可增量的用户短文本聚类方法研究

张 仪,陈 国,张再跃
(江苏科技大学 计算机科学与工程学院,江苏 镇江 212003)

摘 要:随着大数据时代的到来,用户短文本数据呈爆炸性增长,充分利用聚类分析技术获取短文本中的有用信息显得十分重要。聚类分析作为一种重要的知识发现手段,是将对象按其特征的相似程度进行归类的过程。为此,提出了一种可增量面向用户短文本聚类方法。该方法包括离线聚类和在线聚类两大类,前者在短文本预处理的基础上,利用无关语词典对短文本中的无关语进行识别和清理,再利用词类词典对短文本进行语义归一化;同时还提出了基于多特征融合的相似度计算方法,以实现对文本的相关性聚类。后者则以离线聚类结果为特征,对在线文本进行在线聚类操作,将离线聚类结果和在线聚类结果进行合并,以生成最终的聚类结果。为验证该方法的有效性与可行性,与基于特征向量的相似度方法进行了对比实验。实验结果表明,该方法的聚类召回率可达 73%,聚类精度达到 87.7%,F 值为 79.6%,均优于基于特征向量的方法。

关键词:短文本;语义归一化;离线聚类;在线聚类
中图分类号:TP301 **文献标识码**:A **文章编号**:1673-629X(2017)11-0083-05
doi:10.3969/j.issn.1673-629X.2017.11.018

Research on Scalable Clustering of User-oriented Short Text

ZHANG Yi, CHEN Guo, ZHANG Zai-yue
(School of Computer Science and Engineering, Jiangsu University of Science and Technology, Zhenjiang 212003, China)

Abstract: With the advent of big data time, data of user short text has growing explosively. Acquisition of useful information from short text with clustering analysis technology is becoming most important. Clustering analysis, as a crucial means of knowledge discovery, is the process of classifying the objects according to their similarity degree of characteristics. Therefore, a scalable clustering method of user-oriented short text is proposed, which is composed of two phases, offline clustering and online clustering. The short text is pre-processed by recognizing and removing irrelevant words with irrelevant words dictionary and normalizing semantics with parts of speech dictionary in offline clustering. A similarity calculation method is proposed based on fusion of mutli-features to conduct correlation clustering on text. Then in the online clustering, the online texts are clustered via taken results of offline clustering as features. Results of clustering are produced by integration of the results from offline clustering with those of online clustering. In order to verify its effectiveness and feasibility, the contrast experiments are conducted. Experimental results show that it has achieved recall rate in clustering by 73%, clustering accuracy by 87.7% and value of F-measure by 79.6%, which is superior to feature vector method.

Key words: short text; semantic normalization; offline clustering; online clustering

1 概 述

随着互联网的快速发展和大范围普及,文本信息已成为一种重要的信息来源,如电子邮件、新闻、网页等,但是文本自身所具备的无序性、多样性和广泛性,使得从大量的文本中获取有用的知识成为一个难题,因此文本挖掘应运而生。文本挖掘是指从大量的文本数据中挖掘出潜在的、事先未知的、对用户有用的知识的过程[1],主要方法有:关联分析、总结、分类、聚类,其中文本聚类[2]是文本挖掘最常用的方法。文本挖掘是数据挖掘[3-4]应用的一个具体领域,二者既有联系又有区别。文本挖掘是通过处理非结构化的文档内容,发现文档数据集中潜在的关系和知识,从而为用户提供有用的信息。文本聚类的效果如何主要取决于文本表示模型和文本聚类方法。

收稿日期:2016-07-28 **修回日期**:2016-11-09 **网络出版时间**:2017-07-19
基金项目:国家自然科学基金资助项目(61371114,61170156);江苏科技大学海洋装备研究院自培育项目(HZ2016004)
作者简介:张 仪(1987-),男,硕士研究生,研究方向为自然语言理解、知识获取;张再跃,教授,博士,研究方向为知识表示与获取。

聚类分析[5]是将相似的对象放到不同的组中,使每个组中的成员对象拥有一些相似的属性或特征,这体现了“物以类聚”的自然规律。文本聚类则是根据文档内容的相似度将文档分为若干个簇,使得每个簇中的文档内容的相似度尽可能大,不同簇中的文档相似度尽可能小。

文本表示模型就是为半结构化或非结构化的文本定义一个形式化的数学模型[6],用这个模型有效地表示文本。其中最具代表性的有向量空间模型[6]、概率模型[7]、概念链模型[8]、图模型[9]等。向量空间模型(Vector Space Model,VSM)是近年来应用较多且效果较好的方法之一[9],其最大优点在于模型简单,可直接应用于聚类算法。文本聚类方法有很多,比如以贝叶斯理论为基础,用概率的方法进行聚类[10],也可以将文本表示成特征向量,用距离[11]表示文本的相似度进行聚类。

目前,国内外对文本聚类的研究主要集中在文本特征的提取、聚类算法的提出、对聚类结果的评价和聚类结果的表示。文本聚类领域的研究在国外起步较早且发展较快,现在已经有很多研究成果应用在文本挖掘、搜索系统和邮件过滤等领域。文本聚类可以自动提取多个文档主题信息,消除冗余,从而自动生成一篇简明扼要的摘要,其中哥伦比亚大学开发的多文档自动文摘系统 Newsblaster[12]做得比较出色。例如 IBM 的 Intelligent Miner for Text 允许企业从文本信息中获取有价值的客户信息,它扩展了 IBM 的数据采集功能,可以从文本文档和数据源获取信息,其功能包括识别文档语言,建立姓名、用语或其他词汇的词典,提取文本的涵义,将类似的文本分组,并根据内容将文档归类。

相较于国外,国内文本聚类研究起步则较晚些,研究主要集中在科研院所和高等院校,并且取得了不错的成绩。中国科学院计算技术研究所对文本挖掘和知识检索的研究做出了巨大贡献,提出的基于 HMM 模型[13]的中文分词算法已经应用到多个实际系统。另外,钟国祥和王刚[14]提出了利用本体描述文本,根据本体件的语义相似度衡量文本间的相似度,算法称为 TCBO(Text Clustering Based on Ontology);李晓光等[15]提出了一种基于信息论的潜在概念获取与文本聚类方法。

在问答系统等自然语言理解应用系统中,用户大量的咨询文本一般都是简单语句,或者是由简单语句构成的复合句,称为用户短文本。与传统的文本聚类相比,短文本聚类中的挑战在于:含有的特征词较少,容易造成描述概念信号弱、特征稀疏等问题[16],从而影响短文本的相似度计算;由于自然语言具有高度的灵活性,特别是用户短文本中都是口语化的表示,其中包含了很多与短文本要表达的本质含义无关的词语,这会对短文本的语义产生干扰,影响短文本的相似度计算;由于语言表达的多样性,同样的事物可以使用不同的词语表示,短文本语料中有大量意义相近,但是词性不同的词语或短语,这也会影响短文本的语义相似度计算,从而影响聚类的准确性。

为了解决以上短文本聚类中存在的问题,提出一种短文本聚类系统的框架,包括基于无关语识别和词类归一化的相似度计算方法、基于离线聚类与在线聚类的聚类方法。

2 系统框架与实现

2.1 系统框架

由于聚类的数据量大,如果直接利用聚类,算法效率很低,无法满足线上应用的需要。因此提出一种可增量的用户短文本聚类系统,该系统将聚类分为离线聚类和在线聚类两部分,系统框架如图 1 所示。

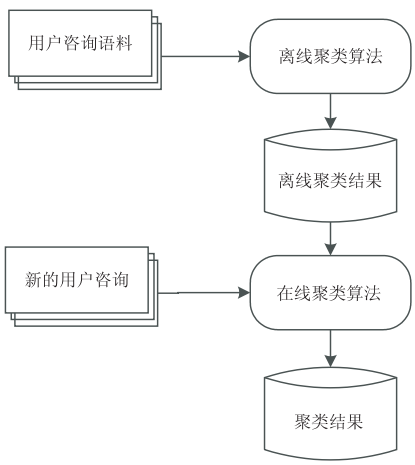

图 1 系统框架

2.2 离线聚类

短文本中含有很多与句子本身语义无关的成分,并且这些成分影响了短文本之间的句子相似度。这些与语义无关的词语,称为无关语。而相似度的准确性决定了聚类结果,因此首先对句子进行无关语识别,清除句子中的无关语。建立语义词类对短文本进行语义归一化,然后利用短文本之间的相似度,建立相似度图,最后利用聚类算法对短文本进行聚类。图 2 是离线聚类的框架图。

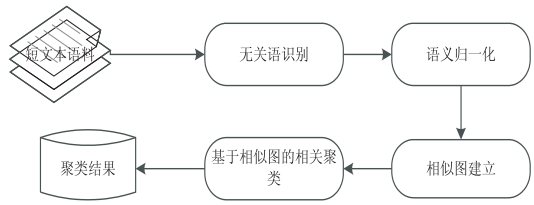

图 2 离线聚类框架

系统框架分为 4 个部分:

(1)短文本干扰项的预处理:在无关语词典的支持下,对短文本中的无关语进行识别,从而对短文本中的语义干扰项进行清理;

(2)基于词类的语义归一化:汉语中相同的语义有多种表达法,并且词的形式也不一样,因此需要对相同语义的词或短语进行归一化,从而提高句子间的语义相似度;

(3)相似度图的建立:经过预处理和语义归一化后,求得短文本之间的相似度,建立短文本的相似度图;

(4)短文本聚类算法:利用层次聚类算法对短文本进行聚类。

2.2.1　无关语识别

在短文本句中,存在大量的词语与短语,而这些词语本身对句子的语义没有实质的意义,从短文本中去掉这些词,短文本的语义不发生实质变化。例如:S_1 =“你好,请问一下电脑蓝屏如何处理?”

在 S_1 中“你好”、“请问一下”等词语对 S_1 的语义没有意义,因此在相似度计算前需要对这些词进行清除。

定义 1(无关语):不影响短文本语义的词语或句子成分。

经过长期的积累收集和获取无关语,构成无关语词典。文本预处理利用无关语词典建立双数组 Trie 结构,利用前向最大匹配进行无关语识别,然后清理短文本中的无关语。例如:S_2 =“电脑蓝屏怎么办呀?”,如果直接计算 S_1 与 S_2 的相似度,其相似度较低。但是经过无关语识别和清理后:

S_1 对应的句子为 S_1' =“电脑蓝屏如何处理”,而 S_2 清理后对应为 S_2' =“电脑蓝屏怎么办”,经过处理后,计算 S_1 与 S_2 的相似度就转换成计算 S_1' 与 S_2' 的相似度,而 S_1' 与 S_2' 的相似度更能表示这两个原短文本的相似度。

2.2.2　语义归一化

由于自然语言的随意性,并且汉语中存在大量的意义相近、但词性不同的词语或短语,完全不同的句子表达的意思却是相同的。例如:

S_2 =“我的笔记本电脑运行很慢。”

S_3 =“手提跑起程序来不动。”

在 S_2 与 S_3 中,意思是相近的,但是句子中的词完全不一致,利用传统的相似度算法计算 S_2 与 S_3 的相似度很低,因此很难将 S_2 与 S_3 聚类到一起。

将一些领域相关的表示同义的词与短语收集起来,形成词语语义类,利用词语语义类对短文本进行语义归一化。例如,在 S_2 与 S_3 中,定义如下词类:

“笔记本电脑|手提|手提电脑|笔记本|个人计算机|……”

“运行|跑程序|跑起程序|运行程序|……”

“慢|很慢|非常慢|不动|……”

利用上述词类,可以将 S_2 与 S_3 进行语义归一化,从而使 S_2 与 S_3 语义相同。

2.2.3　短文本相似度计算

(1)构造的特征向量空间为 $V=\{X_1,X_2,\cdots,X_n\}$,句子 S_1 的特征向量为 $V_1=\{\omega_1,\omega_2,\cdots,\omega_n\}$,$\omega_i$ 表示特征词 X_i 在句子 S_1 中出现的次数,句子 S_2 的特征向量为 $V_2=\{\varphi_1,\varphi_2,\cdots,\varphi_n\}$,$\varphi_i$ 是特征词 X_i 在句子 S_2 中出现的次数,则 S_1 与 S_2 间的特征向量的相似度为:

$$\mathrm{Sim}_1(S_1,S_2)=\frac{\sum_{i=1}^{n}\omega_i\varphi_i}{\sqrt{\sum_{i=1}^{n}\omega_i^2}*\sqrt{\sum_{i=1}^{n}\varphi_i^2}} \tag{1}$$

(2)计算句子间的 2-Gram 相似度,分别求出句子 S_1 和 S_2 的 2-Gram 序列。

$\mathrm{Seq}_1=\{Bw_1,w_1w_2,\cdots,w_{n-1}w_n,w_nE\}$,$\mathrm{Seq}_2=\{Bw_1',w_1'w_2',\cdots,w_{n-1}'w_n',w_n'E\}$,其中 B 和 E 是特殊的符号,分别表示句子的开始和句子的结束,则 S_1 和 S_2 间的 2-Gram 相似度为:

$$\mathrm{Sim}_2(S_1,S_2)=\frac{|\ \mathrm{Seq}_1\cap\mathrm{Seq}_2\ |}{|\ \mathrm{Seq}_1\cup\mathrm{Seq}_2\ |} \tag{2}$$

(3)计算咨询间的搭配相似度。对句子进行搭配分析,获取句子中的搭配对,其中 Col_1 为 S_1 的词的搭配的集合,Col_2 为 S_2 的词的搭配的集合,则 S_1 和 S_2 间的搭配相似度为:

$$\mathrm{Sim}_3(S_1,S_2)=\frac{|\ \mathrm{Col}_1\cap\mathrm{Col}_2\ |}{|\ \mathrm{Col}_1\cup\mathrm{Col}_2\ |} \tag{3}$$

(4)通过多特征的相似度融合算法计算咨询间的相似度:

$$\mathrm{Sim}(S_1,S_2)=w_1*\mathrm{Sim}_1(S_1,S_2)+w_2*\mathrm{Sim}_2(S_1,S_2)+w_3*\mathrm{Sim}_3(S_1,S_2) \tag{4}$$

其中,w_1,w_2,w_3 分别表示这三种相似度的权重,且满足 $w_1+w_2+w_3=1$。

2.2.4　短文本聚类算法

由于短文本相似度矩阵非常巨大,并且很多相似度值为 0(或极小),对这些零元素进行计算和存储会造成程序计算和存储空间的浪费。对短文本进行多次实验,发现短文本相似度小于某个阈值(α)的点非常多,因此采用基于相似度稀疏矩阵的短文本聚类方法,相似度低于 α 的点被排除。首先通过相似度阈值 α 筛选构造了短文本相似度稀疏矩阵,采用文献[17]中的关联聚类(Correlation Clustering)方法对短文本相似度图进行聚类。关联聚类算法是一种随机算法,主要是

基于同簇中的不相似的句子数量和不同簇中的相似句子数量的最小化代价函数。修改原始算法,对边是否剪枝或参与聚类增加权值,加入到代价函数中。算法易于初始化。利用不同的随机数进行多次多重关联聚类,当代价函数最小时就认为是最后的聚类结果。关联聚类的一个重要特征是不需要告诉聚类算法簇的数量,而短文本语料也很难估计有多少个类别。短文本聚类算法如下所述:

算法 1:Offline-clustering。

输入:短文本集合 $D = \{S_1, S_2 \cdots\}$,文本相似度矩阵 $\boldsymbol{X}$,相似度阈值 *,聚类阈值 *

输出:文本集合 D 的一个聚类 $C = \{c_1, c_2 \cdots\}$,c_i 中的元素为 D 中的元素,最终更新后的文本相似度稀疏矩阵为 $\boldsymbol{X}'$。

1:Begin

2:排除 $\boldsymbol{X}$ 中小于 * 的元素,形成短文本相似度稀疏矩阵 $\boldsymbol{X}'$;

3:在 $\boldsymbol{X}'$ 中寻找最大的且大于 * 的一对点 V_1 与 V_2,若找到执行 3,否则执行算法 2;

4:将 V_1 和 V_2 看成一个新簇,更新 $\boldsymbol{X}'$,将更新后的相似度稀疏矩阵记为 $\boldsymbol{X}''$;

5:将 V_1 和 V_2 合并为新簇 NewCluster;

6:利用以下更新 NewCluster 与其他点的相似度:

(| m_cluster[nRowIndex] | * fSimRow +

| m_cluster[nColIndex] | * fSimCol)/

(| m_cluster[nRowIndex] | +| m_cluster [nColIndex] |)

7:Repeat (1)-(5) Until 满足预先设定的终止条件;

8:End

2.3 在线聚类

在线聚类是基于离线聚类结果基础上进行的,从而可以减少聚类的时间。离线聚类后,给每个类都标记了一类号,利用离线聚类的结果作为聚类特征对用户短文本进行在线聚类,然后对离线聚类和在线聚类结果进行合并生成聚类结果。算法如下所述:

算法 2:Online-clustering。

输入:在线新的短文本集合 $D = \{S_1, S_2 \cdots\}$,文本相似度矩阵 $\boldsymbol{X}$,相似度阈值 *,聚类阈值 *;

输出:文本集合 D 的一个聚类 $C = \{c_1, c_2 \cdots\}$,c_i 中的元素为 D 中的元素,最终更新后的文本相似度稀疏矩阵为 $\boldsymbol{X}''$。

1:Begin

2:计算当前在线文本和离线聚类后的类 cluster(i)之间的相似度,通过以下方法求得:

遍历离线类 cluster(i)中的每一条咨询,利用式(4)通过多特征的相似度融合算法计算新咨询 q 和离线类咨询 q_i 间的相似度 $\mathrm{Sim}(q, q_i)$,$q_i \in \mathrm{cluster}(i)$,cluster(q)和类 cluster(i)之间的相似度为:

$$\mathrm{Sim}_2(\mathrm{cluster}(q), \mathrm{cluster}(i)) = \frac{\sum_{i=1}^{|\mathrm{cluster}(i)|} \mathrm{Sim}(q, q_i)}{|\mathrm{cluster}(i)|}$$

其中,$|\mathrm{cluster}(i)|$ 表示类 i 中的咨询的数量。

3:获得的在线用户咨询与离线聚类后的每个类的相似度,形成相似度图;

4:遍历相似度图,找到相似度最大的边,如果相似度最大的边的相似度满足阈值条件 $\mathrm{Sim}_2(\mathrm{cluster}(q), \mathrm{cluster}(i)) > \beta$,则将该咨询加入到 cluster(i)中,否则,如果没有找到满足条件的类,则将该咨询形成一个新的类 newcluster;

5:利用以下更新 NewCluster 与其他点的相似度;

6:Repeat (2)-(5) Until 满足预先设定的终止条件;

7:End

3 实验结果及分析

为了评价短文本聚类方法的效果,收集了某领域 QA 系统中的用户咨询日志,短文本句子数量 2 万行,然后人工对用户短文本进行分类,形成测试集。

聚类效果的评价指标[18]如下:

聚类召回率为:

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{5}$$

聚类准确率为:

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{6}$$

F 值为:

$$F = \frac{2PR}{P + R} \tag{7}$$

其中,TP 表示被正确聚在一起的文本;FN 表示被错误分开的文本;FP 表示被错误聚在一起的文本。

实验包括两个,一个采用基于特征词向量的相似度[4]来计算短文本间的相似度,然后利用提出的聚类算法进行聚类,另一个是利用文中提出的相似度计算方法。实验结果对比见表 1。

表 1 实验结果对比

相似度方法	R	P	F
基于特征向量的方法	0.65	0.68	0.664
文中方法	0.73	0.877 0	0.796

基于离线聚类的结果,对在线咨询和离线聚类进行在线聚类,系统能快速响应,结合快速咨询去重,大大降低了聚类的算法复杂度,且聚类结果的准确率达 85% 以上。可见,文中方法的系统响应快、精度符合实际应用需求,有效性和准确性高,具有较高的实用性,尤其适用于领域问答系统中。

从实验结果可以观察出,文中方法对 2 万行文本聚类比采用基于特征向量的方法效果要好,主要原因是短文本中的特征词较少,并且用户短文本中语义无关项比较多,从而对句子的语义产生干扰,使基于特征向量的相似度计算方法不准确。而文中方法能有效地用于用户短文本聚类,但对于包含多个分句的长文本聚类的效果不是很好,原因之一是用户文本的随意性,

多个分句中包含多个主题,从而导致聚类不准确。

4　结束语

针对短文本中含有的特征词少,容易造成描述概念信号弱、特征稀疏以及聚类耗时等问题,提出了一种可增量的用户短文本聚类方法。该方法基于将离线聚类和在线聚类相结合的聚类框架,通过离线聚类算法,利用语义无关词典和词类词典对用户咨询进行语义预处理,从而实现语义的归一化,依据基于多特征的相似度计算结果构建相似度图,根据相似度图对用户文本进行离线聚类,进而利用离线聚类的结果作为聚类特征,对在线用户文本进行在线聚类,对离线聚类和在线聚类结果进行合并,以生成聚类结果。实验结果表明,该方法在聚类召回率、精度以及 F -值方面要优于基于特征向量的方法。

参考文献:

[1] 杨占华. 聚类分析研究及其在文本挖掘中的应用[D]. 成都:西南交通大学,2014.

[2] Dhillon I S, Modha D S. Concept decompositions for large sparse text data using clustering[J]. Machine Learning, 2001, 42(1-2): 143-175.

[3] Han Jiawei, Kamber M. Data mining concepts and techniques[M]. 北京:机械工业出版社,2001.

[4] Apte C, Liu Bing, Pednault E P D, et al. Business applications of data mining[J]. Communications of ACM, 2002, 45(8): 49-53.

[5] Xu R, Wunsch D. Survey of clustering algorithms[J]. IEEE Transactions on Neural Networks, 2005, 16(3): 645-678.

[6] Salton G, Wong A, Yang C S. A vector space for automatic indexing[J]. Communications of the ACM, 1975, 18(11): 613-620.

[7] Fuhr N. Probabilistic models in information retrieval[J]. The Computer Journal, 1992, 35(3): 243-255.

[8] 宋韶旭. 基于语义关联的文本聚类方法[D]. 北京:清华大学,2006.

[9] 王永成. 中文信息处理技术及其基础[M]. 上海:上海交通大学出版社,1990.

[10] Ramoni M, Sebastiani P. Introduction to the robust Bayesian classifier[R]. [s. l.]: [s. n.], 1999.

[11] Cheeseman P, Stutz J. Bayesian Classification (AutoClass): theory and results[C]//Proceedings of advances in knowledge discovery and data mining. Menlo Park, CA, USA: American Association for Artificial Intelligence, 1996: 153-180.

[12] Hatzivassiloglou V. Simfinder: a flexible clustering tool for summarization[C]//Proceedings of NAACL workshop on automatic summarization. Pittsburgh, USA: Association for Computational Linguistics, 2001: 4-14.

[13] Rabiner L R. A tutorial on hidden Markov models and selected applications in speech recognition[J]. Proceedings of the IEEE, 1989, 77(2): 257-286.

[14] 王　刚,钟国祥. 一种基于本体相似度计算的文本聚类算法研究[J]. 计算机科学,2010,37(9):222-224.

[15] 李晓光,于　戈,王大玲,等. 基于信息论的潜在概念获取与文本聚类[J]. 软件学报,2008,19(9):2276-2284.

[16] Macqueen J. Some methods for classification and analysis of multivariate observations[C]//Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. [s. l.]: [s. n.], 1967: 281-297.

[17] Bansal N, Blum A, Chawla S. Correlation clustering[J]. Machine Learning, 2004, 56(1-3): 89-113.

[18] 曲维光,陈小荷,吉根林. 基于框架的词语搭配自动抽取方法[J]. 计算机工程,2004,30(23):22-24.