文章编号:1001 - 9081(2005)07 - 1570 - 03

一个基于关联规则的多层文档聚类算法

宋江春,沈钧毅,宋擒豹
(西安交通大学 电子与信息工程学院,陕西 西安 710049)
(sjchun@163.net)

摘 要:提出了一种新的基于关联规则的多层文档聚类算法,该算法利用新的文档特征抽取方法构造了文档的主题和关键字特征向量。首先在主题特征向量空间中利用频集快速算法对文档进行初始聚类,然后在基于主题关键字的新的特征向量空间中利用类间距和连接度对初始文档类进行求精,从而得到最终聚类。由于使用了两层聚类方法,使算法的效率和精度都大大提高;使用新的文档特征抽取方法还解决了由于文档关键字过多而导致文档特征向量的维数过高的问题。

关键词:文档挖掘;文档聚类;关联规则;文档主题特征向量;文档关键字特征向量
中图分类号: TP311.11 **文献标识码**:A

Multi-level document clustering algorithm based on association rules

SONG Jiang-chun, SHEN Jun-yi, SONG Qing-bao
(*School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an Shaanxi* 710049, *China*)

Abstract: A multi-level document clustering algorithm was proposed based on association rules. It constructed ducument feature vector of topic and keyword by using a new method of document feature extraction. Firstly, it found the initial ducument clusters by using fast algorithm of finding frequent item sets in topic vector space, then in keyword vector space, re-clustered the initial clusters according to the cluster distance and the link intensity. For processing initial clustering by using classical fast frequent item sets, the efficiency and the precision of the algorithm were highly increased. The new method of ducument feature extraction is also used to solve the problem that the dimention of the keyword vector space is too high with increasing of keywords in document.

Key words: document mining; document clustering; association rule; document topic feature vector; document keyword feature vector

0 引言

由于电子形式的信息量的飞速增长,如电子出版物,互联网,电子邮件等,文本数据库得到了迅速发展。传统的信息检索技术已不适应日益增加的大量文本数据处理的需要,典型的是,大量文档中只有很少一部分与用户有关。而不清楚文档的内容,就很难形成有效的查询和从数据中分析和提取有用的信息。我们需要有关的工具来完成不同文档的比较,多文档的关联性等。数据挖掘研究主要是针对结构数据的,如关系的,事务的和数据仓库的数据。而文档数据库中存储最多的是半结构甚至是无结构数据,因此,对半结构或无结构数据的建模和实现就成为了文档数据挖掘的一个重要组成部分[1~2]。

文档聚类主要有基于概率的方法和基于距离的方法。基于概率的方法[3]以贝叶斯概率为理论基础,用概率的分布方式描述聚类结果,可处理类间相互重叠的情况,缺点是当特征空间的维数较高或特征值之间出现较强的相关性时,聚类的精度和效率均不能令人满意。基于距离的方法[4],典型的有 k-均值和 BIRCH 算法等,它们都以特征向量表示文档,再将文档看成向量空间中的一个点,通过计算点之间的距离来进行聚类,比较形象直观,缺点是当文档维数较高时聚类的质量和算法的效率都明显地下降。

我们认为,文档由一系列主题构成,而每一个主题包含若干个关键字。首先利用向量空间模型(Vector Space Model, VSM)对文档进行结构化处理,用文档主题特征向量形成文档主题事务矩阵。然后用经典的频集快速发现算法得到文档的初步聚类,再根据基于主题关键字的特征向量定义的类间距离和连接度对聚类进一步求精。我们给出了相应的算法,并对算法的有效性、可伸缩性以及算法的时间复杂度进行了研究。

1 文档的结构化表示

在向量空间模型(VSM)[5]中,文档被看作是由一组关键字 $\{k_1, k_2, \cdots, k_n\}$ 构成的。同时,对每一个关键字 k_i,根据它在文档中的重要性赋予一个权值 w_i,可以把 w_i 看成是 k_i 的坐标值,即一个文档就可以看成是 n 维空间的一个点。这样就将文档的聚类问题转化为 n 维向量空间中点的聚类问题。本文中,我们认为,文档由一系列主题 $\{t_1, t_2, \cdots, t_n\}$ 构成,而每一个主题包含若干个关键字 $t_i = \{k_{i1}, k_{i2}, \cdots, k_{imi}\}$。首先,用文档主题特征向量的值来表示文档在向量空间中的坐标值,下面对文档结构化过程中用到的概念进行具体的定义和描述。

1.1 建立文档事务矩阵

定义 1 文档的主题关联度

收稿日期:2005 - 02 - 03;**修订日期**:2005 - 04 - 01 **基金项目**:国家自然科学基金资助项目(60173058)

作者简介:宋江春(1962 -),男,四川成都人,工程师,博士研究生,主要研究方向:数据库与数据挖掘; 沈钧毅(1939 -),男,江苏扬州人,教授,博士生导师,主要研究方向:数据库理论、数据挖掘、工作流; 宋擒豹(1966 -),男,陕西华县人,副教授,博士,主要研究方向:数据仓库与数据挖掘.

关联度表示文档和某一主题的关联程度,文档 D_j 和主题 t_i 之间的关联度可按下式计算:

$$\alpha_{ij} = \frac{\| N_{ij} \|}{\| \cup D_j \|} \times w_{ij} \tag{1}$$

其中,$\| \cup D_j \|$ 为 D_j 中有效词的总数,$\| N_{ij} \|$ 表示 t_i 在 D_j 中出现的次数,w_{ij}表示主题 t_i 在文档 D_j 中的重要程度。

注:我们对 α_{ij} 进行规范化处理,仍记作 α_{ij},$\sum_{j=1}^{n} \alpha_{ij} = 1$,$1 \leqslant i \leqslant m$,$m$ 为 D 中文档个数。

定义 2　文档主题特征向量

设 D 是文档的集合,$D_i \in D$,称:

$$\overrightarrow{\alpha_i} = [\alpha_{i1}, \alpha_{i2}, \cdots, \alpha_{in}]^{\mathrm{T}} \tag{2}$$

为文档 D_i 的主题特征向量。我们认为,文档主题特征向量的值 α_{ij} 即为文档 D_i 在主题向量空间中的坐标值。

此时,可以将文档作为事务,而将文档主题特征向量作为事务项,建立如下的文档主题事务矩阵,以适用关联规则。

设 $m = \| D \|$ 表示文档的个数,把定义 2 中的文档特征向量 $\overrightarrow{\alpha_i} = [\alpha_{i1}, \alpha_{i2}, \cdots, \alpha_{in}]^{\mathrm{T}}$ 作为列向量构成的矩阵 $\boldsymbol{M}_{n\times m}$ 称为文档事务矩阵。

$$\boldsymbol{M}_{n\times m} = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1j} & \cdots & \alpha_{1m} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2j} & \cdots & \alpha_{2m} \\ \vdots & \vdots & & \vdots & & \vdots \\ \alpha_{i1} & \alpha_{i2} & \cdots & \alpha_{ij} & \cdots & \alpha_{im} \\ \vdots & \vdots & & \vdots & & \vdots \\ \alpha_{n1} & \alpha_{n2} & \cdots & \alpha_{nj} & \cdots & \alpha_{nm} \end{bmatrix} \tag{3}$$

1.2　相关算法

算法:建立规范化文档主题事务矩阵

输入:文档集 $\{D_i\}$

输出:文档规范化主题事务矩阵

功能:建立规范化文档主题事务矩阵

方法:

```
for each D_j ∈ D
  for each t_i
    computing α_ij according to formula (1)
  endfor
endfor
for each D_j ∈ D
  for each t_i
    α_ij = α_ij / Σ_{j=1}^{m} α_ij
  endfor
endfor
```

本算法分为两部分,第一部分计算文档和主题的关联度,第二部分对文档主题特征向量进行规范化,其算法的时间复杂度为 $O(n \times m)$。

2　基于关联规则的文档聚类算法

根据上面已经建立的文档主题事务矩阵,就可以对文档进行聚类了。首先在文档主题向量空间中利用经典的关联规则发现算法进行文档的初始聚类,然后在基于关键字的新的特征向量空间中利用类间距和连接度对初始聚类进行求精,得到最终聚类。

2.1　发现初始文档类

在文档主题事务矩阵中,我们将文档看作事务,将文档主题特征向量视为事务项。如果某些文档主题特征向量经常一起出现在某些文档中,那么,对应的文档也应该是相似的。也就是说,根据关联规则挖掘算法得到的频集,就可以得到对应的文档集,并将其作为初步的文档分类结果。为此,我们对支持度 - 文档集在整个文档集中出现的频度,作如下的定义:

定义 3　支持度

设 D 是文档的集合,任意 $X \subset D$,其支持度定义为:

$$s(X) = \frac{1}{\| X \|} \sum_{i=1}^{n} \sum_{j=1}^{\| X \|} \alpha_{ij} \tag{4}$$

其中,$\| X \|$ 是 X 中文档的个数。

该定义将 X 的支持度定义为 X 中文档与主题关联的平均值。利用定义的支持度 $s(X)$,我们可以利用 Agrawal 等人提出的频集发现快速算法[6]获得主题频集。然后扫描整个数据库就可以得到对应的文档集。设得到的初始聚类为 $C = \{C_i\}$。

2.2　文档再聚类

设一个主题包含若干个关键字 $t_i = \{k_{i1}, k_{i2}, \cdots, k_{imi}\}$,其中每个主题的关键字的个数可以不同,用 l_i 表示向量 t_i 的长度。w'_{ir} 表示关键字 k_{ir} 在主题 t_i 中的重要程度。

构造新的基于关键字的特征向量,在新的特征向量空间中依据类间距和连接度对文档再聚类。首先给出相关的定义。

定义 4　文档的关键字关联度

文档 D_j 和主题 t_i 的关键字集合 $\{k_{i1}, k_{i2}, \cdots, k_{imi}\}$ 之间的关联度可按下式计算:

$$\beta_{ij} = \frac{\left\| \sum_{r=1}^{l_i} N_{ir}^{(j)} \times w'_{ir} \right\|}{\left\| \cup D'_j \right\|} \tag{5}$$

其中,$\left\| \cup D'_j \right\|$ 为 D_j 中关键字的总数,$\left\| N_{ir}^{(j)} \right\|$ 表示 k_{ir} 在 D_j 中出现的次数,w'_{ir} 表示关键字 k_{ir} 在主题 t_i 中的重要程度。

和前面一样,我们对 β_{ij} 进行规范化处理,仍记作 β_{ij}。

定义 5　文档关键字特征向量

设 D 是文档的集合,$D_i \in D$,称:

$$\overrightarrow{\beta_i} = [\beta_{i1}, \beta_{i2}, \cdots, \beta_{in}]^{\mathrm{T}} \tag{6}$$

为文档 D_i 的文档的关键字特征向量。

从以上定义可以看出,关键字特征向量仍然是 n 维的,和关键字的总数无关,从而使特征向量的维数降低,为提高文档聚类的时间效率打下了基础。

定义 6　类间距离

设 D 是文档的集合,$G_a, G_b \subset D$,G_a, G_b 的类间距离定义为:

$$D(G_a, G_b) = \frac{1}{\| G_a \| \| G_b \|} \sum_{i=1}^{\| G_a \|} \sum_{j=1}^{\| G_b \|} d(D_i, D_j) \tag{1}$$

其中,$\| G_a \|$,$\| G_b \|$ 为 G_a, G_b 中文档的个数,$d(D_i, D_j)$ 表示 D_i 与 D_j 之间在文档关键字特征向量空间中的欧几里德距离。文档关键字特征向量的值 β_{ij} 即为文档 D_i 在文档关键字特征向量空间中的坐标值。

定义 7　连接度

设 D 是文档的集合,任意 $D_i \in D$,$C \subset D$,D_i 和 C 的连接度定义为:

$$p(D_i, C) = \frac{1}{\| C \|} \sum_{j=1}^{\| C \|} \frac{\sum_{k=1}^{n} \beta_{ik} \beta_{jk}}{\left\| \overrightarrow{\beta_i} \right\| \left\| \overrightarrow{\beta_j} \right\|} \tag{8}$$

从以上定义可以看出,类间距离表示类间耦合性,连接度则表示类的内聚性。

对关联规则发现的文档类在新的文档关键字特征向量空间中进行重新聚类以提高聚类的精度,再聚类分为两个步骤:

1) 计算不同文档类之间的类间距,对类间距大于指定阈值的类进行合并;

2) 计算同一文档类中文档和类的连接度,对连接度小于指定阈值的类进行拆分。

2.3 相关算法

算法:文档聚类算法

输入:初始聚类 $C = \{C_i\}$;δ:耦合阈值;γ:内聚阈值

输出:最终文档聚类

功能:对初始文档类进行再聚类

```
for each C_i ∈ C do
  for each C_j ∈ C do
    computing D(C_i, C_j) according to (7);
    if D(C_i, C_j) ≤ δ then
      C_i = C_i ∪ C_j;
      delete C_j from C;
    endif
  endfor
endfor
for each C_i ∈ C do
  for each D_j ∈ C_i do
    computing p(D_j, C_i) according to (8)
    if p(D_j, C_i) < γ then
      for each C_k ∈ C do
        computing p(D_j, C_k) according to (8)
        ifp(D_j, C_k) ≥ γ then
          C_j = C_j -{ D_j };
          C_k = C_k +{ D_j };
        endif
      endfor
    endif
  endfor
endfor
ANS = ∪ C_i
```

该算法分为合并和剔除两部分,合并过程的时间复杂度为 $O(\|C\| \times \|C\|)$。剔除过程在剔除连接度不够的文档时,重新计算了它和其他文档的连接度,其时间复杂度为 $O(\|C\| \times \|C\| \times \sum_i \|C_i\|)$,因此,该算法总的时间复杂度为 $O(\|C\|^2(1 + \sum_i \|C_i\|))$。

3 实例测试及实验结果

为了对本文算法进行评价,我们对算法进行了实例测试,并将它和 k-均值聚类算法进行了比较。整个实验在 PⅣ 2.0 计算机的 Windows 2000 平台上进行。

首先测试算法的精确度,用搜索引擎 Google 从 Internet 上根据不同的主题及主题关键字进行了 50 次搜索,下载每次搜索到的前 20 个文档构成 50 个文档类,每个文档类有 1 000 个文档。用人工的方法剔除无关文档,同样将它们分成 50 个文档类,并以此作为分类准确性的基准,然后采用本文算法和 k-均值算法对这个文档类进行聚类。

由于不同聚类算法产生的类的数目很可能不同,为了使比较更公平,选用了各自质量最好的 40 个类进行比较,图 1 即为由不同算法产生的 40 个类的平均精度的对比情况。由于本文算法允许类间重叠,并采用了类确认技术,因此平均精度高于 k-均值算法。

然后测试算法的可伸缩性,同样用 Google 从 Internet 上根据不同的主题及主题关键字进行了 50 次搜索,不过每次下载搜索到的前 5 个文档,并以 5 的增幅递增,形成的文档数依次为 250,500,750,1000,1250,1500,1750,2000,2250 的 9 个文档集。然后分别采用本文算法和 k-均值算法对它们进行聚类,并计算它们的平均聚类时间,得到图 2 所示的结果。图 2 表明,随着文档数的增加,两种算法的平均执行时间都在增加,不过本文算法增长的趋势较为缓慢,说明本文算法的可伸缩性较好。其原因在于本文算法采用了新的文档特征向量抽取方法,降低了文档特征向量的维数,同时又先进行了基于频集发现快速算法的初始聚类,减少了数据处理的工作量。

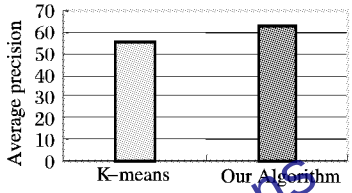

图 1 两种算法准确率比较

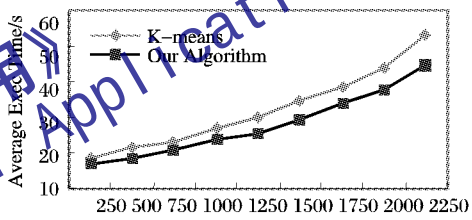

图 2 两种算法执行时间比较

4 结语

本文中,用新的文档特征抽取方法和经典的关联规则发现算法,提出了一个新的文档多层聚类算法。由于利用了频集快速算法进行了初始聚类和类确认技术,使算法的效率和精度都大大提高。本文还使用新的文档特征抽取方法解决了由于文档关键字过多而导致文档特征向量的维数过高的问题。实验表明,我们的算法无论在精度还是在可伸缩性上都优于 k-均值算法。

参考文献:

[1] HAN J, KAMBER M. Data Mining: Concepts and Techniques[M]. Morgan Kaufmann Publisher, 2000.

[2] FELDMAN R, DAGAN I. Knowledge discovery in textual database (KDT)[A]. Proceedings of 1st International Conference on Knowledge Discovery and Data Mining[C]. Montreal, Canda, 1995.

[3] ACKERMAN M, BILLSUS D, GAFFNEY S, *et al*. Learning Probabilistic User Profiles[J]. AI Magazine, 1997, 18(2): 47 - 56.

[4] KAUFMAN L, ROUSSUW PJ. Finding Groups in Data: An Introduction to Cluster Analysis[M]. New York: John Wileyb and Sons, 1990.

[5] SATON G, WONG A, YANG CS. A vector space model for automatic indexing[J]. Communications of ACM, 1975, 18(11): 613 - 620.

[6] AGRAWAL R, STRIKANT R. Fast algorithm for mining association rules[A]. Proceedings of 20th International Conference on Very Large Database[C]. Santiago, Chile, 1994. 487 - 499.