(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 107862070 B
(45)授权公告日 2021.08.10

(21)申请号 201711170964.0

(22)申请日 2017.11.22

(65)同一申请的已公布的文献号
申请公布号 CN 107862070 A

(43)申请公布日 2018.03.30

(73)专利权人 华南理工大学
地址 511458 广东省广州市南沙区环市大道南路25号华工大广州产研院

(72)发明人 陆以勤 夏儒斐 黄国洪

(74)专利代理机构 广州粤高专利商标代理有限公司 44102
代理人 何淑珍

(51)Int.Cl.
G06F 16/35(2019.01)
G06F 16/33(2019.01)
G06F 40/284(2020.01)

(56)对比文件
CN 105022840 A,2015.11.04
JP 5545876 B2,2014.07.09
CN 106919619 A,2017.07.04
CN 105095477 A,2015.11.25
Wei Du.Tracking by cluster analysis of feature points using a mixture particle filter.《IEEE Conference on Advanced Video and Signal Based Surveillance,2005》.2005,第165-170页.
秦恺.不完全语义认知过程中信息特征正确识别仿真.《计算机仿真》.2017,第34卷(第2期),第242-245.

审查员 周循

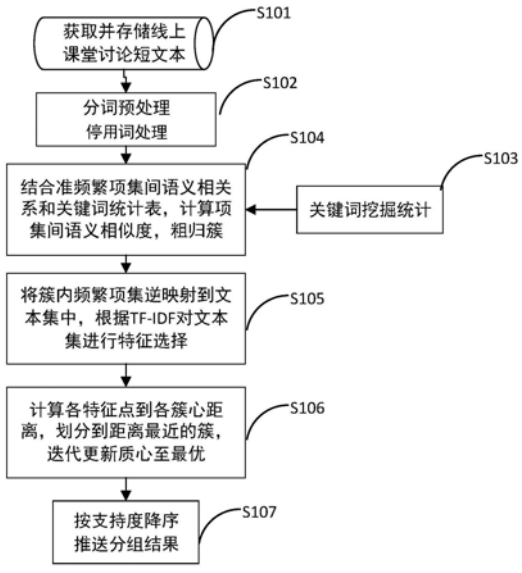
权利要求书3页 说明书6页 附图4页

(54)发明名称

基于文本聚类的线上课堂讨论短文本即时分组方法及系统

(57)摘要

本发明公开基于文本聚类的线上课堂讨论短文本即时分组方法及系统。该方法包括:对文本数据进行分词预处理和停用词预处理;获取各文本项关键词,统计存储于关键词表keyTable;对预处理后的文本集,进行频繁项集挖掘,过滤各子项准频繁项集,结合关键词表定义准频繁项集相似度计算规则,粗归簇;将各簇最靠近簇心的点逆映射到文本集,计算各簇内文本词集TF-IDF值,按距离迭代更新质心至最优;将获取的K个簇,即时分组推送。本发明采用的结合关键词表定义准频繁项集相似度计算规则有效提高线上讨论短文本聚类准确度;采用准频繁项集过滤策略有效提高归簇效率,加速聚类方法;把线上课堂讨论过的文本信息内容,自动归纳成多个主题,并把文本内容按主题分组。



CN 107862070 B

1.一种基于文本聚类的线上课堂讨论短文本即时分组方法,其特征在于,包括如下步骤的组合:

S101、获取并存储线上课堂讨论短文本数据;

S102、对文本数据,进行文本分词预处理和停用词预处理;

S103、获取各文本项关键词,存储于关键词表统计表keyTable;

S104、对预处理过后的文本集,进行频繁项集挖掘,过滤各子项的准频繁项集,结合关键词统计表定义准频繁项集相似度计算规则,粗归簇;

所述S104中结合关键词统计表定义准频繁项集相似度计算规则用于粗归簇,具体包括:

关键词统计表中各关键词K_i对语义相似度的贡献值以逆文档频数N_i来度量,i表征关键词编号,取1～n,n为文本数量;通过包含各关键词的文本个数统计,表征该关键词类别区分能力;若$N_i>n/2$,将该关键词K_i标记为基础词;否则标记为一般关键词;

在线上课堂讨论短文本中,对于同一个题目,讨论内容基于一些基础词发表不同观点;基础词和关键词运用在准频繁项集相似度计算中主要用于区分相同大前提下的小区别;词集中每个词对应几个概念,每个概念由几个义原来描述;对于两个概念s_{1i}和s_{2i},此处下标i表征各概念中义原编号,$Sim(S_{1i},S_{2i})$表示s_{1i}和s_{2i}两个概念之间的语义距离:

$$Sim(S_{1i},S_{2i})=\frac{\alpha}{\alpha+d_{\min}} \tag{1}$$

其中,d_{min}为s_{1i}、s_{2i}两概念第一义原在中文知识库层次体系中的最小距离;α取1.6;定义词语间语义相似度计算公式如下:

$$Sim(w_1,w_2)=\begin{cases}\max\limits_{i=1..n,j=i..m} Sim(S_{1i},S_{2j}) & ,\text{若 w1,w2 收录于词库中}\\ 1 & ,\text{若 w1,w2 未收录于词库中且相同}\\ 0 & ,\text{若 w1,w2 未收录于词库中且不同}\end{cases} \tag{2}$$

准频繁项集间任意两集合t_1和t_2,若t_1和t_2含k个相同的基础词:

$$Sim(t_1,t_2)=\frac{\delta\cdot\sum\max Sim(w_i,w_j)+\Delta(l-s)}{l-k} \tag{3}$$

其中,w_i,w_j不是相同的基础词,δ取1.5,用于区分相同基础关键词大前提下不同表述内容,否则:

$$Sim(t_1,t_2)=\frac{\sum\limits_{i=1\cdots n,j=1\cdots m}\max Sim(w_i,w_j)+\Delta(l-s)}{l} \tag{4}$$

其中,Δ以较小常数0.1定义任一非空值可空值相似度,l和s分别为较长和较短的两个项的长度;

S105、将各组最靠近簇心的点逆映射到文本集,计算各簇内文本词集TF-IDF值,根据TF-IDF提取文本的特征,获取文本特征向量;

S106、计算各特征点到各簇簇心距离,划分到距离最近的簇,迭代更新质心至最优;

S107、获取文本词汇特征向量的K个簇,即时分组推送各簇内容,按支持度降序排列。

2.根据权利要求1所述的一种基于文本聚类的线上课堂讨论短文本即时分组方法,其

特征在于，步骤S102及S103中文本分词预处理及关键词挖掘使用汉语词法分词系统ICTCLAS、基于HTTP协议的开源中文分词系统HTTPCWS或简易中文分词系统SCWS；S102中停用词预处理判定条件为：剔除特殊符号、中英文单字、常见噪声字词；停用词处理使用静态停用词表或基于统计学习的停用词表。

3.根据权利要求1所述的一种基于文本聚类的线上课堂讨论短文本即时分组方法，其特征在于，所述S103获取各文本项关键词存储于关键词表统计表keyTable，关键词统计表keytable记录所有关键词逆文档频数统计。

4.根据权利要求1所述的一种基于文本聚类的线上课堂讨论短文本即时分组方法，其特征在于，所述S104中频繁项集挖掘采用fp-growth算法，对得到的频繁项集过滤各子项的准频繁项集，具体包括：

扫描预处理过后的文本集获取各项集并计算其频繁度，过滤低于阈值的项，将过滤后的频繁项集写入表中按降序排列；二次扫描数据，将原始数据中的文本词项压缩到相同前缀路径共用的树中，构建fp-tree；对表中各项依次从fp-tree中获取条件模式基，累加条件模式基上该项的频繁度，过滤低于阈值的项，构建条件fp-tree；递归挖掘每个条件fp-tree，累加后缀频繁项集，直到找到fp-tree为空或fp-tree只有一条路径；

分析挖掘得到的所有频繁项集，是包含各频繁子项的所有集合的集，遍历虑除各频繁子项最大频繁项集的所有子集，得到包含各频繁子项最大频繁项集但不具有包含关系的集合作为准频繁项集。

5.根据权利要求1所述的一种基于文本聚类的线上课堂讨论短文本即时分组方法，其特征在于，所述S104中粗归簇为根据语义相似度粗归簇，步骤如下：

1)选取当下簇中最长的准频繁项作为第i个质心Ci，i表征质心编号；

2)遍历准频繁项集依次与各质心比较；

3)判断是否有交集，若有则返回2)，否则选取为下一个质心；

4)判断是否有6个质心，若有则计算各准频繁项和各质心相似度，归入相似度最大的簇直至处理完全，否则返回1)。

6.根据权利要求5所述的一种基于文本聚类的线上课堂讨论短文本即时分组方法，其特征在于，所述处理完全时簇个数等于6个。

7.根据权利要求1所述的一种基于文本聚类的线上课堂讨论短文本即时分组方法，其特征在于，所述S105中将各组最靠近簇心的点逆映射到文本集，逆映射过程基于SQL记录；所述S105中根据TF-IDF提取文本的特征，获取文本特征向量包括：计算各文本向量中特征词文件词频TF和逆文档频率IDF，设定TF-IDF阈值条件，选取满足条件的特征词做特征词；所述S106中计算各特征点到各簇簇心的距离，距离采用余弦距离；质心迭代更新基于簇内数据点距离均值。

8.根据权利要求1所述的一种基于文本聚类的线上课堂讨论短文本即时分组方法，其特征在于，所述S107中，按支持度将序排列，支持度以该簇内文本数量表征。

9.用于权利要求1～8任一项所述方法的一种基于文本聚类的线上课堂讨论短文本即时分组系统，其特征在于，通过计算机硬件及类似spark的高效大数据并行计算平台上的编程软件实现，包括如下模块：

线上课堂讨论短文本获取模块，以递增文本编号文本内容相对应的形式存储；

中文分词模块，对获得的线上课堂讨论短文本内容进行中文切词，得到线上课堂讨论短文本所有词集，然后做停用词处理；

关键词统计模块，对线上课堂讨论短文本依次获得每个编号对应文本的关键词存储于keyTable中，统计keyTable中各关键词出现频数合并统计存储；

聚类模块，挖掘线上课堂讨论短文本词集的频繁项集，过滤准频繁项集，结合keyTable计算准频繁项集相似度，粗归簇，依据频繁项集和文本间逆向关系确定簇心数据点；计算各数据点到初始聚类中心点的余弦距离，归于距离最近的簇，迭代直至最优；

即时分组模块，将按聚类结果分成的组按支持度降序依次排列；得到线上课堂讨论短文本即时分组内容推送。

基于文本聚类的线上课堂讨论短文本即时分组方法及系统

技术领域

[0001] 本发明涉及计算机技术领域，具体涉及一种基于文本聚类的线上课堂讨论短文本即时分组方法及系统。

背景技术

[0002] 集成了互联网和传统教育资源的在线云课堂平台兴起于近几年，各大高校、教育机构纷纷开设云课堂在线平台。云课堂为用户创造了一个即时的网络互动课堂，因其高效、便捷、即时性等特点而深受在线学习者欢迎。互动部分中，线上课堂讨论内容实现即时分组可使课上讨论内容条理更明确清晰，可有效提高在线学习者的阅读效率，常采用数据挖掘的方法进行操作。

[0003] 现有技术中，对无标记文本内容分组的常用方法是文本聚类，对同主题文档进行冗余消除、信息融合处理。在中文线上课堂讨论中大量存在10至50有效中文词组组成的短文本信息。现有对短文本的聚类方法主要基于传统的聚类方法，可分为层次法、划分法、基于密度的方法、基于网格的方法和基于模型的方法。在使用传统的聚类方法对短文本进行数据化时，常用的向量空间模型因具有向量维度高、特征稀疏、语义信息不丰富等特点而影响了聚类的准确度。

[0004] 在传统聚类方法中，K-means算法以其简洁、快速和较好的准确度而得以广泛运用。K-means算法是基于数据点到初始聚类中心的某种距离作为优化的目标函数，利用迭代运算调整聚类中心至目标函数最优。算法的初始中心，对聚类结果有较大的影响，但是传统的K-means算法初始中心由随机函数获得。且传统的K-means算法不可预测聚类类别数目。

发明内容

[0005] 本发明为解决上述技术问题，提出了一种基于文本聚类的线上课堂讨论短文本即时分组方法及系统。通过文本预处理、关键词挖掘、准频繁项集粗归簇结合TF-IDF计算簇间文本距离迭代更新质心，调研明确聚类个数，一定程度上克服了传统聚类算法不能准确应用于线上课堂讨论短文本的问题。

[0006] 本发明提供的基于文本聚类的线上课堂讨论短文本即时分组方法，包括：

[0007] 获取并存储线上课堂讨论短文本数据；

[0008] 对文本数据，进行分词预处理和停用词预处理；

[0009] 获取各文本项关键词，统计存储于关键词表统计表keyTable；

[0010] 对预处理过后的文本集，进行频繁项集挖掘，过滤各子项的准频繁项集，结合关键词统计表定义准频繁项集相似度计算规则，粗归簇；

[0011] 将各组最靠近簇心的点逆映射到文本集，计算各簇内文本词集TF-IDF值，根据TF-IDF提取文本的特征，获取文本特征向量；

[0012] 计算各特征点到各簇簇心距离，划分到距离最近的簇，迭代更新质心至最优。

[0013] 获取所述文本词汇特征向量的K个簇，即时分组推送各簇内容，按支持度降序排

列。

[0014] 进一步地，所述文本分词预处理及关键词挖掘使用汉语词法分词系统ICTCLAS、基于HTTP协议的开源中文分词系统HTTPCWS或简易中文分词系统SCWS；停用词预处理使用静态停用词表或基于统计学习的停用词表。其中，停用词判定条件为：剔除特殊符号、中英文单字、常见噪声字词。

[0015] 进一步地，所述获取各文本项关键词存储于关键词表统计表keyTable，关键词统计表keytable记录所有关键词逆文档频数统计。

[0016] 进一步地，所述频繁项集挖掘采用fp-growth算法。对得到的频繁项集过滤各子项的准频繁项集。包括：

[0017] 扫描预处理过后的文本集获取各项集并计算其频繁度，过滤低于阈值的项，将过滤后的频繁项集写入表中按降序排列。二次扫描数据，将原始数据中的文本词项压缩到相同前缀路径共用的树中，构建fp-tree。对表中各项依次从fp-tree中获取条件模式基，累加条件模式基上该项的频繁度，过滤低于阈值的项，构建条件fp-tree。递归挖掘每个条件fp-tree，累加后缀频繁项集，直到找到fp-tree为空或fp-tree只有一条路径。

[0018] 分析挖掘得到的所有频繁项集，是包含各频繁子项的所有集合的集，遍历滤除各频繁子项最大频繁项集的所有子集，得到包含各频繁子项最大频繁项集但不具有包含关系的集合作为准频繁项集。

[0019] 进一步地，所述结合关键词统计表定义准频繁项集相似度计算规则用于粗归簇。包括：

[0020] 对关键词统计表中各关键词K_i(i表征关键词编号，取1～n，n为文本数量)对语义相似度的贡献值以逆文档频数N_i来度量；通过包含各关键词的文本个数统计，表征该关键词类别区分能力；若N_i＞n/2，将该关键词K_i标记为基础词；否则标记为一般关键词。

[0021] 在线上课堂讨论短文本中，对于同一个题目，讨论内容大致基于一些基础词发表不同观点。基础词和关键词运用在准频繁项集相似度计算中主要用于区分相同大前提下的小区别。词集中每个词对应几个概念，每个概念由几个义原来描述。对于两个概念s_{1i}和s_{2i}(i表征各概念中义原编号)，$Sim(S_{1i},S_{2i})$表示s_{1i}和s_{2i}两个概念之间的语义距离：

[0022] $$Sim(S_{1i},S_{2i})=\frac{\alpha}{\alpha+d_{\min}} \quad (1)$$

[0023] 其中，d_{min}为s_{1i}、s_{2i}两概念第一义原在中文知识库层次体系中的最小距离。α取1.6。定义词语间语义相似度计算公式如下：

[0024] $$Sim(w_1,w_2)=\begin{cases}\max\limits_{i=1\ldots n,j=i\ldots m} Sim(S_{1i},S_{2j}) & \text{(若 w1,w2 收录于词库中)}\\ 1 & \text{(若 w1,w2 未收录于词库中且相同)}\\ 0 & \text{(若 w1,w2 未收录于词库中且不同)}\end{cases} \quad (2)$$

[0025] 准频繁项集间任意两集合t_1和t_2，若t_1和t_2含k个相同的基础词：

[0026] $$\mathrm{Sim}(t_1,t_2)=\frac{\delta\cdot\sum\max Sim(w_i,w_j)+\Delta(l-s)}{l-k} \quad (3)$$

[0027] 其中，w_i，w_j不是相同的基础词，δ取1.5，用于区分相同基础关键词大前提下不同表述内容。否则：

[0028] $$Sim(t_1,t_2)=\frac{\sum_{i=1\cdots n,j=1\cdots m}\max Sim(w_i,w_j)+\Delta(l-s)}{l} \quad (4)$$

[0029] 其中，Δ以较小常数0.1定义任一非空值可空值相似度，l和s分别为较长和较短的两个项的长度。

[0030] 进一步地，所述根据根据语义相似度粗归簇步骤如下：

[0031] 1)选取当下最长准频繁项作为第i个质心Ci(i表征质心编号)；

[0032] 2)遍历准频繁项集依次与各质心比较；

[0033] 3)判断是否有交集，若有则返回2)，否则选取为下一个质心；

[0034] 4)判断是否有6个质心，若有则计算各准频繁项和各质心相似度，归入相似度最大的簇直至处理完全，否则返回1)；

[0035] 进一步地，所述将各组最靠近簇心的点逆映射到文本集，逆映射过程基于SQL记录。

[0036] 进一步地，所述根据TF-IDF提取文本的特征，获取文本特征向量包括：计算各文本向量中特征词文件词频TF和逆文档频率IDF，设定TF-IDF阈值条件，选取满足条件的特征词做特征词。

[0037] 进一步地，所述各计算特征点到各簇心的距离，该距离采用余弦距离；质心迭代更新基于簇内数据点距离均值。

[0038] 所述按支持度将序排列，支持度以该簇内文本数量表征。

[0039] 所述文本数据包括在规定时间内提交的所有讨论内容。

[0040] 计算过程及即时推送基于类似spark的大数据并行计算平台，其在文本处理、相似度计算、聚类过程处理上的高效快速特性为即时性提供了保障。

[0041] 本发明还提供一种线上课堂讨论短文本即时分组系统，通过计算机硬件及类似spark的大数据并行计算平台上的编程软件实现，包括如下模块：

[0042] 线上课堂讨论短文本获取模块，以递增文本编号文本内容相对应的形式存储。

[0043] 中文分词模块，对获得的线上课堂讨论短文本内容进行中文切词，得到线上课堂讨论短文本所有词集，然后做停用词处理。

[0044] 关键词统计模块，对线上课堂讨论短文本依次获得每个编号对应文本的关键词存储于keyTable中。统计keyTable中各关键词出现频数合并统计存储。

[0045] 聚类模块，挖掘线上课堂讨论短文本词集的频繁项集，过滤准频繁项集，结合keyTable计算准频繁项集相似度，粗归簇，依据频繁项集和文本间逆向关系确定簇心数据点。计算各数据点到初始聚类中心点的余弦距离，归于距离最近的簇，迭代直至最优。

[0046] 即时分组模块，将按聚类结果分成的组按支持度降序依次排列。得到线上课堂讨论短文本即时分组内容推送。

[0047] 与现有技术相比，本发明的优点和有益效果在于：

[0048] (1)本发明的线上课堂讨论短文本即时分组的方法及系统从当前主流云课堂线上课堂讨论需求出发，偏向于已有教育资源中的定向问题讨论。定义了基础关键词，有效区分了具有相同基础关键词大前提下细化讨论部分内容。采用结合关键词表和准频繁项集语义距离计算语义相似度，以语义相似度作为粗归簇标准有效克服了传统短文本聚类方法中语

义信息贡献值低的问题。

[0049] (2)本发明的线上课堂讨论短文本即时分组的方法及系统利用频繁项集挖掘,过滤准频繁项集,利用语义相似度粗归簇确定了初始簇群,有效克服了传统K-means方法因初始中心随机影响聚类准确性的问题。

[0050] (3)本发明的线上课堂讨论短文本即时分组的方法及系统通过对线上课堂约1000道小学语文类问题的平均约每道题2000条讨论结果调研分析,明确聚类个数取6个最合适,增强了线上课堂讨论短文本即时分组的有效性。

[0051] (4)本发明的线上课堂讨论短文本即时分组的方法及系统使用了类似spark的大数据并行计算平台,有效提高了文本处理、相似度计算以及聚类的速度,为即时性提供了保障。

附图说明

[0052] 图1是本发明的线上课堂讨论短文本即时分组方法流程图;

[0053] 图2是本发明的线上课堂讨论短文本即时分组系统模型图;

[0054] 图3是本发明中聚类过程示意图;

[0055] 图4是本发明中聚类粗归簇流程图。

具体实施方式

[0056] 针对在线上课堂讨论短文本中使用传统聚类方法时,文本特征量稀疏同时语义贡献度低导致的短文本聚类准确度低的问题,本发明实施例提供一种线上课堂讨论短文本即时分组方法,基于频繁项集挖掘,过滤准频繁项集,利用语义相似度粗归簇确定了初始簇群,基于调研统计结果自适应确定聚类个数,基于TF-IDF计算簇内文本间距离迭代更新质心,有效提高K-means算法在短文本聚类时的准确率,使聚类结果更接近于实际需求。

[0057] 如图1所示,本发明实施例提供的一种线上课堂讨论短文本即时分组方法包括:

[0058] S101:获取并存储线上课堂讨论短文本数据。具体地,对每条发言,以递增文本编号与文本内容相对应的形式存储在sparkSQL表filesDivide中。

[0059] S102:对所有文本数据,进行分词预处理和停用词预处理。具体地,使用中科院NLPIR系统进行中文分词;使用静态停用词表进行停用词过滤。

[0060] S103:获取各文本项关键词,统计存储于关键词统计表keyTable。具体地,读取filesDivide,获取表中各文本项关键词,新建一列,存储在相应文本编号后面。统计各关键词逆文档频数存储于表keyTable中。

[0061] 如图3所示,本发明实施例提供一种聚类过程示意图;

[0062] S104:对预处理过后的文本集,进行频繁项集挖掘,过滤各子项的准频繁项集,结合关键词统计表定义准频繁项集相似度计算规则,粗归簇;

[0063] a)具体地,使用fp-growth算法挖掘文本集频繁项集。两次扫描数据库,将原始数据中的事务压缩到相同前缀路径共用的树中,构建fp-tree;递归挖掘fp-tree获取频繁项集。

[0064] b)具体地,对高度冗余的频繁项集,遍历滤除各频繁子项最大频繁项集的所有子集,得到包含各频繁子项最大频繁项集但不具有包含关系的集合作为准频繁项集。

[0065] c)具体地，以逆文档频数N_i来度量对关键词统计表中各关键词K_i（i＝1.2.....n，i表征关键词编号，n为文本数量)对语义相似度的贡献值。结合线上课堂讨论中针对围绕有指向性问题进行作答类题目的局限性，标记逆文档频数N_i＞n/2的该关键词为基础词；否则标记为一般关键词。对准频繁项集间任意两集合t_1和t_2，以如下方式计算项间相似度：

[0066] 若t_1和t_2含k个相同的基础词：

[0067] $$Sim(t_1,t_2)=\frac{\delta\cdot\sum\max Sim(w_i,w_j)+\Delta(l-s)}{l-k} \quad (3)$$

[0068] 其中，w_i，w_j不是相同的基础词(此处i表征基础词编号)，δ取1.5，用于区分相同基础关键词大前提下不同表述内容。否则：

[0069] $$Sim(t_1,t_2)=\frac{\sum_{i=1\cdots n,j=1\cdots m}\max Sim(w_i,w_j)+\Delta(l-s)}{l} \quad (4)$$

[0070] 其中，Δ以较小常数0.1定义任一非空值可空值相似度，l和s分别为较长和较短的两个项的长度。

[0071] 如图4所示，本发明实施例提供聚类粗归簇流程图；

[0072] d)具体地，根据语义相似度归簇步骤如下：

[0073] d1)选取当下最长准频繁项作为第i个质心Ci(i表征质心编号)；

[0074] d2)遍历准频繁项集依次与各质心比较；

[0075] d3)判断是否有交集，若有则返回d2)，否则选取为下一个质心；

[0076] d4)判断是否有6个质心，若有则计算各准频繁项和各质心相似度，归入相似度最大的

[0077] 簇直至处理完全，否则返回d1)

[0078] S105:将各组最靠近簇心的点逆映射到文本集，计算各簇内文本词集TF-IDF值，根据TF-IDF提取文本的特征，获取文本特征向量；

[0079] e)具体地，将各簇中准频繁项集逆映射到文本集。对于各簇选取簇中最长的准频繁项集，在包含该准频繁项集的文本中随机选取一个作为该簇质心。

[0080] f)具体地，计算各簇内文本中词集TF-IDF值，选取大于TF-IDF阈值的词做该文本中的特征词。本实施例中选择TF-IDF阈值为0.2。

[0081] S106:计算各特征点到各簇簇心距离，划分到距离最近的簇，迭代更新质心至最优。

[0082] g)具体地，根据数据点间余弦距离度量数据点间距离：

[0083] $$\cos\theta=\frac{\vec{x}\cdot\vec{y}}{\|x\|\cdot\|y\|} \quad (5)$$

[0084] 计算各簇内数据点间的余弦距离，划分到距离最近的簇。根据数据点距离均值迭代更新质心，至簇不再发生改变。

[0085] S107:获取所述文本词汇特征向量的K个簇，即时分组推送各簇内容。具体地，组间按组内容支持度降序排列；每组将最靠近该簇中心的文本放在第一位置，其他簇内文本随

机排列。

[0086] 如图2所示，本发明实施例提供的一种线上课堂讨论短文本即时分组系统，通过计算机硬件及及spark平台上的编程软件实现，包括：

[0087] 线上课堂讨论短文本获取模块201，用于获取课堂讨论短文本，文本数据包括在规定时间内提交的所有讨论内容。对每条发言，以递增文本编号与文本内容相对应的形式存储。

[0088] 中文分词模块202，用于对获取的线上课堂讨论短文本内容进行中文切词和停用词处理。得到有效短文本词集。

[0089] 关键词统计模块203，对线上课堂讨论短文本，依次获得每个编号对应文本的关键词；统计各关键词逆文档频数存储于keyTable中。

[0090] 聚类模块204，挖掘线上课堂讨论短文本词集的频繁项集，过滤准频繁项集，结合keyTable计算准频繁项集相似度，粗归簇，依据频繁项集和文本间逆向关系确定簇心数据点点。计算各数据点到初始聚类中心点的余弦距离，归于距离最近的簇，迭代直至最优。

[0091] 即时分组模块205，将按聚类结果分成的组按支持度降序依次排列。得到线上课堂讨论短文本即时分组内容推送。

[0092] 在本申请方法中涉及到的各阈值的设置均根据实验效果和经验选取。在具体实施情况中，根据文本数量、内容及文本预处理情况应对阈值做相应调节，使效果最优。

[0093] 提供以上实例仅仅为描述发明目的，而非限制本发明适用范围。凡在本发明原则范围内，所做的数量修改、等同替换等，均应包含在本发明权利要求范围之内。

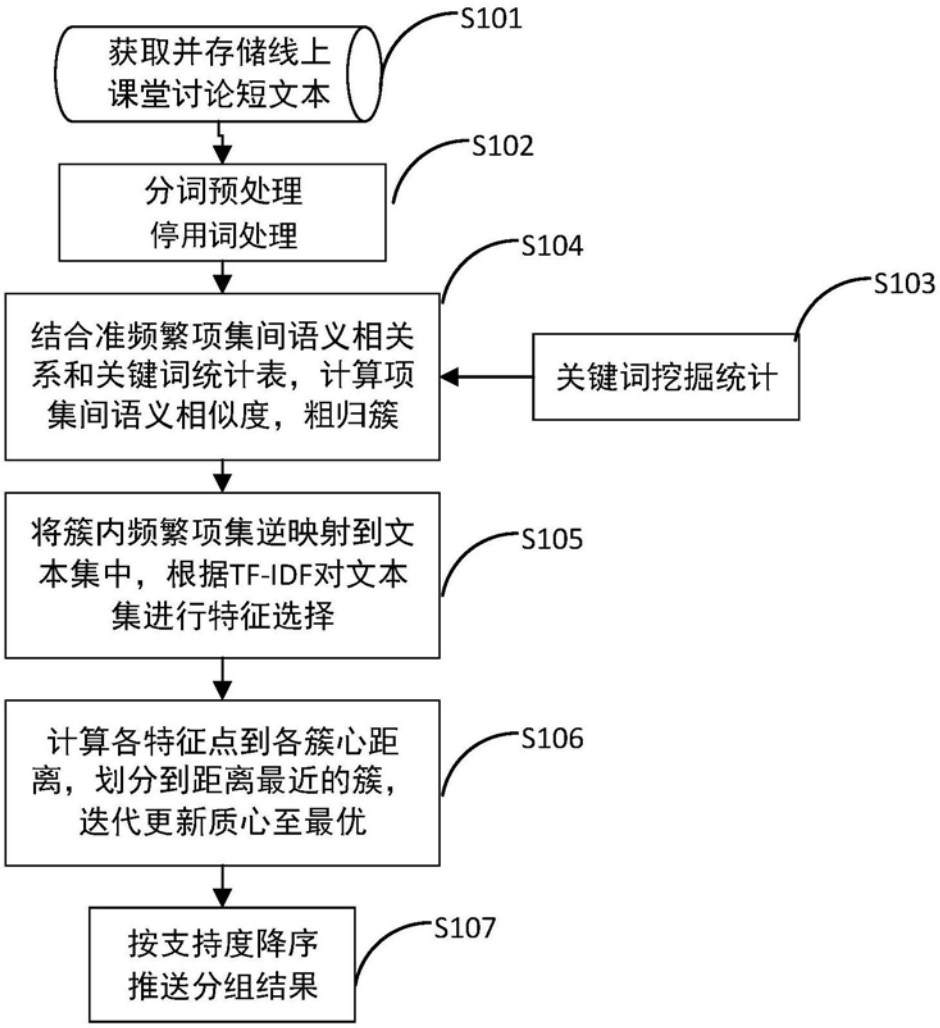

S101
获取并存储线上课堂讨论短文本
S102
分词预处理
停用词处理
S104
结合准频繁项集间语义相关系和关键词统计表，计算项集间语义相似度，粗归簇
S103
关键词挖掘统计
将簇内频繁项集逆映射到文本集中，根据TF-IDF对文本集进行特征选择
S105
计算各特征点到各簇心距离，划分到距离最近的簇，迭代更新质心至最优
S106
按支持度降序推送分组结果
S107

图1

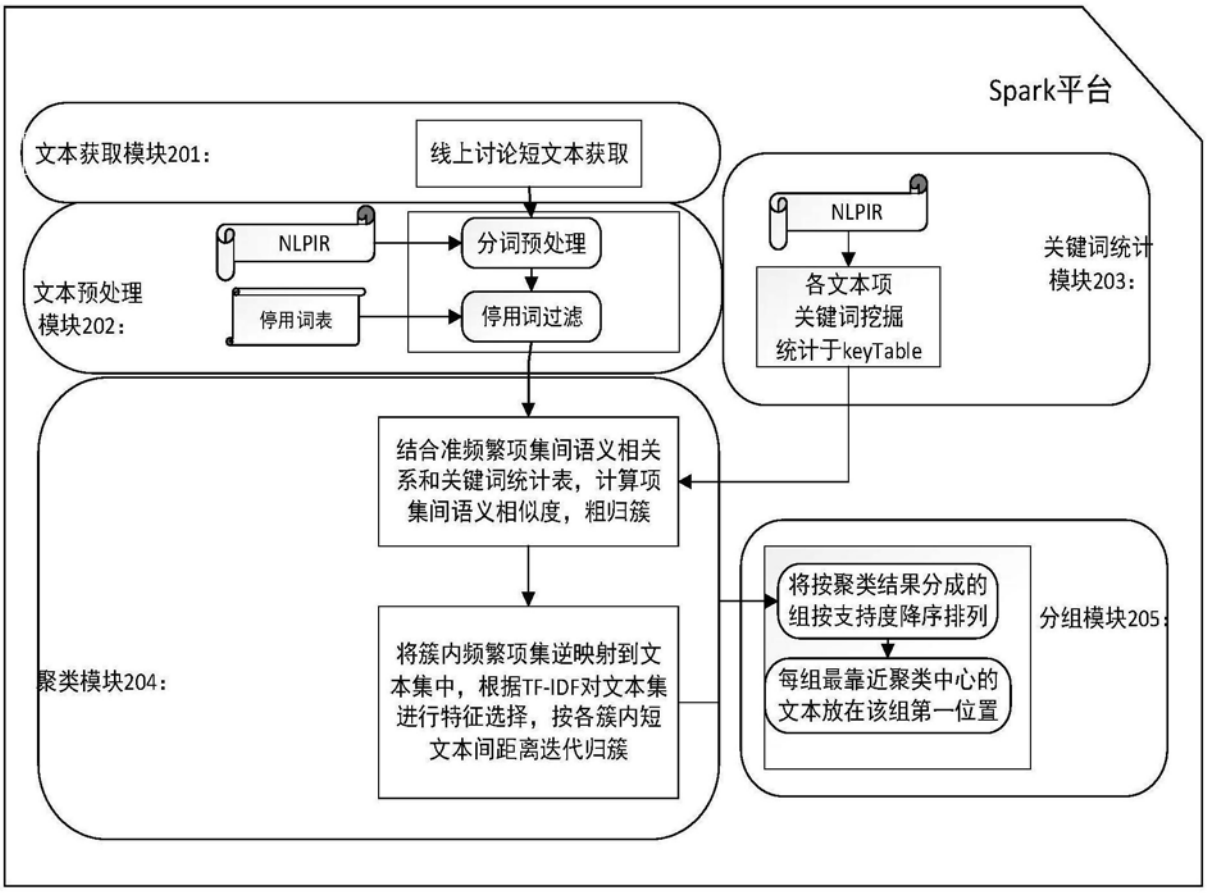

Spark平台
文本获取模块201：
线上讨论短文本获取
文本预处理
模块202：
NLPIR
分词预处理
停用词表
停用词过滤
NLPIR
各文本项
关键词挖掘
统计于keyTable
关键词统计
模块203：
结合准频繁项集间语义相关系和关键词统计表，计算项集间语义相似度，粗归簇
聚类模块204：
将簇内频繁项集逆映射到文本集中，根据TF-IDF对文本集进行特征选择，按各簇内短文本间距离迭代归簇
将按聚类结果分成的组按支持度降序排列
每组最靠近聚类中心的文本放在该组第一位置
分组模块205：

图2

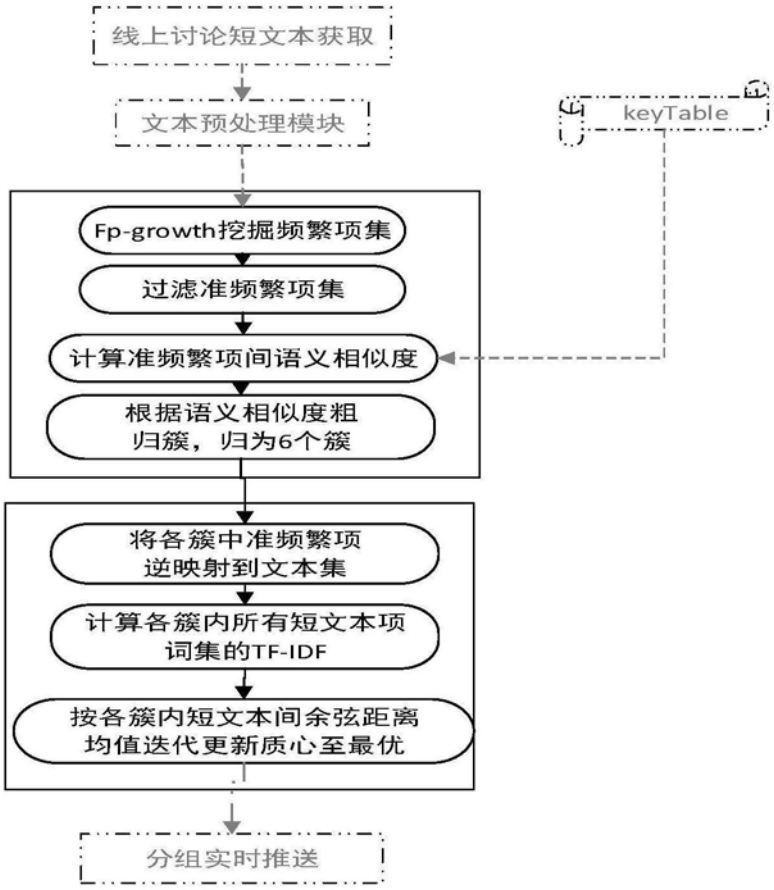

线上讨论短文本获取
文本预处理模块
keyTable
Fp-growth挖掘频繁项集
过滤准频繁项集
计算准频繁项间语义相似度
根据语义相似度粗
归簇，归为6个簇
将各簇中准频繁项
逆映射到文本集
计算各簇内所有短文本项
词集的TF-IDF
按各簇内短文本间余弦距离
均值迭代更新质心至最优
分组实时推送

图3

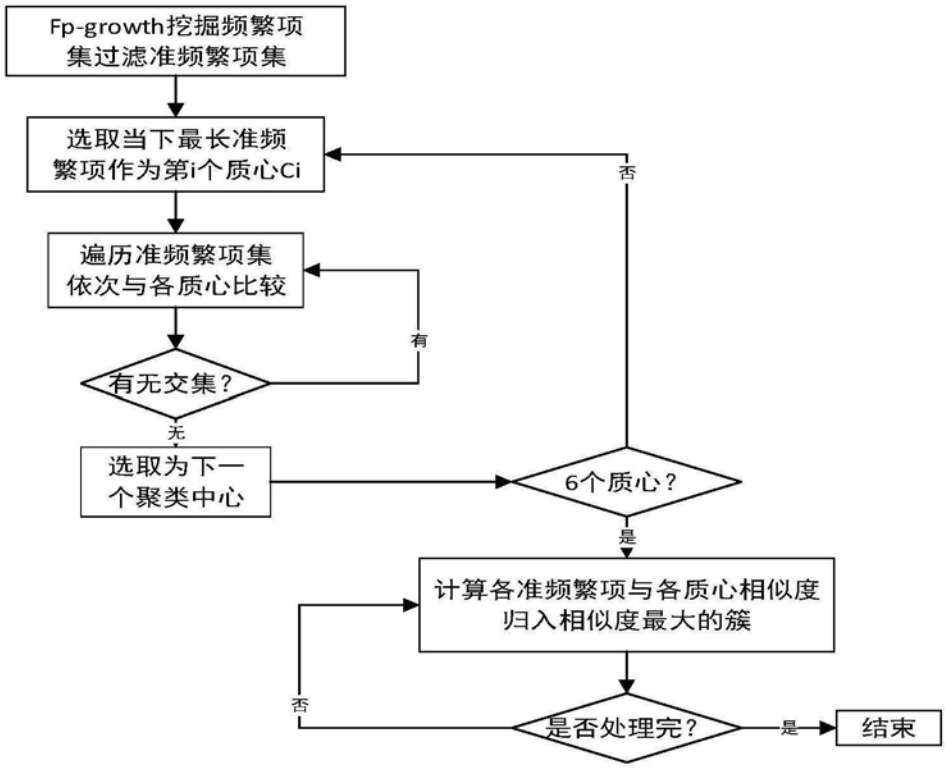

Fp-growth挖掘频繁项集过滤准频繁项集
选取当下最长准频繁项作为第i个质心Ci
遍历准频繁项集依次与各质心比较
有无交集？
有
无
选取为下一个聚类中心
6个质心？
否
是
计算各准频繁项与各质心相似度归入相似度最大的簇
是否处理完？
否
是
结束

图4