(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 105447119 A
(43)申请公布日 2016.03.30

(21)申请号 201510783415.5

(22)申请日 2015.11.16

(71)申请人 北京京东尚科信息技术有限公司
地址 100080 北京市海淀区杏石口路65号西杉创意园西区11C楼东段1-4层西段1-4层
申请人 北京京东世纪贸易有限公司

(72)发明人 黄菲菲

(74)专利代理机构 中科专利商标代理有限责任公司 11021
代理人 宋焰琴

(51)Int.Cl.
G06F 17/30(2006.01)
G06F 17/27(2006.01)

权利要求书1页 说明书5页 附图1页

(54)发明名称

一种文本聚类方法

(57)摘要

本发明公开了一种文本聚类方法，包括：在出现频繁的词汇中找出成对约束实例，出现频繁的词汇集可以从每篇文档中权重最大的那个特征词抽取出来，从中找出正约束集和负约束集；根据K最近邻集来对约束集进行扩充；根据约束集的划分结果进行聚类。本发明的方法加入了半监督聚类算法对特征词进行聚类，不仅降低了向量空间的维度，提高了实验的效率，而且在少量监督信息的指导下，使特征词的聚类更加合理、可靠；此外，本发明使层次协同聚类对文本和特征词进行聚类，聚类的效果得到改善。

CN 105447119 A

1.一种文本聚类方法，包括以下步骤：

对文本集先采用半监督聚类方法对特征词进行聚类，在出现频繁的词汇中找出成对约束实例，出现频繁的词汇集从每篇文档中权重最大的那个特征词抽取出来，从中找出正约束集和负约束集；

根据K最近邻集对所述正约束集和负约束集进行扩充；

根据约束集的划分结果对特征词进行聚类。

2.如权利要求1所述的文本聚类方法，其中所述根据K最近邻集对所述正约束集和负约束集进行扩充的原则为：

距离正约束对其中一个对象的距离非常近，距离另外一个对象的距离小于此正约束对的距离，并且此对象和正约束对的两个对象都不属于负约束集，那么此对象和其中一个对象的关系属于正约束关系；以及

距离负约束对其中一个对象的距离非常近，距离另外一个对象的距离大于此负约束对的距离，并且此对象和负约束对的两个对象都不属于正约束集，则此对象和其中一个对象的关系属于负约束关系。

3.如权利要求1所述的文本聚类方法，其中所述根据约束集的划分结果进行聚类的步骤之后，还包括将每类的特征词合并为一个属性的步骤，其中合并的方法采用数值求和方式。

4.如权利要求3所述的文本聚类方法，其中每类的特征词合并为一个属性的步骤，其中所述数值求和方式的计算公式为：

$$cw_{ip} = \frac{\sum_{t_j \in A_p} w_{ij}}{\max_{g<m \& h<k} cw_{gh}} \qquad 1 \le p \le k$$

；

其中，cw_{ip}为合并后并归一化的属性值，也就是第p个特征词簇在第i个文本中的权重之和的值。

5.如权利要求1所述的文本聚类方法，其中对特征词进行聚类后，继续采用层次协同的聚类方法对文本进行聚类。

6.如权利要求5所述的文本聚类方法，其中所述采用层次协同的聚类方法为K-means聚类方法。

7.如权利要求5所述的文本聚类方法，其中所述采用层次协同的聚类方法对文本进行聚类的步骤包括：

把所述文本的向量空间映射到每一维子空间中，每一维实际就是这个矩阵的一个列，然后在这每一维的子空间中用K-means的方法对所述文本进行聚类，其中平均权重最高的那一个簇为高权重文本簇，这样样本空间有多少维就有多少个相对应的聚类结果，就有若干个相对应的高权重文本簇，对高权重文本簇进行比较，交集的个数越高，两个相对应的特征词之间的相似度就越高，然后利用这些特征词之间的相似度用K-means算法对特征词进行聚类。

一种文本聚类方法

技术领域

[0001] 本发明涉及语义分析技术领域，更具体地涉及一种文本聚类方法。

背景技术

[0002] 信息化时代的今天，网络文本呈现出海量的特性，从搜索到的海量文本中提取有效信息或获取当前热点信息，需要对文本聚类，使同一个文本簇中的文本间相似度尽量的高，不同簇中的文本间相似度尽量的低。

[0003] 文本聚类中，常用特征词来表达文本的特性，最常用的模型是向量空间模型。向量空间模型中，每一个文本用一个向量表示，向量中的每一个值表示每一个特征词在文本中的权重。文本向量空间模型是一个矩阵模型，矩阵的行表示文本，矩阵的列表示特征词属性，矩阵中的数值表示对应列的特征词在对应行的文本中的权值。

[0004] 向量空间模型中的权值是指特征词能代表文本特性的能力，是一个文本不同于其他文本的特性。如果特征词在这个文本中出现的次数越高，在其他文本中出现的次数越低，那么这个特征词在这个文本中的权重就越高，这个特征词就越能代表这个文本的特性。

[0005] 层次聚类算法是通过分解数据集合来构建树形层次结构，具体可以分为分裂(自顶向下)算法和凝聚(自底向上)算法。分裂算法是将所有的数据集合看做一个簇，一步一步的将簇分解，然后逐层向下，每个层次都将分裂其中一个簇，直到每一个数据对象都是单独一个簇或者满足条件为止。相反，凝聚算法是起初将每个数据对象看做一个单独的簇，逐步的合并簇，从底逐层向上，每一步都合并最相似的两个簇，最终将所有的簇合并为一个簇或者满足条件为止。

[0006] 文本聚类中常用的还有协同聚类算法，其是将文本和特征词同时进行聚类的方法，该算法中将行对象和列对象同时聚类或者交替聚类，常常运用到文本聚类算法中，文本聚类中的行对象是文本对象，列对象是特征词对象，用此方法聚类不仅可以提高聚类的精确度，而且由于聚类结果的簇中含有文本对象和特征词对象，特征词对象还可以作为文本簇的主题词，从而达到了主题发现的效果。

[0007] 目前，将层次聚类和协同聚类结合起来研究实际问题，可以达到一个好的聚类效果，2010年，Li等人提出了层次协同聚类的算法，这个算法用来解决文本和特征词的共同聚类问题。此方法是将文本和特征词当作叶子节点，利用特征词和文本间的相似性矩阵，用层次聚类法同时聚类文本和特征词。2011年Li等人又提出了用层次协同聚类的算法来对音乐信息进行协同聚类，利用艺术家和音乐风格之间的权重信息，对艺术家和音乐风格进行层次协同聚类，用到了层次聚类算法的凝集算法和分裂算法两种来对实际问题进行聚类。

[0008] 由于文本和特征词都是海量的，直接运用层次协同聚类将会增加时间复杂度并且降低精度。而且只考虑了词对于文本的权重值，而忽略了特征词之间的语义关系。例如，假设一篇文档中出现“高血糖”，而另外一篇文档中出现“高血脂”，如果单单从特征词的共现性来判断两篇文档的相似度，这两个特征词被认为是完全独立的，这两篇文档可能不会属于同一个类，实际上都是属于医学类别，这里就容易出现聚类的错误。

发明内容

[0009] 有鉴于此，本发明的目的在于提出一种文本聚类方法。本发明优选的文本聚类方法可以兼采几种聚类方法的优点，改善聚类效率和精度，同时利用少量的监督信息指导聚类，使聚类的效果明显得到改善。

[0010] 为了实现上述目的，本发明提出了一种文本聚类方法，包括以下步骤：

[0011] 对文本集先采用半监督聚类方法对特征词进行聚类，在出现频繁的词汇中找出成对约束实例，出现频繁的词汇集从每篇文档中权重最大的那个特征词抽取出来，从中找出正约束集和负约束集；

[0012] 根据K最近邻集对所述正约束集和负约束集进行扩充；

[0013] 根据约束集的划分结果对特征词进行聚类。

[0014] 其中，所述根据K最近邻集对所述正约束集和负约束集进行扩充的原则为：

[0015] 距离正约束对其中一个对象的距离非常的近，距离另外一个对象的距离小于此正约束对的距离，并且此对象和正约束对的两个对象都不属于负约束集，那么此对象和其中一个对象的关系属于正约束关系；以及

[0016] 距离负约束对其中一个对象的距离非常的近，距离另外一个对象的距离大于此负约束对的距离，并且此对象和负约束对的两个对象都不属于正约束集，则此对象和其中一个对象的关系属于负约束关系。

[0017] 基于上述技术方案可知，本发明的方法加入了半监督聚类算法对特征词进行聚类，找到特征词的簇，对特征词进行合并，不仅降低了向量空间的维度，提高了实验的效率，而且在少量监督信息的指导下，使特征词的聚类更加合理、可靠；此外，本发明对文本和特征词进行层次协同聚类之前，扩充文本特征词对象集的相似度矩阵，通过找出文本之间的语义关系，特征词之间的语义关系，构造了协同矩阵，对原有的只含有特征词和文本之间的相似度的矩阵，扩充为含有文本之间的相似度、特征词之间的相似度、含有两种对象之间的相似度的协同矩阵，使层次协同聚类所含有的两种类型的对象任意两两之间的相似度包含在协同矩阵里面，聚类的效果得到改善。

附图说明

[0018] 图1为K最邻近集扩充正约束集的示意图；

[0019] 图2为K最邻近集扩充负约束集的示意图。

具体实施方式

[0020] 为使本发明的目的、技术方案和优点更加清楚明白，以下结合具体实施例，并参照附图，对本发明作进一步的详细说明。

[0021] 聚类分析(Clustering Analysis)是根据事物之间的内在联系对其进行归类，分成逐个事物的集合，又称簇(Cluster)，聚类的结果使同一个簇中的事物之间尽量相似，不同簇的对象之间尽量相异。常用的聚类分析算法包括层次聚类、协同聚类、半监督聚类等，下面分述之。

[0022] 层次聚类算法是通过分解数据集合来构建树形层次结构，具体可以分为分裂(自

顶向下)算法和凝聚(自底向上)算法。分裂算法是将所有的数据集合看做一个簇,一步一步的将簇分解,然后逐层向下,每个层次都将分裂其中一个簇,直到每一个数据对象都是单独一个簇或者满足条件为止。相反,凝聚算法是起初将每个数据对象看做一个单独的簇,逐步的合并簇,从底逐层向上,每一步都合并最相似的两个簇,最终将所有的簇合并为一个簇或者满足条件为止。

[0023] 协同聚类,又称双向聚类,或联合聚类,指对数据集的对象和属性同时聚类或者交替进行聚类,相互协作,最终达到收敛。数据集的对象和属性常常用数据矩阵的方法表示,协同聚类就是对数据矩阵的行和列同时或者交替进行聚类,实现行聚类和列聚类的彼此约束。协同聚类和传统的聚类方法最大的不同是聚类的对象含有两种类型的数据,一种是样本点,一种是样本的属性。

[0024] 半监督聚类是指利用少量的监督信息来指导聚类分析,从而提高聚类的性能。少量监督信息是指样本的类标记或者样本点之间的相似约束信息。利用样本的类标记进行聚类的算法一般是学习少量的监督信息,从中得出聚类的种子,指导聚类的完成。

[0025] 向量空间模型中的权重是指特征词能代表文本特性的能力,是一个文本不同于其他文本的特性。如果特征词在这个文本中出现的次数越高,在其他文本中出现的次数越低,那么这个特征词在这个文本中的权重就越高,这个特征词就越能代表这个文本的特性。

[0026] 半监督聚类特征词即用先验信息指导聚类那些关系极为密切的特征词,先验信息包括约束实例和类别标记,这里用到的先验信息为成对约束实例。

[0027] 本发明公开了一种半监督层次协同文本聚类方法,包括:首先在出现频繁的词汇集中找出成对约束实例,这个过程可以是人工找出成对约束集,出现频繁的词汇集可以从每篇文档中权重最大的那个特征词抽取出来,从中找出正约束集和负约束集。然后根据K最近邻集来对约束集进行扩充,接着根据约束集的划分结果进行聚类。

[0028] 成对约束实例也就是样本间的关联约束,分为正约束(Must-link)和负约束(Cannot-link)。

[0029] 设正约束集为M,负约束集为C,M和C具有对称性和传递性,如下公式(1)、公式(2),利用这两个属性可以对M集合和C集合进行扩充。

[0030] $$\begin{cases}(x_i, x_j) \in M \Leftrightarrow (x_j, x_i) \in M \\ (x_i, x_j) \in C \Leftrightarrow (x_j, x_i) \in C\end{cases} \quad (1)$$

[0031] $$\begin{cases}(x_i, x_j) \in M \,\&\, (x_j, x_k) \in M \Leftrightarrow (x_i, x_k) \in M \\ (x_i, x_j) \in M \,\&\, (x_j, x_k) \in C \Leftrightarrow (x_i, x_k) \in C\end{cases} \quad (2)$$

[0032] 对于以上扩充后的少量信息仍然无法满足需要,利用最近K临近集原则对M集合和C集合再进行扩充。K临近集法扩充约束集有如下假设:距离正约束对最近的那两个对象,如果不属于负约束集,则它们在距离上相互靠近,属于正约束集;距离负约束集最近的两个对象,如果不属于正约束对,则在距离上它们彼此疏远,属于负约束集。K临近集法扩充约束集的优点是以最小的开销,引入数据点的空间分布信息,从而构建合理庞大的约束集。

[0033] 扩充的集合M原则是距离正约束对其中一个对象的距离非常的近,距离另外一个对象的距离小于此正约束对的距离,并且此对象和正约束对的两个对象都不属于负约束

集，那么此对象和其中一个对象的关系属于正约束关系。

[0034] 如图1所示，$(x_i, x_j) \in M$，$dis(x_i, x_j)$为x_i和x_j之间的距离，如果x_q为x_i的最近K邻近集里的其中一个对象，如果$dis(x_q, x_j) < dis(x_i, x_j)$，并且$(x_i, x_q) \notin C$，$(x_q, x_j) \notin C$，那么$(x_q, x_j) \in M$。

[0035] 例如正约束对x_i代表特征词“篮球”，x_j代表特征词“足球”，(篮球，足球)为正约束集的成员，经过距离计算发现“篮球”的最近K邻域有个特征词“操场”，用x_q表示，而这个特征词“操场”和“足球”的距离比“篮球”和“足球”之间的距离还要近，并且(足球，操场)和(篮球，操场)都不在负约束集中，因此把(足球，操场)纳入正约束集，这样从而扩充了正约束集。

[0036] 同理，扩充的集合C原则是距离负约束对其中一个对象的距离非常的近，距离另外一个对象的距离大于此负约束对的距离，并且此对象和负约束对的两个对象都不属于集合M，则此对象和其中一个对象的关系属于负约束关系。

[0037] 如图2所示，$(x_i, x_j) \in C$，$dis(x_i, x_j)$为x_i和x_j之间的距离，如果x_q为x_i的最近K邻近集里的其中一个对象，如果$dis(x_q, x_j) > dis(x_i, x_j)$，并且$(x_q, x_j) \notin M$，那么$(x_q, x_j) \in C$。

[0038] 例如正约束对x_i代表特征词“乐曲”，x_j代表特征词“数学”，(乐曲，数学)为负约束集的成员，经过距离计算发现“乐曲”的最近K邻域有个特征词“歌曲”，用x_q表示，而这个特征词“歌曲”和“数学”的距离比“乐曲”和“数学”之间的距离还要近，并且(歌曲，数学)不在正约束集中，因此把(歌曲，数学)纳入负约束集，这样从而扩充了负约束集。

[0039] 根据以上原理，扩充约束集的步骤为：

[0040] (1)通过约束集的对称性和传递性扩充集合M和集合C；

[0041] (2)通过K最邻近集扩充集合M；

[0042] (3)通过K最邻近集扩充集合C；

[0043] (4)循环以上步骤直到收敛。

[0044] 以上收敛后的正约束集具有自反性、对称性和传递性，因此正约束集也就是一个划分，也就是一个聚类结果。聚类的结果簇表示为$A_1, A_2, \cdots, A_k$，共聚k类。

[0045] 特征词合并

[0046] 特征词聚类之后，每类的特征词合并为一个属性，合并的方法用数值求和方式，计算方法如公式(3)所示，cw_{ip}为合并后并归一化的属性值，也就是第p个特征词簇在第i个文本中的权重之和的值，w_{ij}为第i篇文本中第j个词的权重。

[0047] $$cw_{ip} = \frac{\sum_{t_j \in A_p} w_{ij}}{\max_{g<m \& h<k} cw_{gh}} \qquad 1 \le p \le k \quad (3)$$

[0048] 公式(3)用了数值求和的方式计算特征簇在文本中的权重，可以看出一个特征词簇在一个文本中的权重越大，说明这个文本中含有的这个特征词簇中的特征词对象越多，并且这些特征词在这个文本中的权重越大。

[0049] 属性合并之后，新产生的向量空间矩阵如下公式(4)，$t_1, t_2, \cdots, t_k$分别表示第1，2，…，k个特征词簇，不再表示一个特征词，$d_1, d_2, \cdots, d_k$分别表示第1，2，…，k个文本，而从这

里可以看出向量空间模型的维度已经变为k维了。

$$W_{m\times k}=\begin{array}{c} \\ d_1 \\ d_2 \\ \vdots \\ d_m \end{array}\begin{array}{c} \begin{array}{cccc} t_1 & t_2 & \cdots & t_k \end{array} \\ \begin{bmatrix} cw_{11} & cw_{12} & \cdots & cw_{1k} \\ cw_{21} & cw_{22} & \cdots & cw_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ cw_{m1} & cw_{m2} & \cdots & cw_{mk} \end{bmatrix} \end{array} \quad (4)$$

[0050]

[0051] 特征词簇合并成新的属性之后，既降低了向量空间的维度，有利于层次协同聚类高效的进行，而且还利用少了的监督信息将特征词进行了聚类，使得一意多词的现象减少了，使聚类能更加有效的进行。

[0052] 由此可见，本发明方法的关键在于在文本聚类前对特征词进行聚类，找到特征词的簇，对特征词进行合并；以及扩充文本特征词对象集的相似度矩阵，通过特征词之间的相似性，找出文本之间的语义关系。

[0053] 作为本发明方法的另一个优选实施例，其中把这个向量空间映射到每一维子空间中，每一维实际就是这个矩阵的一个列，然后在这每一维的子空间中用K-means的方法对文本进行聚类，其中平均权重最高的那一个簇为高权重文本簇，这样样本空间有多少维就有几个相对应的聚类结果，就有若干个相对应的高权重文本簇，对高权重文本簇进行比较，交集的个数越高，两个相对应的特征词之间的相似度就越高，然后利用这些特征词之间的相似度用K-means算法对特征词进行聚类。

[0054] 以上所述的具体实施例，对本发明的目的、技术方案和有益效果进行了进一步详细说明，应理解的是，以上所述仅为本发明的具体实施例而已，并不用于限制本发明，凡在本发明的精神和原则之内，所做的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。

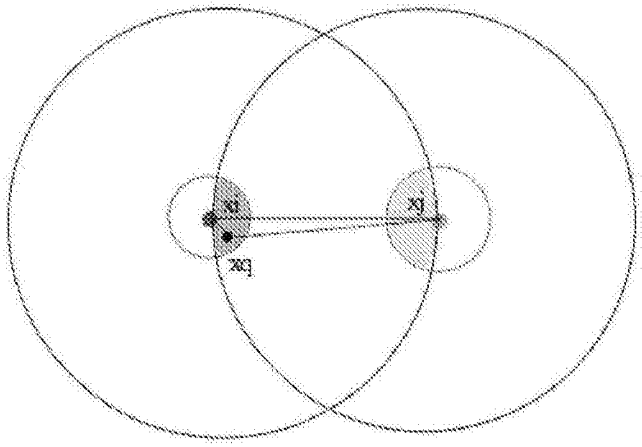

图1

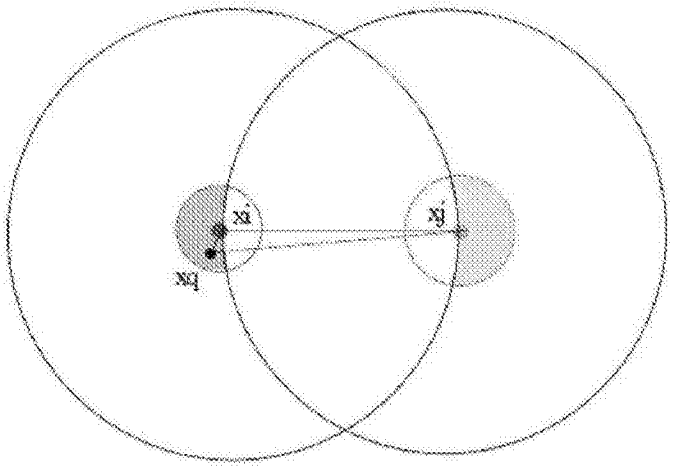

图2