一种基于关联规则的中文概念集生成算法*)

赵 心 蔡 智 洪 流 蔡庆生

(中国科学技术大学计算机科学技术系 合肥230027)

摘 要 本文提出了一种基于关联规则的中文概念集生成算法。该算法首先产生文档的中文关键词集,采用向量空间模型VSM(vector space model)表示文档;然后以中文关键词为事务项,以中文文档为事务,采用成熟的关联规则算法发现中文关键词频繁集;再生成原始概念集并对原始概念集进行聚类,最终实现了中文概念集的自生成,同时该算法能引入增量更新的特性,对概念集进行增量更新。通过实验,表明该算法能有效地生成中文概念集,可以用之于对表示中文文档的高维特征向量的语义降维,具有一定的使用价值。

关键词 关联规则,中文概念集,文本聚类,增量更新

A Chinese Conception Sets-Generating Algorithm Based on Association Rule

ZHAO Xin CAI Zhi HONG Liu CAI Qing-Sheng

(Department of Computer Science and Technology, University of Science and Technology of China, Hefei 230027)

Abstract This paper proposes a Chinese conception sets-generating algorithm based on association rule. The authors adopt VSM(vector space model) to represent document. Each keyword corresponds to a transaction and each document corresponds to a transaction-item. The authors use the association rules discovery algorithm to find the frequency Chinese conception sets, and use k-means clustering algorithm to cluster the original conception sets. Noteworthily, this algorithm can introduce the characteristic of increment updating to generate conception sets. The experiment results show this algorithm is effective for generating the Chinese conception sets.

Keywords Association rule, Chinese conception sets, Text clustering, Increment updating

1 引言

随着信息技术的发展和互联网的普及,中文的文本资源在近几年呈现出了几何级数的增长,如何利用海量的中文文本资源,从中发现有用信息,就显得颇为重要。

在对海量中文文本进行聚类等分析时,需要用向量空间模型VSM(vector space model)[1]来表示中文文本,经过滤去停用词及高频无用词等加工,特征向量的维度也达到数千维,在如此高维的特征空间中执行聚类等操作无论是精度还是速度都不会令人满意。如果能够利用概念集来生成表示文本的概念集向量,那么可以显著提高文本聚类的性能,实现基于概念的文本聚类[1]。

在进行中文信息检索时,传统的基于关键词的信息检索的查全率较低,如果将基于关键词的信息检索扩展为概念查询,也就是根据给定关键词所在的概念集,执行概念匹配,从而扩展了查询,提高了信息检索的查全率。

目前中文概念集的生成主要有两种方法:基于概率分布的概念生成算法[2]和基于自组织映射神经网络的概念生成算法(Self-Organization Map, SOM)[7],但这两种算法都不能很好地解决增量更新的概念生成的问题。本文提出了一种基于关联规则的中文概念集生成算法,首先采用向量空间模型VSM(vector space model)表示文档,将关键词和文档之间的关系描述成事务的形式,然后采用关联规则挖掘算法生成关键词之间的强关联规则,再根据强关联规则产生原始概念集簇,最后对原始概念集簇进行聚类产生最终中文概念集,同时该算法引入了关联规则增量更新的特性,可以对概念集进行增量更新。实验结果表明本文提出的算法能高效地生成中文概念集,同时可以自由设定概念集的个数,具有很好的扩展性。

本文第2节给出了文档集的结构化表示,第3节给基于关联规则的中文概念集生成算法的关键问题和实现,第4节给出了本算法结果的分析,最后是结语。

2 文档集的结构化表示

与数据库中的结构数据相比,Web 文档具有有限的结构,或者根本就没有结构,即使具有一些结构,也是着重于格式而非文档内容。另外,文档的内容是人类所使用的自然语言,计算机难以处理其语义。向量空间模型 VSM(vector space model)[1]是近年来应用最多并且效果较好的文档表示方法,同时该方法还可以对向量进行优化。本文就是采用VSM 来表示中文文档内容。

将中文文档库中的文档切分成词,当文档库比较大时,分词后可能得到甚至高达几十万个词条,在如此高维的向量空间中进行概念聚类,其效率和准确度显然不能令人满意,因此我们需要对关键词进行约简。首先根据停用词表和高频词表,过滤去停用词也就是文档中无实际意义的词和高频无效词

*)本文得到国家自然科学基金项目资助(No. 90104030)。赵 心 硕士研究生,主要研究领域:知识发现、数据仓库。蔡 智 博士,主要研究领域:人工智能,知识发现。洪 流 博士生,主要研究领域:人工智能,知识发现。蔡庆生 教授,博导,主要研究领域:人工智能、机器学习、知识发现。

(如"可以","能够"等),初步生成中文文档库的关键词集。根据经验,20个关键词就能很好地表达一篇文档的特征,因此我们计算出一篇文档中每个词条的频率和位置的加权权重,取前20个权重比较大的词条来作为代表该文档的关键词集。考虑到一个词在所有文档中出现的频率如果过低,那么包含该词的概念集的置信度就不高,基于这点考虑,本文将在少于3篇文档中出现的词条忽略。经过上述处理,生成了文档库的关键词集。

我们利用文档库的关键词集来形成文档的向量空间,对于文档 d 可表示为一个范化特征向量:

$$V(d)=(t_1,\omega_1(d);\cdots;t_i,\omega_i;\cdots;t_n,\omega_n)$$

其中 t_i 为词条项,$\omega_i(d)$ 为 t_i 在文档 d 中的权值,$\omega_i(d)$ 一般被定义为 t_i 在 d 中出现概率 $tf_i(d)$ 的函数,即 $\omega_i(d)=\Psi(tf_i(d))$。本文的核心思想是通过挖掘关键词在文档中的共现率来发现中文概念集,只关心关键词 t_i 在文档 d 中是否出现,所以权值函数采用布尔函数。

$$\Psi=\begin{cases}1, tf_i(d)\geqslant 1\\ 0, tf_i(d)=0\end{cases}$$

这样,我们就可以生成了表示文档库的关键词-文档矩阵:

$$\lambda_{m\times n}=\begin{bmatrix}\omega_1(d_1) & \omega_2(d_1) & \cdots & \omega_i(d_1) & \cdots & \omega_n(d_1)\\ \omega_1(d_2) & \omega_2(d_2) & \cdots & \omega_i(d_2) & \cdots & \omega_n(d_2)\\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots\\ \omega_1(d_j) & \omega_2(d_j) & \cdots & \omega_i(d_j) & \cdots & \omega_n(d_j)\\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots\\ \omega_1(d_m) & \omega_2(d_m) & \cdots & \omega_i(d_m) & \cdots & \omega_n(d_m)\end{bmatrix}$$

若 $\omega_i(d_j)=1$,则说明词条 t_i 是文档 d_j 的关键词之一。如果以文档为事务,以关键词为事务项,那么矩阵 $\lambda_{m\times n}$ 就可以看做是事务矩阵,可以在矩阵 $\lambda_{m\times n}$ 上进行关联规则的挖掘。

3 基于关联规则的概念集生成

根据文档库建立了关键词-文档事务矩阵 $\lambda_{m\times n}$,设定支持度和置信度的阈值,运用关联规则挖掘算法,可以发现关键词频繁集,然后从关键词频繁集生成原始概念集,再对原始概念集执行聚类,就可以生成需要的中文概念集。

3.1 根据关联规则发现原始概念集

关联规则挖掘发现大量数据中项集之间有趣的关联或相关联系。设 $I=\{i_1,i_2,\cdots,i_m\}$ 是表示 n 个不同数据项组成的项集。设任务相关的数据 D 是事务集,对于 D 中的每一个事务 T,有 $T\subseteq I$,也就是 T 是 I 中项的集合。设 A 是一个项集,事务 T 包含 A 当且仅当 $A\subseteq T$。关联规则是形如 $A\Rightarrow B$ 的蕴涵式,其中 $A\subset I$,$B\subset I$,并且 $A\cap B=\varnothing$。规则 $A\Rightarrow B$ 在事务集 D 中成立,具有支持度 s 和置信度 c,定义为:

$$s(A\Rightarrow B)=P(A\cup B);\ c(A\Rightarrow B)=P(B|A)$$

同时满足最小支持度阈值 s_{min} 和最小置信度阈值 c_{min} 的规则称为强规则,也就是我们从事务集 D 中挖掘出的感兴趣的规则。

包含 k 个项的项集称为 k-项集,设 k-项集 C 在事务集 D 中的出现频率为 $\sigma(C)=\|\{t|t\in D,C\subseteq t\}\|$,如果 $\sigma(C)\geqslant\|D\|\times s_{min}$,那么称 C 为频繁 k-项集。在频繁项集上可以产生满足最小支持度和最小置信度的强关联规则。

根据经验,在相当比例的文档中共同出现的关键词在语义上相近,我们把关键词作为事务项,将文档看做事务,就可以在一定的最小支持度阈值 s_{min} 和最小置信度阈值 c_{min} 下发现关键词之间强关联规则。我们可以认为强关联规则下的关键词在语义上相近,下面给出了强关联规则下的关键词的两个定义。

定义1(概念派生词) 设由中文文档库生成的关键词集为 $W=\{t_1,t_2,\cdots,t_n\}$,对于 $t_i,t_j\in W$,如果 $P(t_j|t_i)>c_{min}$,$P(t_i\cup t_j)>s_{min}$,并且 $(P(t_i)-\alpha)(P(t_j)-\alpha)\geqslant 0$($\alpha$ 为频率阈值,$P(t)>\alpha$ 说明关键词 t 为频繁词),那么我们定义 t_j 为 t_i 的概念派生词,记做 $t_i\rightarrow t_j$。

定义2(概念等价词) 如果对于 $t_i,t_j\in W$,如果同时满足 $t_i\rightarrow t_j$ 和 $t_j\rightarrow t_i$,那么我们定义 t_j 和 t_i 互为概念等价词,记做 $t_i\leftrightarrow t_j$。

基于上面给出的两个定义,我们给出概念集簇的定义如下:

定义3(概念集簇) 设关键词集 $E=\{t_1,t_2,\cdots,t_m,w_1,w_2,\cdots,w_k\}$,如果对于任意的 $1\leqslant i\leqslant m$,存在 $1\leqslant j\leqslant m(i\neq j)$ 满足 $t_i\leftrightarrow t_j$;对于任意的 $1\leqslant j\leqslant k$,存在 $1\leqslant i\leqslant m$ 满足 $t_i\rightarrow t_j$,那么关键词集 E 就称为概念集簇。

概念集簇也就是从大量文档中基于关键词的共现率分析来获得的关键词集。为了生成概念集簇,首先需要从文档-关键词事务矩阵 $\lambda_{m\times n}$ 中发现所有2-项强关联规则,这里我们采用关联规则挖掘的 Apriori 算法[3,4]来发现满足本文的最小支持度 s_{min} 和最小置信度 c_{min} 的2-项强关联规则集合 F,以发现满足最小支持度的关键词的所有概念派生词。然后从挖掘到的2-项强关联规则集合 F 中生成概念集簇,也就是原始概念集。Apriori 算法是很成熟的关联规则发现算法,在这里就不加更详细的说明,下面给出从2-项强关联规则也就是派生概念词对中生成概念集簇的算法。

算法1 生成概念集簇算法

```
输入:F:由λm×n生成的2-项强关联规则集合
输出:原始概念集 G={G1.G2,…,Gl},Gi(1≤i≤l)为概念集簇
功能:生成原始概念集
(1)从 F 中生成所有的具有概念派生词的关键词集合 H;
(2)done={};//done 表示所有已经根据其派生概念生成概念集簇的
            关键词集合
(3)i=0;
(4)for(each keywordH)
(5){
(6)       if (keyword not in done)
(7)       {
(8)              i=i+1;
(9)              Gi 为 keyword 的所有概念等价词和概念派生词以
                   及其概念等价
(10)             词的概念派生词的集合;
(11)             将 keyword 和其概念等价词加入到集合 done 中;
(12)      }
(13)}
```

生成 G_i 的时间复杂度为 $O(\|F\|)$,因此算法的时间复杂度为 $O(\|F\|\times\|H\|)$。

3.2 对原始概念集聚类

由于文档库中具有足够支持度并且具有概念派生词的关键词数量可能会比较大,因此,会导致原始概念集 G 中的概念集簇数目过大,这就需要对原始概念集簇中的概念集进行概念聚类,以生成更为简约的概念集。在这里我们采用 k-means 算法对概念集簇进行聚类[5,6]。首先利用 VSM 来表示每个原始概念集,向量的分量为布尔值,0表示概念集中不包含该关键词,1表示包含该关键词,我们可以根据实际情况设定聚类中心的个数 N,也就是结果概念集的数目来对原始概念集簇进行聚类,以达到了进一步对概念进行规约的目的。

3.3 概念集的增量更新

对生成概念集进行维护需要面对一个很重要的问题,就是概念集的增量更新,当然,每次对文档集进行更新后,我们

可以采用上面的过程来重新生成概念集，但这在效率上是无法接受的。

采用关联规则增量式更新算法 IUA(Incremental Updating Algorithm)[8]可以高效地实现关联规则的增量更新。对于增量更新产生的2-项强关联规则，也就是概念派生词对 $t_i \rightarrow t_j$，将关键词 t_j 加入由关键词 t_i 所生成的原始概念集以产生新的原始概念集，然后再对原始概念集簇进行增量式聚类[5,6]，最终可以实现概念集的增量更新。

4 实验过程及结果

在这里，我们采用"入世理论研究"文档集进行实验，该文档集包含4092篇中文文本文档和中文 Web 文档，过滤去停用词和高频无效词，计算每个文档中词条频率和位置的加权权重，每个文档取出权重最高的20个关键词，4092篇文档中共取出10936个关键词，然后生成关键词-文档事务矩阵 λ，该矩阵的行列为4092×10936。我们设定最小支持度 $s_{min}=0.3\%$，最小置信度 $c_{min}=30\%$，采用 Apriori 算法，共挖掘得到6471条2-项强关联规则，我们设定频繁阈值 $\alpha=25\%$，也就是对于关键词 A 和 B，如果具有强关联规则 $A \Rightarrow B$，并且 $P(A)<\alpha, P(B)>\alpha$，那么关键词 B 就不是关键词 A 的派生概念。利用我们的概念集簇生成算法，共生成599个概念集簇。对这599个概念集簇执行概念聚类，设定聚类中心为100个，得到的部分结果概念集如下：

表1 部分结果概念集

能源 煤炭 石油 原油 资源 价格 煤 油价
药物 开发 生物 新药 研究 药品 医药 中药 产业 基因 制药 西药 中成药 中医药 知识产权 药 药价 专利 保护 核心 推动
家电 空调 彩电 冰箱 电视机 洗衣机 购买 消费 调查 长虹 背投 康佳 创维 售后服务 销售 消费者 厂家 经销商 品牌
海信 方正 厂商 电脑 国产 联想 领域 手机 诺基亚 摩托罗拉 移动 爱立信 通信 无线 生产 品牌 价格
程序 软件 微软 硬件
数据 系统 传输 通信 用户 语音 互联网 网络 运营 运营商 交换 安全 收费 线 解决方案 客户 应用 上网 电子化 创新 门户网站
降价 车 价格 轿车 汽车 入世 消费者 老百姓 价 关税 国产 进口 工业 汽车业 购车 进口车 经济型轿车 进口轿车 客车 微型 下降 消费 到岸 费用
寿险 保险 保险业 外资 人寿保险 民族 保险费率 监管 开放 投保 险种 中资 对外开放 上市 市场准入 代理 代理人 经纪人 交易 经纪
养老 保险 保障 基金 社会 社会保险 职工 养老金 退休
企业家 股票 国有 经营者 所有者 经理 激励 期权 激励机制 股票期权 收益 报酬 股东 上市 持股 股权 业绩 企业经营者 股份 年薪 收入 职工
税负 所得税 征收 增殖税 财政 税率 税收 税制 优惠 增值税 税 税种 征管 地区 开发 消费税 纳税人 征税 纳税 税务
集成电路 元器件 芯片 电子
铁路 列车 旅客 运输 航空 航空业 民航 航线 飞机 维修 航运 物流 货物 集装箱

在实验中我们采用的 PC 机硬件配置为 PIII667，256M 内存，从关键词-文档事务矩阵中发现所有2-项强关联规则约为5s，从强关联规则生成概念集簇约800s，对概念集簇进行聚类约600s，因此本算法的效率还是比较令人满意的。

结束语 由实验结果可以看出，该算法能较好地发现中文概念集，并且具有较高的效率，同时由于可以引入增量更新，因此具有一定的实用价值。当然结果中有些概念词之间的语义联系并不符合经验，这是由于文档集的数量和质量的限制所造成的，下一步可以尝试对更大规模的文档集进行计算并引入模糊聚类。

参 考 文 献

1 史忠植. 知识发现. 清华大学出版社，2002

2 宫秀军，史忠植. 基于 Bayes 语义模型的半监督 Web 挖掘. 软件学报，2002，13(8)

3 Agrawal R，Imieliski T，Swami A. Mining association rules between sets of items in large database. In：Proc. of ACM SIGMOD Intl Conf on Management of Data(SIGMOD'93)，1993. 207～216

4 Han Jiawei，Kamber M. Data Mining Concepts and Techniques. Morgan Kaufmann，2001

5 Bradley P，Fayyad U，Reina C. Scaling clustering algorithms to large databases. KDD-98，，New York，Aug. 1998

6 Jain A K，Murty M N，Flynn P J. Data Clustering：A Review. ACM Computing Surveys，1999

7 刁倩，王永成，张惠惠. 基于神经网络的中文信息概念联想构造算法. 情报学报，2000

8 冯玉才，冯剑琳. 关联规则的增量式更新算法. 软件学报，1998，9(4)