 由于聚集算法的使用, 每一类中的评论可以代表一种相近的观点, 因此最终选出的 K 条评论可以近似代表评论中不同的 K 种观点.

## 1 算法描述

### 1.1 变量定义

根据给出的数据库, 我们定义 $N$ 为数据库中所有新闻文章的集合, $N$={$n1,n2,...nn$}. 每一篇新闻文章中的评论集合定义为 $C$, $C$={$c1,c2,...cn$}. 那么 $C\_N$ 就代表新闻集合 $N$ 中的所有评论(即是 $C$ 的集合). 本文提出方法的目标, 则是提取出一个集合 $C$ 的子集 $Sc$={$s1, s2, ...sn$}, 这个集合包含有 $K$ 个最具有代表性的评论, 也即是新闻文章的评论摘要, 其中 $K$ 是由用户设定的变量.

### 1.2 评论相似性

本文中介绍的算法中的第一步是对所有评论根据内容相似度聚类, 其中最直接的方法是计算基于评论向量的余弦相似度. 为了得到评论向量, 首先需要抽取关键字并计算其特征值.

**抽取关键字:** 关键字选择在聚类中是非常重要的一步, 不同的关键字常常会导致不同的聚类结果. 在本文设计的系统中, 随着评论数量的上升, 评论词语的数量也在增多, 而使用所有词语作为关键字会导致数据量非常大, 非常耗费空间和时间. 因此, 为了减少数据量, 我们选用在一个评论文件中出现过四次以上的词语作为关键字.

**计算特征值:** TD-IDF 加权法[7]是最常用的计算评论向量特征值的方法. 该方法仅考虑了一条评论中关键字的信息, 但是事实上, 出自于同一个新闻资源的新闻和评论之间, 互相是有联系的, 它们有可能讨论同一个话题, 因此使用 TD-IDF 加权法就会忽视掉它们之间的联系. 此外, 在评论数量非常大的情况下, 单条评论的长度可能会非常短, 因此使用 TF-IDF 加权法计算每条评论的评论向量并不是很合适.

基于以上考虑, 本文提出一种新的加权法 CF-ICF 来计算评论向量. 给出一个评论中的单词 $w$, CF 指该单词在本篇文章的所有评论集合 $C$ 中出现的次数, ICF 指在所有文章的评论集合 $C\_N$ 中该单词的反转频率. 根据该方法, 如果单词 $w$ 在集合 $C$ 中出现了很多次, 那么就说明它对当前文章很重要; 如果单词 $w$ 仅在集合 $C\_N$ 中出现了很多次, 那么就说明它对当前文章并不是很重要. 因此单词 $w$ 的特征值就依据其在 $C$ 中的词频和在 $C\_N$ 中的反转频率计算, 其定义如公式(1).

$$Value(w)=\frac{F(w)}{\sum_i F_i}\times\log\frac{|C\_N|}{|\{j:w\in C\_N_j\}|} \quad (1)$$

其中, $F(w)$代表 $w$ 在 $C$ 中出现的次数, $\sum_i F_i$ 代表所有关键字在 $C$ 中出现的总次数, $|\{j:w\in C\_N\}|$ 代表所有包含关键字 $w$ 的集合 $C$, $|C_N|$代表新闻集合 $N$ 中的所有评论数量.

### 1.3 评论聚类

用于评论聚类的算法有许多种, 如基于密度的 DBSCAN 算法[8], 基于分割的 k-means 算法[9], 基于连通性的分层算法[10]等等. 本文使用基于局部图的 BorderFlow 算法, 该算法使用两条评论间的余弦相似度作为连接边的权重. 这样属于同一类的评论会有更相近的距离和更多的联系从而完成评论的聚类. 本文中选取评论数量最多的三个聚类作为代表性评论. 在

文献[4]中证明了 BorderFlow 算法可以达到更好的聚类效果和更有效地聚类纯度.

在 BorderFlow 算法应用的过程中注意到, 有一些评论非常短, 或者不具有代表性, 这样的评论产生的聚类也会非常小, 无法代表热议的话题. 因此本文在 BorderFlow 算法的基础上忽略这种聚类.

此外, BorderFlow 算法有时会返回重叠的聚类. 为了克服这个问题, 本文使用一种超强化策略. 令 $G=(N,E,w)$代表一张图表, 其中 $N$ 代表一组评论的序号; $E$ 代表评论间的连接边; $w$ 代表连接边的权重, 即两个评论间的余弦相似度. 使用 BorderFlow 算法, 可以得到一组类 $C=\{c1,c2,...cn\}$, $\forall i \in \{1,...,k\}$, $ci$ 代表一组评论. 但是可能 $\forall i,j \in \{1,...k\}$, $c_i \cap c_j \neq \Phi$. 而在使用过超强化策略后, 就将 $k$ 个评论集合转变成 $k$ '个评论集合, 并且 $\forall i,j \in \{1,...k'\}$, $c_i' \cap c_j' = \Phi$.

### 1.4 聚类内评论排序

在完成评论的聚类后, 本文使用类PageRank算法在每个类中对评论进行排名, 并选出最有代表性的评论. PageRank 是 Google 网页搜索引擎研发的一款基于网络图的网页排名分析算法. 在本文中, 我们使用了基于评论图的类PageRank算法对评论进行排名, 以数据库中的评论作为节点, 评论间的余弦相似度作为连接边.

**获取评论图:** 评论代表了读者的观点和反馈, 而不同读者的观点有可能是相似的, 因此评论间应该存在某种联系. 考虑到同一个聚类中的评论互相联系最紧密, 它们间一定有共同的内容. 我们定义如果两个评论的余弦相似度大于一个值 *MAX*(*MAX* 是一个预设值, 如 0.5), 则两个评论间存在联系, 这样两个评论间的联系是双向的. 图 1 中给出的评论图中, 评论 1 和 2,2 和 3,1 和 4 间是存在联系的, 对于余弦相似度大于 *MAX* 的, 在评论图中设为 1, 代表两个评论间有联系; 小于 *MAX* 的设为 0, 代表两个评论间无联系.

| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | 0 | 1 | 0 | 1 |
| 2 | 1 | 0 | 1 | 0 |
| 3 | 0 | 1 | 0 | 0 |
| 4 | 1 | 0 | 0 | 0 |

图 1 评论图

PageRank 算法: 根据评论图, 我们使用 PageRank 算法对评论评分, 如公式(2)中定义.

$$Score(c_i)=\alpha\cdot\frac{1}{|c|}+(1-\alpha)\cdot\sum_{c_j}Score(c_j)\cdot\frac{1}{\sum_{c_k}e(c_j,c_k)} \quad (2)$$

$\alpha$ 代表衰减系数, 本试验中将其设为 0.15, |c|表示与一篇新闻相关的评论数量, $e(c_j,c_k)$ 代表评论图中评论 $C_j$ 到 $C_k$ 的边值.

在本系统中, 我们依据公式(2)计算出的分数来对每个聚类中的评论进行排名, 并取分数最高的几个评论.

### 1.5 聚类间评论排序

在使用了 PageRank 算法完成了每个聚类中的评论排名后, 就可以获得每个类中最具代表性的 K 个评论, 并且这些评论代表的是同一个讨论话题. 接下来, 要对每个类间的评论再次进行排序, 以获得所有话题中最具有代表性的评论, 并最终生成评论摘要.

这一步我们使用 MMR 算法, 这是一种在尽量保持查询相关性和减少冗余的前提下, 重新确定文档序值并最选取最有代表性的几条句子的方法. 本文中, MMR 算法被如下定义:

令 $Cq$ 为新闻文章的查询向量;

令 $Cs$ 为 2.4 中选出的评论集;

令 $C$ 为候选的评论向量.

某一条评论的分数由公式(3)确定.

$$Score(c)=\lambda\times sim(c,c_q)-(1-\lambda)\times\max_{c'\in C_s}(sim(c,c')) \quad (3)$$

在公式(3)中, 当 λ 设为 1 时, 计算的是 C 和 Cq 间的最大关联性; 当 λ 设为 0 时, 计算的是 C 和 Cs 间的最大差异性. 在本文中, sim(c1, c2)函数是根据每个向量集合 C 的余弦相似度计算的, 且 λ 被设为 0.8.

将 1.4 中的所有备选评论用该 MMR 算法计算评分后, 再进行排名, 就可以得到所有讨论话题中最具有代表性的评论, 并可以据此生成评论摘要.

## 2 仿真实验及结果

### 2.1 仿真实验数据库

本实验中用的数据库是从雅虎新闻网上摘取出的 1000 多条新闻及其评论集合, 其中每条新闻平均有 1059 条评论, 整个数据库约有 100 万条评论. 此外, 整个数据集合中包含有四个的文件夹, 分别为 News, newsTXT, Comments, commentsTXT, 每个文件夹中相同文件名的文件是互相关联的. 其中 News 文件夹中包含 HTML 格式的新闻文章, Comments 文件夹中包含

HTML 格式的对应文章评论, newsTXT 和 commentsTXT 文件夹中则包含新闻文章和评论的 TXT 格式文件.

### 2.2 实验结果

选择数据库中一篇名为“Magnitude 3.8 earthquake shakes Southern California, no damage reported”的文章作为测试, 该文章共有 108 条评论. 最终的聚类结果如表 1 所示, 表 2 显示的是系统自动选出的 10 条评论摘要.

表 1 评论聚类结果

| ID | 聚类结果 (评论编号) |
|---|---|
| 0 | 104, 12, 13, 2, 20, 24, 30, 31, 33, 34, 36, 38, 49, 66, 73, 82, 91 |
| 1 | 42, 45 |
| 2 | 18, 23, 28, 29, 4, 40, 43, 46, 47, 48, 51, 56, 68, 69, 71, 85, 9, 92 |
| 3 | 101, 106 |
| 4 | 5, 54, 8, 98 |
| 5 | 1, 103, 107, 108, 14, 19, 25, 26, 27, 32, 35, 39, 41, 44, 50, 52, 55, 57, 60, 61, 7, 70, 74, 84, 88, 93, 97 |

表 2 评论摘要

| No | 评论 |
|---|---|
| 1 | 3.8 earthquake in southern california? I guess you don't have much to write about.. |
| 2 | Yes this is news worthy actually..because this could the foretell signs of the long-predicted California massive earthquake...remember California and the whole United States west coast sits on the so called "pacific ring of fire" It is predicted that California earthquake will be far worse than the Indonesian or Japanese tsunami.Beafraid..be very afraid.. |
| 3 | Wish all of California would sink ninto the Ocean! The world would be better off- no more Hollyweird. |
| 4 | 3.8 is nothing in California. In fact most probably didn't even notice. Lame article. |
| 5 | Lets see you post a laugh when millions die in the biggest earthquake ever to hit the us in the next 90 days |
| 6 | Californians are too Dumb to understand any size earthquake in a forecast of something more terrible to come, just read some/most of these I Q -1 comments....Right America???? |
| 7 | California is always having tremors. They are sitting smack dab on a fault so the're always going to be shaking. Get real and produce something worthy of being called news. |
| 8 | 3.8 is news in Ohio, not California |
| 9 | Must be a slow news day. There are earthquakes in California everyday. This event is nothing special. *yawn* |
| 10 | Are you kidding me? Reporting a 3.8 earthquake? It was pathetic when the news covered 4's!<br /><br />For anyone who's curious, there are roughly 350 earthquakes of this magnitude EVERY DAY. |

从表 2 中可以看出, 选出的 10 条评论摘要基本可以代表读者不同的观点.

### 2.3 数据对比

考虑到我们所提出的方法核心是对评论聚类, 然后对聚类中选出来的评论先进行聚类内排序, 再进行聚类间排名. 这种方法对评论需要进行两次排序, 这样做是否有意义? 从理论上进行分析, 根据 BorderFlow 算法得到的评论数量最多的三个聚类中, 如果任意从中选取几条评论作为该聚类的代表进行聚类间排序, 则很有可能因为该评论在聚类内部的代表性不足而导致整个聚类在参与类间排序时无法得到准确的结果. 而使用本文中的算法. 先对每个聚类内部的评论进行排序, 得到最有代表性的几条评论, 再用这些评论进行聚类间排序, 则可以保证每个聚类的排名不会因为其内部评论的代表性而受到影响.

为了进行实验验证, 我们采取了另外一种方法, 即仅适用 MMR 算法对所有评论进行排序. 这两种方法, 我们首先选取 20 名志愿者, 让他们依据评论是否具有代表性对选出的评论摘要进行从 1 到 5 的打分, 5 分代表所有志愿者都认为该评论摘要非常有代表性, 1 分代表所有志愿者都认为该评论摘要没有代表性. 其次再根据志愿者的打分情况, 使用 NDCG 和 MAP 两种评价方法分别对两种算法生成的 4 条、6 条、8 条、10 条评论摘要进行数据有效性分析. NDCG(normalized DCG) 和 MAP(Mean Average Precision)是最常用的评价排名的指标, 在本文中这两个指标用来衡量志愿者排名和系统生成排名之间的相关度, 数值越高代表与志愿者的排名结果越接近. 实验中的数据是对数据库中任意选出的 50 条新闻的评论进行摘要采集, 并取其平均值得到. 结论如表 3.

表 3 本文算法和 MMR 算法的 NDCG、MAP 值对比

| | NDCG | MAP |
|---|---|---|
| | 4 条评论摘要 | |
| MMR 法 | 0.73 | 0.83 |
| 本文算法 | 0.75 | 0.875 |
| | 6 条评论摘要 | |
| MMR 法 | 0.66 | 0.82 |
| 本文算法 | 0.76 | 0.86 |

| | 8 条评论摘要 | |
|---|---|---|
| MMR 法 | 0.68 | 0.77 |
| 本文算法 | 0.78 | 0.83 |
| | 10 条评论摘要 | |
| MMR 法 | 0.68 | 0.71 |

根据表 3 可以看出, 本文提出的算法其 NDCG 值和 MAP 值均高于仅使用 MMR 的算法, 说明本文提出的对评论进行两次排名可以提取出符合读者直观感受的更有代表性的评论, 实验结果表明本文提出的算法有效性和准确性更高.

## 3 结论

本文提出一种面向新闻评论的摘要采集算法. 首先对所有评论进行聚类, 为此我们设计了一个新的 CF-ICF 算法来获取评论向量; 另外为了避免 BorderFlow 算法出现的类间重叠, 文中对其进行改进, 加入了一种超强化策略; 其次, 使用基于评论图的类 PageRank 算法对每个聚类中的评论进行排名, 选出最有代表性的几条评论作为每个热点话题的代表; 最后, 对上一步选出的所有评论使用 MMR 算法进行再次排序, 选择出名次最高的 K 条评论作为最终的评论摘要. 仿真实验表明, 文中提出的算法能够代表评论者不同的观点, 相比于仅使用 MMR 算法进行排序的系统, 本文提出的两次排序可以提供更好的有效性和准确性, 更符合读者直观感受.

## 参考文献

1 Hu M, Sun A, Lim EP. Comments-oriented document summarization: Understanding documents with readers' feedback. Proc. of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM. 2008. 291–298.

2 Ma Z, Sun A, Yuan Q, et al. Topic-driven reader comments summarization. Proc. of the 21st ACM International Conference on Information and Knowledge Management. ACM. 2012. 265–274.

3 Khabiri E, Caverlee J, Hsu CF. Summarizing user-contributed comments. Fifth International AAAI Conference on Weblogs and Social Media. 2011.

4 Ngomo ACN, Schumacher F. BorderFlow: A local graph clustering algorithm for natural language processing. Lecture Notes in Computer Science, 1970, 5449: 547–558.

5 Page L. The PageRank citation ranking: Bringing order to the web. Stanford InfoLab. 1999: 1–14.

6 Goldstein J, Carbonell J. Summarization: (1) using MMR for diversity - based reranking and (2) evaluating summaries. Proc. of a Workshop on Held at Baltimore. Association for Computational Linguistics. Maryland. 1998. 181–195.

7 Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Information Processing & Management an International Journal, 1988, 24(5): 513–523.

8 荣秋生,颜君彪,郭国强.基于 DBSCAN 聚类算法的研究与实现.计算机应用,2004,24(4):45–46.

9 Hartigan JA, Wong MA. Algorithm AS 136: A k-means clustering algorithm. Applied Statistics, 1979, 28(1): 100–108.

10 Johnson SC. Hierarchial clustering schemes. Psychometrika, 1967, 32(3): 241–248.