(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 103324665 B
(45)授权公告日 2017.05.03

(21)申请号 201310177797.8

(22)申请日 2013.05.14

(65)同一申请的已公布的文献号
申请公布号 CN 103324665 A

(43)申请公布日 2013.09.25

(73)专利权人 亿赞普(北京)科技有限公司
地址 100081 北京市海淀区南大街东北旺北京中关村软件园孵化器1号楼C座三层1322-D

(72)发明人 杜毅 罗峰 黄苏支 李娜

(74)专利代理机构 北京润泽恒知识产权代理有限公司 11319
代理人 赵娟

(51)Int.Cl.
G06F 17/30(2006.01)

(56)对比文件
CN 102110140 A,2011.06.29,说明书第1-5页.
马彬 等.基于线索树双层聚类的微博话题检测.《中文信息学报》.2012,第26卷(第6期),第123-127页.

审查员 高沛沛

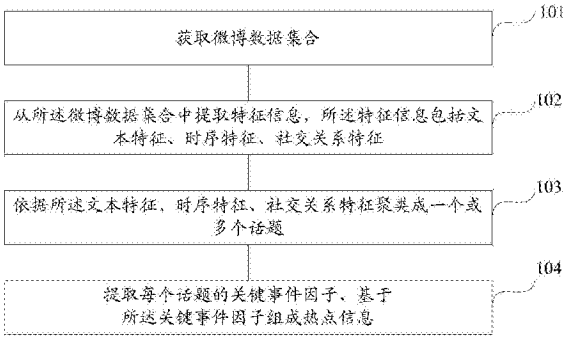
权利要求书2页 说明书10页 附图1页

(54)发明名称

一种基于微博的热点信息提取的方法和装置

(57)摘要

本发明提供了一种基于微博的热点信息提取的方法和装置,其中所述方法包括:获取微博数据集合;从所述微博数据集合中提取特征信息,所述特征信息包括文本特征、时序特征、社交关系特征;依据所述文本特征、时序特征、社交关系特征聚类成一个或多个话题;提取每个话题的关键事件因子,将基于所述关键事件因子组成热点信息。本发明综合考虑了微博数据的特点,可以提高基于微博的热点信息发现的准确度。



CN 103324665 B

1.一种基于微博的热点信息提取的方法，其特征在于，包括：

获取微博数据集合；

从所述微博数据集合中提取特征信息，所述特征信息包括文本特征、时序特征、社交关系特征；其中，所述文本特征包括微博标签、内嵌外部链接对应的标题、微博的纯文本内容；所述社交关系特征包括发布微博的用户信息、微博评论的次数、微博转发的次数；

依据所述文本特征、时序特征、社交关系特征聚类成一个或多个话题；

提取每个话题的关键事件因子，基于所述关键事件因子组成热点信息，其中，所述关键事件因子为与事件主题最相关又能达到内容覆盖最大的关键词；

所述依据文本特征、时序特征、社交关系特征聚类成一个或多个话题的步骤包括：

对所述微博标签、内嵌外部链接对应的标题、微博的纯文本内容进行预处理，构建第一空间矩阵；

依据所述时序特征、社交关系特征构建第二空间矩阵，并对所述第二空间矩阵进行降维；

按照所述第一空间矩阵与第二空间矩阵进行聚类，得到一个或多个话题。

2.根据权利要求1所述的方法，其特征在于，所述对微博标签、内嵌外部链接对应的标题、微博的纯文本内容进行预处理，构建第一空间矩阵的子步骤包括：

对微博标签、内嵌外部链接对应的标题、微博的纯文本内容进行分词；

对所述分词结果中出现的用户标签中的词汇、内嵌外部链接对应的标题中的词汇和人名、地名和机构名进行加权处理；

依据所述分词并加权的结果，构建第一空间矩阵。

3.根据权利要求1所述的方法，其特征在于，所述依据时序特征、社交关系特征构建第二空间矩阵的子步骤包括：

分别对所述时序特征以及社交关系特征添加权重；

依据所述时序特征及社交关系特征和时序特征及社交关系特征对应的权重，构建第二空间矩阵。

4.根据权利要求1-3任一权利要求所述的方法，其特征在于，所述时序特征包括微博发布时间、微博评论时间。

5.根据权利要求1-3任一权利要求所述的方法，其特征在于，所述关键事件因子包括事件最早发布时间、事件发生地名、事件发生人名、事件发生机构名、事件内容关键词、用户情感倾向性。

6.一种基于微博的热点信息提取的装置，其特征在于，包括：

微博数据结合获取模块，用于获取微博数据集合；

特征信息提取模块，用于从所述微博数据集合中提取特征信息，所述特征信息包括文本特征、时序特征、社交关系特征；其中，所述文本特征包括微博标签、内嵌外部链接对应的标题、微博的纯文本内容；所述社交关系特征包括发布微博的用户信息、微博评论的次数、微博转发的次数；

话题聚类模块，用于依据所述文本特征、时序特征、社交关系特征聚类成一个或多个话题；

热点信息组成模块，用于提取每个话题的关键事件因子，将所述关键事件因子组成热

点信息，其中，所述关键事件因子为与事件主题最相关又能达到内容覆盖最大的关键词；

所述话题聚类模块包括：

第一空间矩阵构建子模块，用于对所述微博标签、内嵌外部链接对应的标题、微博的纯文本内容进行预处理，构建第一空间矩阵；

第二空间矩阵构建子模块，用于依据所述时序特征、社交关系特征构建第二空间矩阵，并对所述第二空间矩阵进行降维；

话题生成子模块，用于按照所述第一空间矩阵与第二空间矩阵进行聚类，得到一个或多个话题。

7.根据权利要求6所述的装置，其特征在于，所述第一空间矩阵构建子模块进一步包括如下单元：

分词单元，用于对微博标签、内嵌外部链接对应的标题、微博的纯文本内容进行分词；

加权单元，用于对所述分词结果中出现的用户标签中的词汇、内嵌外部链接对应的标题中的词汇和人名、地名和机构名进行加权处理；

第一空间矩阵构造单元，用于依据所述分词并加权的结果，构建第一空间矩阵。

一种基于微博的热点信息提取的方法和装置

技术领域

[0001] 本发明涉及数据处理领域，特别是涉及一种基于微博的热点信息提取的方法，以及一种基于微博的热点信息提取的装置。

背景技术

[0002] 随着互联网的迅猛发展，如何有效利用网络舆情是一种重要的研究课题，网络舆情是由于各种事件的刺激而产生的通过互联网传播的人们对于该事件的所有认知、态度、情感和行为倾向的集合。在网络舆情的研究过程中，话题(事件)发现或检测是一项重要的技术。

[0003] 话题(事件)发现是指将输入的报道归入不同的话题簇，并在需要的时候建立新的话题簇。从本质上讲这等同于“无指导”的聚类研究，这种聚类多以增量的方式进行，聚类过程可以划分为两个阶段：识别出新事件的出现；将描写先前遇到的话题的报道归入相应的话题簇。

[0004] 目前，在话题(事件)发现方面比较有代表性的研究有：采用凝聚式聚类算法与平均聚类算法相结合的策略，将近似于同一话题模型的相关事件综合在一起作为话题检测的结果，使辅助话题检测系统具备了回溯相关事件的能力。TNO在层次话题检测方面，提出了增量式层次聚类算法，改进了凝聚层次聚类算法，其首先随机抽取小规模样本通过层次聚类构造初期的非循环有向图体系，然后将不对称的聚类结构通过二次分支进行优化，最后将其余报道根据相似性大小融合于非循环有向图体系，其中相似性大于特定阀值的报道被嵌入非循环有向图中已有的话题，而相似性小于特定阀值的报道则确定一个新的话题结构。

[0005] 微博作为新兴的一种传播形态，已经成为人们用以获取信息咨询和发布信息的主要平台之一，用户可以在微博上自由公开的对任何网络舆情热点和事件发表意见和与其他人交流。然而，上述话题检测的方法对于微博话题检测并不适用，主要存在以下缺点：

[0006] 1、数据准确率不高。传统的事件发现(检测)方法是通过构造词汇-文本特征矩阵分析事件，而微博数据的短文本性和文本缺失性会导致特征矩阵高度稀疏，从而使发现(检测)结果的准确率难以令人满意；

[0007] 2、数据检测单一性。微博数据中丰富的社交信息、超文本数据和特有的转发、评论数据为事件发现(检测)提供了更丰富的数据基础，而传统的方法并不能很好地将上述数据综合考虑进去。

[0008] 因此，本发明提出了一种基于微博的热点信息提取机制，能够综合考虑微博数据的特点，提高基于微博的热点信息发现的准确度。

发明内容

[0009] 本发明所要解决的技术问题是提供一种基于微博的热点信息提取的方法，用以综合考虑微博数据的特点，提高基于微博的热点信息发现的准确度。

[0010] 相应的，一种基于微博的热点信息提取的装置，用以保证上述方法在实际中的应用。

[0011] 为了解决上述问题，本发明公开了一种基于微博的热点信息提取的方法，包括：

[0012] 获取微博数据集合；

[0013] 从所述微博数据集合中提取特征信息，所述特征信息包括文本特征、时序特征、社交关系特征；

[0014] 依据所述文本特征、时序特征、社交关系特征聚类成一个或多个话题；

[0015] 提取每个话题的关键事件因子，基于所述关键事件因子组成热点信息。

[0016] 优选地，所述文本特征包括微博标签、内嵌外部链接对应的标题、微博的纯文本内容，所述依据文本特征、时序特征、社交关系特征聚类成一个或多个话题的步骤包括：

[0017] 对所述微博标签、内嵌外部链接对应的标题、微博的纯文本内容进行预处理，构建第一空间矩阵；

[0018] 依据所述时序特征、社交关系特征构建第二空间矩阵；

[0019] 按照所述第一空间矩阵与第二空间矩阵进行聚类，得到一个或多个话题。

[0020] 优选地，所述对微博标签、内嵌外部链接对应的标题、微博的纯文本内容进行预处理，构建第一空间矩阵的子步骤包括：

[0021] 对微博标签、内嵌外部链接对应的标题、微博的纯文本内容进行分词；

[0022] 对所述分词结果中出现的用户标签中的词汇、内嵌外部链接对应的标题中的词汇和人名、地名和机构名进行加权处理；

[0023] 依据所述分词并加权的结果，构建第一空间矩阵。

[0024] 优选地，所述依据时序特征、社交关系特征构建第二空间矩阵的子步骤包括：

[0025] 分别对所述时序特征以及社交关系特征添加权重；

[0026] 依据所述时序特征及社交关系特征和时序特征及社交关系特征对应的权重，构建第二空间矩阵。

[0027] 优选地，所述时序特征包括微博发布时间、微博评论时间。

[0028] 优选地，所述社交特征包括发布微博的用户信息，微博评论的次数、微博转发的次数。

[0029] 优选地，所述关键事件因子包括事件最早发布时间、事件发生地名、事件发生人名、事件发生机构名、事件内容关键词、用户情感倾向性。

[0030] 本发明还公开了一种基于微博的热点信息提取的装置，包括：

[0031] 微博数据结合获取模块，用于获取微博数据集合；

[0032] 特征信息提取模块，用于从所述微博数据集合中提取特征信息，所述特征信息包括文本特征、时序特征、社交关系特征；

[0033] 话题聚类模块，用于依据所述文本特征、时序特征、社交关系特征聚类成一个或多个话题；

[0034] 热点信息组成模块，用于提取每个话题的关键事件因子，将所述关键事件因子组成热点信息。

[0035] 优选地，所述文本特征包括微博标签、内嵌外部链接对应的标题、微博的纯文本内容，所述话题聚类模块包括：

[0036] 第一空间矩阵构建子模块，用于对所述微博标签、内嵌外部链接对应的标题、微博的纯文本内容进行预处理，构建第一空间矩阵；

[0037] 第二空间矩阵构建子模块，用于依据所述时序特征、社交关系特征构建第二空间矩阵；

[0038] 话题生成子模块，用于按照所述第一空间矩阵与第二空间矩阵进行聚类，得到一个或多个话题。

[0039] 优选地，所述第一空间矩阵构建子模块进一步包括如下单元：

[0040] 分词单元，用于对微博标签、内嵌外部链接对应的标题、微博的纯文本内容进行分词；

[0041] 加权单元，用于对所述分词结果中出现的用户标签中的词汇、内嵌外部链接对应的标题中的词汇和人名、地名和机构名进行加权处理；

[0042] 第一空间矩阵构造单元，用于依据所述分词并加权的结果，构建第一空间矩阵。

[0043] 与现有技术相比，本发明具有以下优点：

[0044] 首先，本发明综合考虑了微博数据的特点，在进行基于微博的话题聚类时，提取能够更全面、准确反映微博话题的文本特征、时序特征、社交关系特征，使基于微博的话题聚类更加准确、全面；

[0045] 第二，本发明能够提取与话题最相关的关键事件因子，给出更直观可读的话题热点信息。

附图说明

[0046] 图1示出了一种基于微博的热点信息提取的方法实施例的步骤流程图；

[0047] 图2示出了一种基于微博的热点信息提取的装置实施例的结构框图。

具体实施方式

[0048] 为使本发明的上述目的、特征和优点能够更加明显易懂，下面结合附图和具体实施方式对本发明作进一步详细的说明。

[0049] 参照图1，其示出了一种基于微博的热点信息提取的方法实施例的步骤流程图，具体可以包括以下步骤：

[0050] 步骤101，获取微博数据集合；

[0051] 具体而言，微博，即微博客(MicroBlog)的简称，是一个基于用户关系信息分享、传播以及获取平台，用户可以通过WEB、WAP等各种客户端组建个人社区，以140字左右的文字更新信息，并实现即时分享。微博具有以下一些特点：

[0052] (1)微博信息获取具有很强的自主性、社交选择性，用户可以根据自己的兴趣偏好，依据对方发布内容的类别与质量，来选择是否“关注”某用户，并可以对所有“关注”的用户群进行分类；

[0053] (2)微博宣传的影响力具有很大弹性，与内容质量高度相关，其影响力基于用户现有的被“关注”的数量。用户发布信息的吸引力、新闻性越强，对该用户感兴趣、关注该用户的人数也越多，影响力越大。此外，微博平台本身的认证及推荐亦助于增加被“关注”的数量；

[0054] (3)微博内容短小精悍。微博的内容限定为140字左右，内容简短，不需长篇大论，门槛较低；

[0055] (4)信息共享便捷迅速。可以通过各种连接网络的平台，在任何时间、任何地点即时发布信息，其信息发布速度超过传统纸媒及网络媒体。

[0056] 本发明实施例针对微博数据特有的特点进行话题检测，可以通过开放接口采集微博数据(也可以称为微博帖子)，生成微博数据集合。

[0057] 步骤102，从所述微博数据集合中提取特征信息，所述特征信息包括文本特征、时序特征、社交关系特征；

[0058] 在具体实现中，由于采集到的微博数据集合几乎没有经过任何处理就存入到数据库中，在原始的微博数据中存在很多没有价值的信息，如广告、重复网站导航工具或者一些半结构化的HTML代码，这些没有价值的信息很大程度上影响了话题检测的准确性，因此在进行话题检测前要对原始的微博数据进行处理，从中抽取有价值的信息。

[0059] 在本发明实施例中，基于微博数据的草根性、内容短小精悍等特点，抽取出文本特征、时序特征、社交关系特征等特征信息。

[0060] 其中，所述文本特征可以包括微博标签、内嵌外部链接对应的标题、微博的纯文本内容等内容。具体而言，在微博中，微博标签可以包括微博用户标签以及微博文章标签，微博用户标签是指用户的个性化说明，如“文艺”、“历史人文”、“摄影”等，通过用户标签可以推断用户特征；微博文章标签除了可以用于对微博文章进行分类，还可以标注微博帖子的关键内容。

[0061] 另一方面，受微博发布字数的限制(一般而言，一条微博最多允许发布140个字符)，用户只能用精简的语言发表微博，文本规范性和完整性比较差，为了更好地阐述自己的观点，用户可以在发帖时嵌入超文本链接，如:图片、视频和网页链接等，而所述超文本链接的标题在很大程度上反应了链接内容的关键主题，可以通过解析html标签或通过第三方html解析工具来抽取内嵌链接的标题。

[0062] 所述时序特征可以包括微博发布时间、微博评论时间等内容。通过微博数据集合中抽取的时序特征可以得到某个时间段内用户喜欢什么或者某个时间段内用户在做什么。

[0063] 所述社交关系特征可以包括发布微博的用户信息，微博评论的次数、微博转发的次数、微博用户关注的粉丝数等内容。通过社交关系特征可以得到微博上用户男女比例、年龄比例等。

[0064] 步骤103，依据所述文本特征、时序特征、社交关系特征聚类成一个或多个话题；

[0065] 具体而言，将物理或抽象对象的集合分成由类似的对象组成的多个类的过程被称为聚类。由聚类所生成的簇是一组数据对象的集合，这些对象与同一个簇中的对象彼此相似，与其他簇中的对象相异。可以通过聚类(Cluster)分析算法对所述文本特征、时序特征、社交关系特征进行聚类，聚类分析是由若干模式(Pattern)组成的，通常模式是一个度量(Measurement)的向量，或者是多维空间中的一个点，聚类分析以相似性为基础，在一个聚类中的模式之间比不在同一聚类中的模式之间具有更多的相似性。

[0066] 在本发明的一种优选实施例中，所述步骤103可以包括如下子步骤：

[0067] 子步骤S11，对所述用户标签、内嵌外部链接对应的标题、微博的纯文本内容进行预处理，构建第一空间矩阵；

[0068] 在具体实现中，所述第一空间矩阵可以称为词汇-文本矩阵，传统的词汇-文本矩阵构造往往考虑使用某种特征选择算法，典型的特征选择算法如采用TF-IDF算法，词频(Term Frequency，TF)表示某一词条在某一文档中出现的频率，反文档频率(Inverse Document Frequency，IDF)表示包含该词条的文档数占总文档数的比重的倒数。TF-IDF算法的基本思想是词条的重要性随着它在文件中的出现次数成正比增加，但同时会随着它在文档库中的出现频率成反比下降。然而由于微博数据的“草根性”、“随意性”等特点使得微博数据的用语灵活多变，同一个语义可能出现多种表达方式，因此这种传统的特征选择算法并不太适用于微博数据。

[0069] 针对上述问题，本发明实施例结合微博数据特有的文本特征，综合考虑内嵌链接URL、用户标签和命名实体等因素提出了相应的加权方案，对传统的TF-IDF算法进行改进，构造出更能反映微博内容的词汇-文本特征矩阵。

[0070] 在本发明的一种优选实施例中，所述子步骤S11可以包括如下子步骤：

[0071] 子步骤S111，对用户标签、内嵌外部链接对应的标题、微博的纯文本内容进行分词；

[0072] 在实际中，微博的转发评论等功能使得微博中的信息具有重复性，并且由于自然语言不仅由主要表达文本意思的名称、动词和形容词组成，还包含一些对文本表达意思价值不大的可以去除的代词、冠词、连词、介词和标点符号等。为了减少后续处理的计算量，提高算法的执行效率和话题检测的精确度，需要对微博数据的文本特征进行数据预处理，所述预处理可以包括中文分词、词性标注等。

[0073] 中文分词指的是将一个汉字序列切分成一个一个单独的词，分词就是将连续的字序列按照一定的规范重新组合成词序列的过程。中文分词是文本挖掘的基础，通过中文分词不仅可以达到电脑自动识别语句含义的效果。常用的中文分词算法可分为三大类：基于字符串匹配的分词方法、基于理解的分词方法和基于统计的分词方法；按照是否与词性标注过程相结合，又可以分为单纯分词方法和分词与标注相结合的一体化方法。本领域技术人员可以根据实际需要采用上述任一种或几种算法均是可以的，本发明实施例在此不作限制。

[0074] 应用于本发明实施例，可以将所述中文分词的结果组织成词汇集合。

[0075] 子步骤S112，对所述分词结果中出现的用户标签中的词汇、内嵌外部链接对应的标题中的词汇和人名、地名和机构名进行加权处理；

[0076] 在文本中，不同的词对文本表达主要意思的贡献不同，为了体现不同词汇在文本或者话题中的重要程度，体现不同词汇区分各个文本含义的能力，需要对文本特征中的词汇添加不同的权重。

[0077] 应用于本发明实施例，可以对以下三点内容进行加权：

[0078] 1)考虑微博标签的影响。微博帖子中的标签，很大程度上反映着该帖子的主题，在标签中出现的词汇比在微博的纯文本内容(去除微博标签tag、内嵌链接URL外的内容)中出现的词汇的权重值大；

[0079] 2)考虑内嵌外部链接URL的影响。由于微博发帖字数的限制，用户往往在帖子中嵌入外部链接，以用来阐明自己的观点，因此链接到的网页内容也能反映该帖子的主题。本发明实施例将内嵌链接对应网页的标题补充到帖子内容中，并对出现标题中的词汇进行加

权；

[0080] 3)经过分词后，对出现的词汇频率进行统计，求排序在前的N个词汇作为事件的关键词，即构成命名实体的要素。对命名实体(人名、地名、机构名)进行加权，使得聚类后的话题中尽可能出现命名实体，构成事件的两要素(地点和人物)。

[0081] 在具体实现中，通常使用TF-IDF算法对词汇进行加权，TF-IDF(term frequency-inverse document frequency)是一种用于资讯检索与资讯探勘的常用加权技术。TF-IDF是一种统计方法，用以评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度，字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。TF-IDF的主要思想是：如果某个词或短语在一篇文章中出现的频率TF高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用来分类。

[0082] 需要说明的是，在进行权重计算时本领域技术人员使用现有技术中任一一种计算方法均是可以的，本发明对此无需加以限制。

[0083] 子步骤S113，依据所述分词并加权的结果，构建第一空间矩阵。

[0084] 实际上，本发明实施例中的词汇-文本矩阵(第一空间矩阵)是一种空间向量模型，权重计算方法经常会和余弦相似性(cosine similarity)一同使用于向量空间模型中，用以判断两份文本之间的相似性。

[0085] 构建空间向量模型(Vector Space Model，简称VSM)就是通过把一个文本转化为一个空间向量实现把语言处理问题转化成易于计算的数学问题。文本特征中每个词汇对应于向量的每个维度，所述词汇集合转化的全部维度构成了整个第一空间矩阵，每个词汇对文本的代表性用每一维的权重表示。

[0086] 在具体实现中，由于所述文本特征经过分词后得到的词汇数量是巨大的，因此表示文本的空间向量的维数往往是高维的，使得在聚类时计算量巨大，并且很多情况下是稀疏矩阵，另外，虽然可以利用词汇的权值量化向量，但无法刻画文本的语义，仅仅是统计了词汇的频率而已，加之由于微博数据的“草根性”、“随意性”、“短文本性”等特点，使得其用语灵活多变，同一个语义可能出现多种表达方式，而且根据不同的语境或其他因素，原本不同的单词也有可能表示相同的意思，从而造成聚类的准确性不高。潜在语义分析(Latent Semantic Analysis，简称LSA)是处理上述问题的常用技术，其主要思想就是寻找到能够很好解决实体间词法和语义关系的数据映射，映射高维向量到潜在语义空间，使其降维。具体而言，LSA的出发点是认为所述词汇集合中的词汇与词汇之间存在某种关联，即存在某种潜在的语义结构，这种潜在的语义结构隐含在文本词汇的上下使用模式中，通过对所述词汇-文本矩阵的奇异值分解(任何矩阵都有奇异值，奇异值分解是线性代数和矩阵论中一种重要的矩阵分解法)计算，并提取K个最大奇异值以及其对应的奇异矢量构成新的词汇-文本矩阵来表示原来的词汇文本矩阵。

[0087] 子步骤S12，依据所述时序特征、社交关系特征构建第二空间矩阵；

[0088] 在本发明的一种优选实施例中，所述子步骤S12可以包括如下子步骤：

[0089] 子步骤S121，分别对所述时序特征以及社交关系特征添加权重；

[0090] 子步骤S122，依据所述时序特征及社交关系特征和时序特征及社交关系特征对应的权重，构建第二空间矩阵。

[0091] 在本发明实施例中，所述第二空间矩阵也是一种空间向量模型，所述第二空间矩阵可以为社交关系矩阵，也可以为时序矩阵、社交关系矩阵。当所述第二空间为社交关系矩阵时，所述社交关系矩阵由所述社交关系特征以及时序特征构造而成；当所述第二空间矩阵为时序矩阵以及社交关系矩阵时，所述时序矩阵由时序特征构造而成，所述社交关系矩阵由社交关系特征构造而成。

[0092] 对所述时序特征以及社交关系特征进行加权得到每个时序特征或文本特征对应的向量，每个向量对应空间向量模型的每个维度，所有的时序特征和/或社交关系特征转化成的全部维度构成整个第二空间矩阵。

[0093] 在具体实现中，可以借鉴LSA算法对所述第二空间矩阵进行降维，得到新的第二空间矩阵。

[0094] 子步骤S13，按照所述第一空间矩阵与第二空间矩阵进行聚类，得到一个或多个话题。

[0095] 具体而言，所述子步骤S13的过程是进行话题(事件)检测或发现的过程。话题(事件)发现指在将输入的报道归入不同的话题簇，并在需要的时候建立新的话题簇，从本质上讲这等同于“无指导”的，即系统无法预先知道该有多少话题簇、什么时候建立这些话题簇的聚类研究，但只允许有限的向前看。话题识别可以看作是一种按事件的聚类，这种聚类多以增量的方式进行，聚类过程可以划分为两个阶段：识别出新事件的出现；将描写先前遇到的话题的报道归入相应的话题簇。

[0096] 话题发现(检测)任务可细分为：在线话题发现、新事件发现、事件回溯发现和层次话题发现等研究子任务。在线话题发现(On-line Topic Detection，简称OTD)的主要任务是发现新话题并收集后续相关报道；新事件发现(New Event Detection，简写为NED)是辅助话题检测(TD)的重要组成部分，与首次报道检测(First Topic Detection)任务很相似，唯一的区别在于前者提交的最新事件可能相关于历史上的某一话题，后者必须输出话题最早的相关报道；事件回顾检测(Retrospective news event detection，简写为RED)的主要任务是回顾过去所有发生过的新闻报道，并从中检测出未被识别到的相关新闻事件；层次话题检测(Hierarchical Topic detection，简写为HTD)是面向话题检测中两种不恰当的假设提出的，其中一个假设是所有报道与相关话题的近似程度都在一个层次上，而另一个假设是每篇报道只可能相关于一个话题，HTD通常可以采用基于一个根节点的非循环有向图描述话题包含的层次结构。

[0097] 在本发明实施例中，在对微博数据进行事件检测时，综合考虑了微博的语义相似性、社交关系相似性和时序相似性，以提高聚类的准确性。

[0098] 在具体实现中，在构建好微博数据的第一空间矩阵以及第二空间矩阵后，通过计算微博的第一空间矩阵和/或第二空间矩阵的相似性可以得出两个微博数据结合是否相近，其中，所述第一空间矩阵的相似性可以称为语义相似性，所述第二空间矩阵的相似性可以称为社交关系相似性和/或时序相似性，把所述语义相似性、社交关系相似性和/或时序相似性都大于预设阈值的微博聚集在一起，形成数量较多凝聚度高、构成文本量有限，特征项较多的话题。

[0099] 步骤104，提取每个话题的关键事件因子，基于所述关键事件因子组成热点信息。

[0100] 应用于本发明实施例，所述步骤104的过程是事件摘要的过程，所述关键事件因子

可以称为事件元素，所述事件元素可以包括事件最早发布时间、事件发生地名、事件发生人名、事件发生机构名、事件内容关键词、用户情感倾向性等。

[0101] 具体而言，基于事件检测和聚类的结果提取既能与事件主题最相关又能达到内容覆盖最大的关键词来组成关键词(what)、命名实体(who、where)、事件的最早发帖时间(when)和用户情感倾向性(how)(4W1H)。从而得到更直观可读的事件摘要。

[0102] 其中，摘要是以提供事件内容梗概为目的，不加评论和补充解释，简明、确切地记述事件重要内容的短文。其基本要素包括事件的主要对象和范围，采用的手段和方法，得出的结果和重要的结论，有时也包括具有情报价值的其它重要的信息。

[0103] 进一步地，可以获取排序在前的N个事件摘要作为组成热点信息。例如，“动车脱轨”、“生命奇迹-伊伊”等都可以作为热点信息。

[0104] 本发明实施例综合考虑到微博数据的特征信息(包括转发次数、评论次数、内嵌外部链接、用户标注标签等)，借鉴LSA算法来计算与所述微博帖子间的时序特征和微博用户构成的社交关系特征对应的时序相似性和社交关系相似性，提出了基于微博数据语义相似性、时序相似性和社交关系相似性的事件发现算法。在进行事件摘要过程中，通过提取既能与该事件主题最相关，又能达到内容覆盖最大的关键词(what)、命名实体(who、where)、事件的最早发帖时间(when)和用户情感倾向性分析(how)总结出事件的4W1H要素，从而得到精确的热点信息。

[0105] 需要说明的是，对于方法实施例，为了简单描述，故将其都表述为一系列的动作组合，但是本领域技术人员应该知悉，本发明并不受所描述的动作顺序的限制，因为依据本发明，某些步骤可以采用其他顺序或者同时进行。其次，本领域技术人员也应该知悉，说明书中所描述的实施例均属于优选实施例，所涉及的动作并不一定是本发明所必须的。

[0106] 参照图2，其示出了一种基于微博的热点信息提取的装置实施例的结构框图，具体可以包括以下模块：

[0107] 微博数据结合获取模块201，用于获取微博数据集合；

[0108] 特征信息提取模块202，用于从所述微博数据集合中提取特征信息，所述特征信息包括文本特征、时序特征、社交关系特征；

[0109] 话题聚类模块203，用于依据所述文本特征、时序特征、社交关系特征聚类成一个或多个话题；

[0110] 其中，所述文本特征可以包括微博标签、内嵌外部链接对应的标题、微博的纯文本内容等；所述时序特征可以包括微博发布时间、微博评论时间等；所述社交特征可以包括发布微博的用户信息，微博评论的次数、微博转发的次数等。

[0111] 在本发明的一种优选实施例中，所述话题聚类模块203可以包括如下子模块：

[0112] 第一空间矩阵构建子模块，用于对所述微博标签、内嵌外部链接对应的标题、微博的纯文本内容进行预处理，构建第一空间矩阵；

[0113] 在本发明的一种优选实施例中，所述第一空间矩阵构建子模块进一步可以包括如下单元：

[0114] 分词单元，用于对微博标签、内嵌外部链接对应的标题、微博的纯文本内容进行分词；

[0115] 加权单元，用于对所述分词结果中出现的用户标签中的词汇、内嵌外部链接对应

的标题中的词汇和人名、地名和机构名进行加权处理；

[0116] 第一空间矩阵构造单元，用于依据所述分词并加权的结果，构建第一空间矩阵。

[0117] 第二空间矩阵构建子模块，用于依据所述时序特征、社交关系特征构建第二空间矩阵；

[0118] 在本发明的一种优选实施例中，所述第二空间矩阵构建子模块进一步可以包括如下单元：

[0119] 权重添加单元，用于分别对所述时序特征以及社交关系特征添加权重；

[0120] 第二空间矩阵构造单元，用于依据所述时序特征及社交关系特征和时序特征及社交关系特征对应的权重，构建第二空间矩阵。

[0121] 话题生成子模块，用于按照所述第一空间矩阵与第二空间矩阵进行聚类，得到一个或多个话题。

[0122] 热点信息组成模块204，用于提取每个话题的关键事件因子，将所述关键事件因子组成热点信息。

[0123] 作为本实施例的一种优选示例，所述关键事件因子可以包括事件最早发布时间、事件发生地名、事件发生人名、事件发生机构名、事件内容关键词、用户情感倾向性等。

[0124] 由于所述图2的装置实施例基本相应于前述图1方法实施例，故本实施例的描述中未详尽之处，可以参见前述图2实施例中的相关说明，在此就不赘述。

[0125] 本说明书中的各个实施例均采用递进的方式描述，每个实施例重点说明的都是与其他实施例的不同之处，各个实施例之间相同相似的部分互相参见即可。对于装置实施例而言，由于其与方法实施例基本相似，所以描述的比较简单，相关之处参见方法实施例的部分说明即可。

[0126] 本领域内的技术人员应明白，本发明的实施例可提供为方法、系统、或计算机程序产品。因此，本发明可采用完全硬件实施例、完全软件实施例、或结合软件和硬件方面的实施例的形式。而且，本发明可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质(包括但不限于磁盘存储器、CD-ROM、光学存储器等)上实施的计算机程序产品的形式。

[0127] 本发明是参照根据本发明实施例的方法、设备(系统)、和计算机程序产品的流程图和/或方框图来描述的。应理解可由计算机程序指令实现流程图和/或方框图中的每一流程和/或方框、以及流程图和/或方框图中的流程和/或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理设备的处理器以产生一个机器，使得通过计算机或其他可编程数据处理设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的装置。

[0128] 这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理设备以特定方式工作的计算机可读存储器中，使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品，该指令装置实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能。

[0129] 这些计算机程序指令也可装载到计算机或其他可编程数据处理设备上，使得在计算机或其他可编程设备上执行一系列操作步骤以产生计算机实现的处理，从而在计算机或其他可编程设备上执行的指令提供用于实现在流程图一个流程或多个流程和/或方框图一

个方框或多个方框中指定的功能的步骤。

[0130] 尽管已描述了本发明的优选实施例，但本领域内的技术人员一旦得知了基本创造性概念，则可对这些实施例做出另外的变更和修改。所以，所附权利要求意欲解释为包括优选实施例以及落入本发明范围的所有变更和修改。

[0131] 最后，还需要说明的是，在本文中，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、物品或者设备不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、物品或者设备所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括所述要素的过程、方法、物品或者设备中还存在另外的相同要素。

[0132] 以上对本发明所提供的一种基于微博的热点信息提取的方法和装置进行了详细介绍，本文中应用了具体个例对本发明的原理及实施方式进行了阐述，以上实施例的说明只是用于帮助理解本发明的方法及其核心思想；同时，对于本领域的一般技术人员，依据本发明的思想，在具体实施方式及应用范围上均会有改变之处，综上所述，本说明书内容不应理解为对本发明的限制。

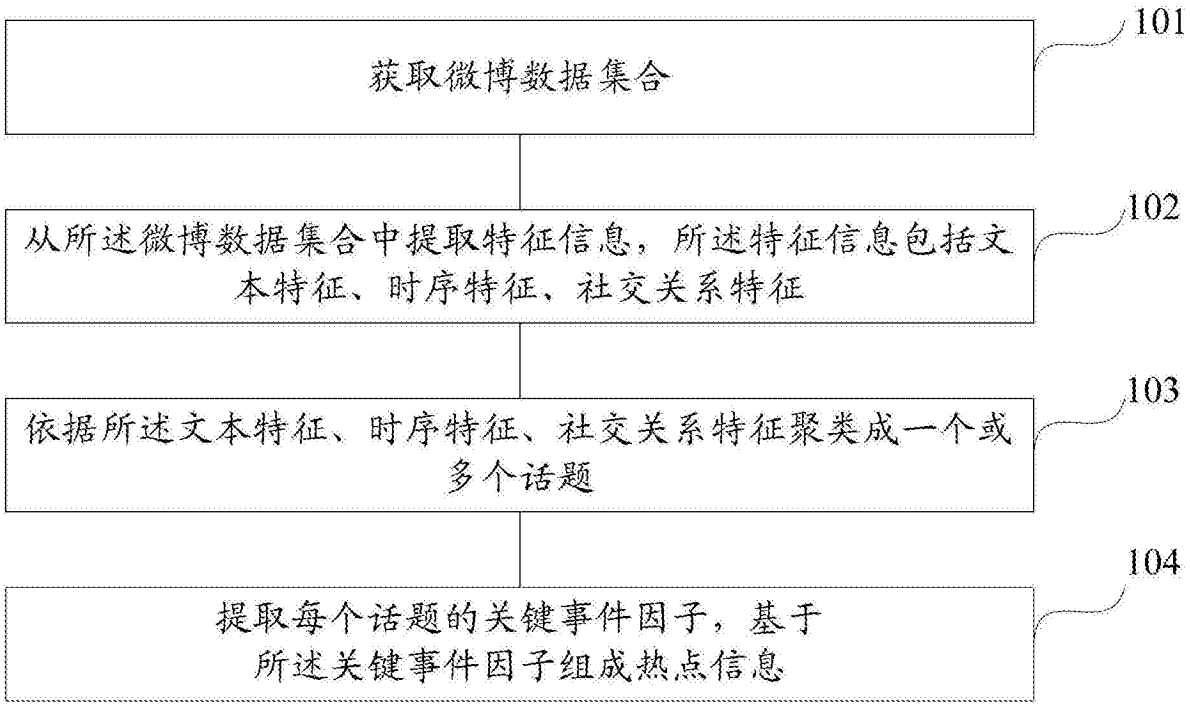

101
获取微博数据集合
102
从所述微博数据集合中提取特征信息，所述特征信息包括文本特征、时序特征、社交关系特征
103
依据所述文本特征、时序特征、社交关系特征聚类成一个或多个话题
104
提取每个话题的关键事件因子，基于所述关键事件因子组成热点信息

图1

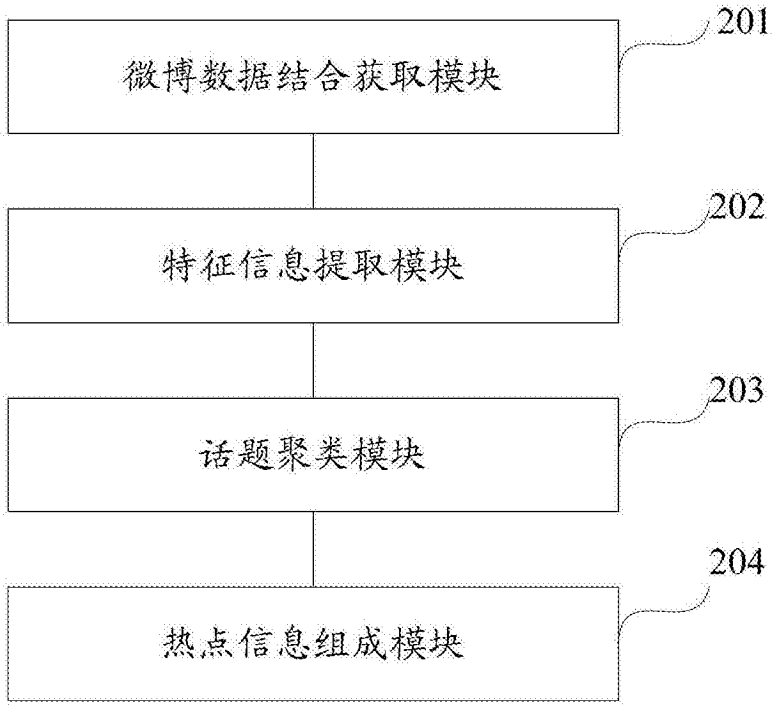

201
微博数据结合获取模块
202
特征信息提取模块
203
话题聚类模块
204
热点信息组成模块

图2