

## 基于频繁特征项集的文档聚类研究

郑小慎

(天津科技大学 海洋科学与工程学院, 天津 300457)

(zxs@tust.edu.cn)

**摘 要:**提出了基于频繁特征项集的文档聚类方法。对预处理后的文档,通过 Apriori 算法找出文档频繁特征项集,依据其子集中频繁特征词语对相关文档进行聚类,该方法能够有效降低特征项的维数,并能够通过频繁特征词语集合对聚类后的类别进行适当的描述。

**关键词:**频繁特征项集;文档聚类;文档挖掘

**中图分类号:** TP311.13 **文献标识码:** A

## Documents clustering based on frequent term sets

ZHENG Xiao-shen

(College of Marine Science and Engineering, Tianjin University of Science and Technology, Tianjin 300457, China)

**Abstract:** Novel method of frequent term sets-based text clustering was presented. For the pre-treated documents, such frequent sets could be efficiently discovered by using the Apriori method, which subset was used for documents clustering according to frequent terms' correlation. This method allows us to reduce drastically the dimensionality of the term sets and provides an understandable description of the discovered clusters by their frequent terms sets.

**Key words:** frequent term sets; text clustering; text mining

### 0 引言

随着文档信息的日益丰富,新的内容层出不穷,预先定义类别已不能包容所有的内容,聚类方法成为文档挖掘的一个新的研究热点。

基于平面划分和基于层次的聚类主要是根据对象间的距离来进行的。通过计算点之间的距离进行聚类,比较形象直观;其缺点是特征向量必须经过规范化处理,以避免由于文档长度不同或各个文档间关键词出现的频度各异而产生的畸变,特别是当向量数据的维数较高时,聚类的质量和算法的性能都明显下降。基于密度的方法能够过滤“噪声”文档,但对于参数的输入很敏感,且计算复杂度高。基于网格的方法对文档数据的输入顺序不敏感,但由于在处理高维数据空间时对方法进行了简化,降低了聚类的精确性。基于模型聚类方法的数据是根据潜在的概率<sup>[1]</sup>分布生成的,其缺点是当文档特征项的维数较高或特征值间呈现出较强的相关性时,聚类精确度和效率均不能令人满意。后缀树文档聚类方法摆脱了向量空间模型的束缚<sup>[2]</sup>,它的基本思想是依靠一棵后缀树来识别含有共同词语或短语的文档,并利用这些信息进行文档聚类。该方法在文档聚类效果上有了一定的提高,但依据后缀树的短语聚类,有时会导致局部最优,而且对结果不能给出一个合理的描述。

针对上述问题,我们提出基于频繁特征项集的文档聚类方法<sup>[3]</sup>。对预处理后的文档,通过关联规则中的 Apriori 算法找出文档频繁特征词语的集合,然后对其子集中与频繁特征词语相关的文档进行聚类,并通过频繁特征词语集合间的相互重叠来确定文档类别之间的重叠。这种方法通过发现频繁特征项集对文档的维数进行约减,提高了算法运行的准确率和

速度;而且频繁特征项集可以对文档类别进行较精确的描述,有效地提高了聚类的性能,并且能够处理文档类之间存在的固有重叠情况。

### 1 基于频繁特征项集的文档聚类

聚类过程的处理对象是文档数据库中的文档,因此我们用特征词语集合的概念,提出基于频繁特征项集的文档聚类方法(Frequent Term Set-based text Clustering, FTSC),其中的特征词语是提取大于一定阈值的特征词语。这种聚类方法对预处理后的文档,通过 Apriori 算法找出文档频繁特征项集,将其子集中的频繁特征词语作为候选聚类,而不是直接对高维的文档向量进行聚类,能够有效降低特征项的维数,同时,一个频繁特征项集的子集对应一个文档类别,可以为聚类类别提供比较准确的描述。

我们提出的 FTSC 方法将频繁特征词语集合作为候选聚类,是一种能够有效地降低文档向量维数的新颖方法。其运行过程是自底向上,首先从一个空的集合开始,不断地从剩余的频繁特征词语集合中选择元素,直到整个集合中所有满足条件的元素都被选择。在计算的每一步,FTSC 算法选择的元素都是和其他的聚类候选元素相比具有最小类别重叠信息熵的频繁特征词语。迭代计算的过程中,剩余候选聚类文档重叠信息熵的值均被重新计算。

依据 Apriori 算法及其改进方法发现频繁词语集合,可以得到大于最小支持度的文档集合  $D$  的所有频繁特征词语集合  $F$ ,集合  $F$  中的一个子集  $F_i$  构成一个聚类类别,它所包含的特征词语可以看作是对文档集合  $D$  中该类别的一种描述。在聚类的过程中,希望得到的聚类是互相重叠数量最少的类别。设  $f_j$  为文档集合  $D$  的文档  $D_j$  对应的所有频繁特征项集的数目,即:

$$f_j = |\{F_j \in R \mid F_j \subseteq D_j\}|$$

收稿日期:2005-10-31;修订日期:2006-01-09

基金项目:天津市高等学校科技发展基金项目(20051505);天津科技大学引进人才科技启动基金(20050420)

作者简介:郑小慎(1973-),女,河北献县人,副教授,博士,主要研究方向:信息处理、遥感、信息检索。

其中  $|$  表示集合的数目,  $R$  表示频繁特征词语集合  $F$  的一个子集。 $F_j$  是文档  $D_j$  中子集  $R$  内的特征词语集合。

当类别  $C_i$  和其他类别之间的重叠越少时, 类别  $C_i$  中文档  $D_j$  的  $f_j$  值也就越小。理想情况下,  $f_j$  对应的类别  $C_i$  中的文档只属于一个类别, 也就是说  $f_j = 1$ 。所有文档均支持类别  $C_i$ , 这样  $C_i$  与其他候选类别的重叠为 0。定义类别  $C_i$  的标准重叠为该类别  $(f_i - 1)$  的均值, 用  $SO(C_i)$  表示:

$$SO(C_i) = \sum_{D_j \in C_i} (f_j - 1) / |C_i|$$

标准重叠虽然容易计算, 但是频繁项集具有单调性(频繁项集的子集仍是频繁项集), 当文档支持  $m$ -项集时, 那么它必然支持相应的  $(m-1)$ -项集,  $(m-2)$ -项集,  $\dots$ , 2-项集, 1-项集等。因此一个由许多项描述的候选项集的标准重叠常常会大于由少许项描述的候选项集的标准重叠。从而, 频繁项集的项数越少, 它的标准重叠也就越小。

为了减小这方面的影响, 利用信息熵来定义类别的重叠。信息熵表示文本对一部分类别的支持度与其对剩余参考类别的支持度的比值分布。例如, 用  $f_j$  描述文档  $D_j$  在候选类别中的分布, 用  $p_j = \frac{1}{f_j}$  表示文档  $D_j$  属于一个特定候选类别的概率。当  $f_j = 1$  时,  $p_j = 1$ , 也就是说文档  $D_j$  只属于一个类别。当  $f_j$  增大时,  $p_j$  值会相应减小, 也就是说文档  $D_j$  属于该类别的概率减小, 可能存在类别重叠现象。如果用  $EO(C_i)$  表示类别  $C_i$  的重叠信息熵, 则:

$$EO(C_i) = \sum_{D_j \in C_i} -\frac{1}{f_j} \cdot \ln\left(\frac{1}{f_j}\right)$$

当  $f_j = 1$  时, 重叠信息熵为 0, 即类别  $C_i$  中的文档  $D_j$  只属于这一个类别, 与其他类别之间不存在重叠现象。

为了发现具有最小聚类重迭的聚类, 使用贪婪算法来进行文档的聚类。该算法的时间复杂度由频繁特征词语聚类的内在复杂性决定。

FTSC 方法的聚类过程如下, 其中 DetermineFrequentTermsets 是发现有最小支持度的文档集合  $D$  中所有频繁特征词语集合的有效方法。

```

FTSC(database D, float minsup)
SelectedTermSets: = {};
n: = |D|;
RemainingTermSets := DetermineFrequentTermsets(D, minsup);
while |cov(SelectedTermSets)|  $\neq$  n do
for each set in RemainingTermSets do
Calculate overlap for set;
BestCandidate: = element of Remaining
TermSets with minimum overlap;
SelectedTermSets := SelectedTermSets  $\cup$  {BestCandidate};
RemainingTermSets := RemainingTermSets - {BestCandidate};
Remove all documents in cov(BestCandidate) from D and from the
coverage of all of the RemainingTermSets;
return SelectedTermSets and the cover of the elements of
SelectedTermSets;
```

表 1 说明了一个含有 16 篇文档的数据库中, FTSC 方法的第一步过程, 即依据最小的重叠信息熵, 得到了由 {体育, 篮球, 运动} 来描述的聚类类别, 文档  $D_8, D_{10}, D_{11}, D_{15}$  是从文档数据集中选出的属于该类的文档。得到聚类的同时将文档  $D_8, D_{10}, D_{11}, D_{15}$  从文档数据库集合中移出。因此,

FTSC 方法形成的聚类不存在类别间的重叠, 同时可以返回聚类的描述。

表 1 16 篇文档应用 FTSC 方法的第一步聚类过程

频繁特征词语集合	候选聚类	EO
{体育}	{ $D_1, D_2, D_4, D_5, D_6, D_8, D_9, D_{10}, D_{11}, D_{13}, D_{15}$ }	2.98
{运动}	{ $D_1, D_3, D_4, D_6, D_7, D_8, D_{10}, D_{11}, D_{14}, D_{15}, D_{16}$ }	3.0
{篮球}	{ $D_2, D_7, D_8, D_9, D_{10}, D_{12}, D_{13}, D_{14}, D_{15}$ }	2.85
{排球}	{ $D_1, D_2, D_6, D_7, D_{10}, D_{11}, D_{12}, D_{14}, D_{16}$ }	2.73
{体育, 运动}	{ $D_1, D_4, D_6, D_8, D_{10}, D_{11}, D_{15}$ }	1.97
{运动, 排球}	{ $D_1, D_6, D_7, D_{10}, D_{11}, D_{16}$ }	1.72
{体育, 篮球}	{ $D_2, D_8, D_9, D_{10}, D_{11}, D_{15}$ }	1.72
{体育, 排球}	{ $D_1, D_2, D_6, D_{10}, D_{11}$ }	1.34
{运动, 篮球}	{ $D_7, D_8, D_{10}, D_{14}, D_{15}$ }	1.47
{篮球, 排球}	{ $D_2, D_7, D_{10}, D_{12}, D_{14}$ }	1.47
{体育, 运动, 排球}	{ $D_1, D_6, D_{10}, D_{11}$ }	0.98
{体育, 篮球, 运动}	{ $D_8, D_{10}, D_{11}, D_{15}$ }	0.9

如果对 FTSC 方法进行简单修改, 不把已选出的与频繁词语集合相关的文档从文档数据库中移出, 那么该方法也可以产生重叠聚类。

## 2 实验结果

为了更全面地评价基于频繁特征项集的文档聚类方法的性能, 首先利用 NTCIR-3 文档数据集对 FTSC 方法进行定性评价, 然后利用 NTCIR-3 和 Reuters-21578<sup>[4]</sup> 两个文档数据集对 FTSC 方法的聚类结果进行定量评价, 最后将其聚类结果与 K-means 方法的聚类结果进行比较。

### 2.1 FTSC 方法的定性评价

下面采用信息熵的方法在 NTCIR-3 文档集合上定性评估 FTSC 方法的聚类结果。

设  $k = \{k_1, \dots, k_k\}$  是文档数据库中的标准类别集合, 聚类得到的类别  $C_j$  的信息熵定义为:

$$E(C_j) = \sum_j \frac{n_j}{|D|} \sum_i -p_{ij} \log(p_{ij})$$

其中  $|D|$  表示文档数据库中文档总数量,  $n_j$  表示聚类类别  $C_j$  中的文档数量,  $p_{ij}$  是聚类类别  $C_j$  中的文档属于类别  $k_i$  的概率值。聚类类别  $C_j$  的信息熵表示了聚类类别的准确性, 取值在区间  $[0, \log(|k|)]$  上。信息熵的值越小, 产生的聚类类别越准确。

表 2 FTSC 方法对 NTCIR-3 标准测试集的聚类

最小支持度	聚类类别数	最小聚类熵值	取最小聚类熵值时类别中文档的覆盖率(%)	包含全部文档时最终的聚类熵值
0.5	14	0.249	82.5	0.380
1.0	10	0.307	81.3	0.404
1.5	5	0.364	80.0	0.485

图 1 给出了对于 NTCIR-3 文档集合, 当最小支持度分别取 0.5, 1.0 和 1.5 时, 文档聚类的信息熵和文档集合中被聚类文档所占的比例。这些实验表明, 随着聚类数目  $k$  的增加, 聚类的熵值并不是单调地线性递减。因此, 通过该聚类方法可以得到人们乐于接受的聚类结果(而不是每个叶子结点只有一个文档)。

表 2 给出了与每个最小支持度对应的最小聚类熵值, 当选定最小聚类熵值时, 类别中包含文档的覆盖率(约为 81% 左右), 以及聚类的类别数目。如果要求聚类结果包含数据库中的全部文档, 则最终的聚类熵值将会增大 40% 左右。

由以上实验可以看出,随着最小支持度的增加,聚类的类别数减少。

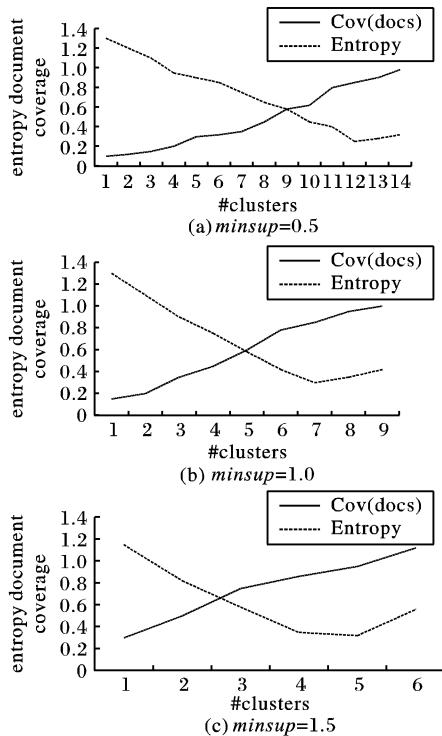


图1 最小支持度取 0.5, 1.0 和 1.5 时的文档聚类

## 2.2 FTSC 方法的定量评价

在 NTCIR-3 和 Reuters-21578 两个文档数据集上采用 F-测量 (F-Measure) 方法对 FTSC 方法聚类结果的定量评价, 并把 K-means 方法与这种方法的聚类结果进行比较。

我们选择的文档集合如表 3 所示。

表3 文档数据集合

数据集	#文本数	#类别数
NTCIR-3	10500	14
Reuters-21578	8654	52

选用 F-测量方法进行聚类结果评价的原因是: F-测量方法综合了精确度和召回率的结果。精确度和召回率是两个相互矛盾的衡量标准, 一般情况下, 精确度会随着召回率的升高而降低, 两者不可兼得。

F-测量的定义如下:

$$F_{\beta}(P, R) = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

其中  $\beta$  是一个调整参数, 用于以不同得权重综合精确度和召回率。当  $\beta = 1$  时, 表示精确度和召回率被平等对待, 此时 F-测量又被称为  $F_1$ , 定义如下:

$$F_1(P, R) = \frac{2PR}{P + R}$$

FTSC 方法虽然通过频繁特征词语集合的挖掘降低了文档的维数, 但通过表 4 的数据可以看出, 其聚类结果与 K-Means 方法差别不大。另一方面, FTSC 方法可以进行文档的重叠聚类, 与不能形成具有重叠关系的文档聚类方法 K-Means 相比, 具有显著的优越性。前面已经说过, 进行文档聚类时, 如果不考虑类别重叠, 可能会由于个别文档的聚类错误导致在将来聚类过程中重大的误差。因此, 基于频繁特征词语集合的 FTSC 方法是实用性强, 聚类效果较好的方法。

表4 K-means 方法和 FTSC 方法在不同文档集合中 F1 值的比较

Data Set	FTSC	K-means
NTCIR-3	0.38	0.43
Reuters-21578	0.47	0.56

## 3 结语

本文针对当前文档聚类方法中, 存在向量维数大, 对类别没有描述的问题, 提出基于频繁特征词语集合文档聚类的 FTSC 方法, 该方法用 Apriori 方法得到频繁特征词语集合, 有效地降低了向量的维数。在此基础上对文档进行聚类, 频繁特征词语集合能够对聚类结果给出适当的描述。实验表明, 基于频繁特征词语集合的文档聚类方法能够得到较好的聚类效果。

### 参考文献:

- [1] ACKERMAN M, BILLISUS D, GAFFNEY S. Learning probabilistic user profiles[J]. AI Magazine, 1997, 18(2): 47-56.
- [2] ZAMIR O, ETZIONI O. Web Document Clustering: A Feasibility Demonstration[A]. Proceedings of ACM SIGIR 98[C]. 1998. 46-54.
- [3] 郑小慎. 文档分类和聚类方法及其在信息检索中应用的研究[D]. 博士学位论文, 天津大学, 2004.
- [4] LEWIS DD. Reuters-21578[DB/OL]. <http://www.daviddlewis.com/resources/testcollections/>, 2005.
- [5] Third IEEE International Conference on Data Mining[C]. 2003.
- [7] (加) HAN JW, KAMBER M. 数据挖掘概念与技术[M]. 范明, 孟小峰, 译. 北京: 机械工业出版社, 2001.
- [8] CHEN B, HAAS PJ, SCHEUERMANN P. A new two-phase sampling based algorithm for discovering association rules[A]. Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining[C]. Andreas, 2002.
- [9] 王春花, 黄厚宽. 利用抽样技术分布式开采可变精度的关联规则[J]. 计算机研究与发展, 2000, 38(9): 1101-1106.
- [10] PRODROMIDIS L, CHAN PK. Meta-learning in distributed data mining systems: Issues and approaches[M]. Advances in Distributed Data Mining, MIT: MIT Press, 2000. 81-113.
- [11] 数据挖掘研究院[EB/OL]. <http://www.dmresearch.net/forum/index.jsp>, 2005.
- [12] SCHUSTER A, WOLFF R. Communication-efficient distributed mining of association rules[A]. Proceedings of the ACM SIGMOD Int'l. Conference on Management of Data[C]. Santa Barbara, California, 2001. 473-484.

(上接第 874 页)

- [2] BRIN S, MOTWANI R, ULLMAN JD, et al. Dynamic itemset counting and implication rules for market basket data[A]. Proceedings ACM SIGMOD International Conference on Management of Data[C]. Tucson, Arizona, USA, 1997.
- [3] SAVASERE A, OMIECINSKI E, NAVATHE S. An efficient algorithm for mining association rules in large databases[A]. Proceedings of 21th Int'l Conference on Very Large Data Base[C]. Switzerland, 1995. 432-444.
- [4] PARK JS, YU PS, CHEN MS. Mining association rules with adjustable accuracy[A]. Proceedings of the fourth Int conf on Knowledge Discovery and Data Mining[C]. New York. 1998.
- [5] CHEUNG DW, HAN JW, NG VT, et al. Fast distributed algorithm for mining association rules[A]. Proceedings of Int'l Conference on Parallel and Distributed Information Systems[C]. Florida, 1998. 31-44.
- [6] SCHUSTER A, WOLFF R, TROCK D. A high performance distributed algorithm for mining association rules[A]. Proceedings of the