一种基于 SVD 和 Rough 集的信息过滤方法*

陈彩云　李治国

南开大学组合数学研究中心，天津 300071

摘 要

本文提出了一种信息过滤方法，即在奇异值分解（SVD）的基础上，运用粗糙集（Rough Sets）理论进行信息过滤。通过对词语×文档矩阵进行奇异值分解得出近似矩阵，改变了一些词语在相应文档中的重要性，从而使得词语更好的体现文档内容。然后运用粗糙集理论中决策表上的规则推理方法，生成我们感兴趣信息的规则库，将未知文档的条件属性与规则库里规则进行相似匹配，进行信息过滤。实验表明，该方法在准确度方面比传统的 VSM 和 LSI 要好。

关键字： 奇异值分解　粗糙集　信息过滤 规则提取

1、引言

随着因特网上信息量的迅速增加，人们往往为了找到自己需要的信息花费大量的时间和精力，如何能够更有效的，更准确的找到自己感兴趣的信息，滤除与自己的需求无关的信息已经成为基于 Internet 网络信息处理的当务之急。随之产生的信息过滤技术正得到越来越广泛的关注，信息过滤系统根据用户的信息需求对动态信息流进行过滤，仅把用户感兴趣的文档传送给用户，可以提高获取信息的效率，对信息过滤主要的需求是对文档与用户信息需求相关性的判断要准确，同时查全率也需要提高。本文提出了一种信息过滤方法，在奇异值分解的基础上，运用粗糙集理论中规则推理方法，建立信息过滤的规则库，对于任意一个未知文档，我们只要将其条件属性与规则库中的规则进行相似匹配，进行过滤。实验证明该方法较传统的向量法和 LSI 方法都要好。

2、粗糙集相关理论

粗糙集是波兰 Z. Pawlak 教授提出的一种数据推理方法[1]。该理论为发现重要数据结构和复杂对象的分类提供了强有力的基础。我们首先描述与本文相关的粗糙集理论中的一些概念。（下面提到的概念和符号源自文献[2]）

2.1 信息系统（Information System）

信息系统由 4 元集组成，记为$\mathrm{S}=<U,\mathrm{Q},V,f>$，其中：

U：由 N 个研究对象$\{x_1,x_2,\ldots x_N\}$组成的非空集合，称为闭域（Closed Universe）；

Q：由 n 个属性$\{q_1,q_2,\ldots,q_n\}$组成的有限非空集合；

*本研究工作得到了教育部、科技部以及国家自然科学基金和国家 973 项目(项目编号 G19980306)的资助。

$V = \bigcup_{q \in Q} V_q$：表示 Q 中所有属性的值域，其中 V_q 是属性 $q \in Q$ 的值域。

$f : U \times Q \to V$：全决策函数（Total Decision Function），使得对于任一 $x \in U$，$q \in Q$，有 $f(x, q) \in V_q$。通过 f 作用，信息系统 S 能用一个有限的数据表表示，表的第 i 行研究对象 x_i 和第 j 列属性 q_j 有对应的值 $f(x_i, q_j)$。

2.2 决策表（Decision Tables）

如果信息系统 $S = < U, Q, V, f >$ 的属性集 Q 可以分成互不相交的条件属性集 C 和决策属性集 D，即满足 $C \cap D = \Phi$ 且 $C \cup D = Q$，满足这样条件的信息系统称为决策表，记 $DT = < U, C \cup D, V, f >$。一般情况下，集合 D 包含多个决策属性，但是在本文中根据研究的需要，我们只包含一个决策属性 d，即 D={d}。通过决策表，我们就可以对数据集进行规则推理。下面的过滤方法就是在决策表的基础上进行规则推理的。

3、奇异值分解（SVD）

给定 m×n 的矩阵 M，可以分解成三个矩阵的乘积 $M = U \times S \times V^T$，其中 U 和 V 分别为 $m \times m$ 和 $n \times n$ 的正交矩阵，S 为对角矩阵，S 的非零对角元 $\delta_i, (i = 1 \ldots r)$ 叫做矩阵 M 的奇异值，r 为非零对角元的个数。

定义 m×n 矩阵 $M_k = U_k \times S_k \times V_k^T$，其中 U_k 由 U 的前 k（$k \le r$）列列向量组成的 m×k 的矩阵，S_k 由 S 的前 k 个最大的奇异值组成的 k×k 的对角矩阵，V_k 由 V 的前 k 列列向量组成的 n×k 矩阵。由此构造的矩阵 M_k 是秩为 k 的矩阵中与 M 距离最近的矩阵，称之为秩为 k 的最好近似矩阵[3]。

4、构造信息过滤方法

第一步：准备数据，建立词语-文档矩阵（Term-Document）[4]M

首先我们收集一定数量的文档数据集。将之分成训练集和测试集，一般情况下，取所有文档的 60%-80%作为训练集，其它的作为测试集。假设有 m 个文档，选取 n 个关键词语，建立词语-文档矩阵 M，矩阵的每一行代表一个文档，每一列代表词语在文档中的出现的频率，即 $M=(m_{ij})$，m_{ij} 表示第 j 个词语在第 i 个文档中出现的频率。

第二步：将该矩阵 M 进行奇异值分解，构造秩为 k 的最好近似矩阵 M_k

我们将矩阵 M 进行奇异值分解，估计文档使用的词语结构。分解 M 得到

$\mathrm{M}=\mathrm{U}\times\mathrm{S}\times\mathrm{V}^{\mathrm{T}}$，再构造秩为 k 的最好近似矩阵 $M_k=U_k\times S_k\times V_k^T$，其中 $k\le r$，r 是非零奇异值的个数。通常情况下，我们面临的数据量是很大的，而使用奇异值分解，使我们找到了 M 的秩为 k 的最好近似矩阵 $\mathrm{M_k}$，从而降低了词语-文档的空间维数。

通过这样的变换，使得原来比较稀松的词语-文档矩阵变得稠密，改变了不同的词语在不同文档中的相对比重，从而使词语能更好的表达文档的内容。同样对于任何一篇新的文章，我们统计这 n 个关键词在该文章中出现的频率，得到 1×n 的向量 P，可以通过公式变换 $D_{\mathrm{p}}=\mathrm{P}\times V_k\times S_k^{-1}$，将 P 转化成词语-文档向量空间的向量的形式。

第三步：构造决策表 DT，生成决策规则

我们用上面预处理过的文档数据来构造决策表。$DT=<U,CUD,V,f>$表示一个决策表，其中闭域 U 是由词语-文档矩阵中 m 个文档$\{x_1,x_2,\ldots,x_m\}$组成，条件属性集 C 由词语-文档矩阵 M 的 n 个词语$\{t_1,t_2,\ldots,t_n\}$作为条件属性构成，决策属性集 D={d}由文档的类别属性构成。值域$\mathrm{V}=\{\mathrm{u_{ij}},1\le i\le m,1\le j\le n+1\}$，其中条件属性的取值我们直接取 M 的最好近似矩阵 $\mathrm{M_k}$ 的值，即 $f(x_i,t_j)=u_{ij}=m_{ij},1\le i\le m,1\le j\le n$，决策属性的取值根据我们感兴趣的或者是有价值的文档的属性决定，可以分别用文档的属性标明，比如军事，财经，体育等等，也可以直接用布尔变量 0，1 表示，0 表示是我们不感兴趣的文档，1 表示是我们感兴趣的文档，即$u_{i,n+1}=f(x_i,d)$的取值根据需求确定。

有此决策表，我们就可以用来进行规则推理：首先定义闭域 U 中任两个文档的相似度φ：

$$\varphi(u,v)=\frac{<u_n,v_n>}{|u_n|\cdot|v_n|}\qquad(1)$$

其中$u,v\in U$，u_n和$,v_n$分别为文档 u、v 对应的 n 个条件属性组成的向量，取定一个常数Φ_0为相似度阈值，如果$\varphi(u,v)\ge\Phi_0$，则认为 u 和 v 是相似的，否则是不相似的。任给 $x_i\in U$，记 $\mathrm{X_i}$ 为闭集 U 中与 x_i 相似文档（包括 x_i）的集合，即 $X_i=\{u\mid(\varphi(u,x_i)\ge\Phi_0,x_i,u\in U\}$。由 $X_i,i=1,2,\ldots,m$ 组成的集合记为 $C^*=\{X_1,X_2,\ldots,X_m\}$。根据决策属性 D 将 U 分成 g 类 $D^*=\{Y_1,Y_2,\ldots,Y_g\}$，决策表上的第 i 个决策规则定义如下：

$$\mathrm{Des_C(X_i)}\Rightarrow\mathrm{Des_D(Y_j)},\ \mathrm{X_i}\in\mathrm{C^*},\ \text{且}\ \mathrm{Y_j}\in\mathrm{D^*},$$

其中 $\mathrm{Des_C(X_i)}$ 和 $\mathrm{Des_D(Y_j)}$ 分别为 $\mathrm{X_i}$ 和 $\mathrm{Y_j}$ 的共性的唯一描述。例如 $\mathrm{Des_C(X_i)}$ 就是 $\mathrm{X_i}$ 的共性，在本文中即是与 x_i 相似度大于等于 Φ_0 的文档。

对于我们感兴趣的某一种类型的信息集 $Y_j \in D^*$ 的规则集表示如下：

$$\{\tau_{ij}\} = \{Des_C(X_i) \Rightarrow \mathrm{Dec}_D(Y_j) \mid X_i \cap Y_j \neq \phi, X_i \in C^*, Y_j \in D^*; i = 1,2,..., m\}$$

决策规则的准确率用公式 $\rho(\tau_{ij}) = \dfrac{\mathrm{card}(X_i \cap Y_j)}{\mathrm{card}(X_i)}$ 来计算，从 $\{\tau_{ij}\}$ 中选取准确率大于准确率阈值 α 的规则组成我们信息过滤的规则库 G，即 $G = \{\tau_{ij} \mid \rho(\tau_{ij}) > \alpha\}$。

决策规则可以用如果…那么语句来表示，即如果某个条件成立，那么就有某个结论成立。在本文中，决策规则我们可以表示为：对于一条新来的文档，如果该文档与规则库中某个规则相对应的文档的相似度大于等于阈值 Φ_0，那么该文档的决策属性就是该规则所对应的决策属性。举例来说，对于新来的文档，如果它与 x_2 的相似度大于 Φ_0，并且 $\tau_{2j} \in G$，那么该文档就是我们感兴趣的。

从上面的叙述中可以看到我们只需把训练集中与规则库对应的文档标识出来就可以了，对于新来的文档只需与这些文档进行相似度计算。用一个 $1 \times M$ 的向量 R 存储标识，R 的每一位只有 0 和 1 两个值。1 表示该文档与规则库的规则对应，0 表示没有规则与该文档对应。也就是说与标识为 1 的文档相似的文档就是我们感兴趣的，否则为我们不感兴趣的文档。

规则提取的算法可以如下描述：

步 1：选取类 D^* 中我们感兴趣的信息 Y_j，对决策表闭域 U 中每一个文档 x_i 重复执行**步** 2 至**步** 4。

步 2：依次将 x_i 与 U 中的所有文档 u（包括 x_i）计算相似度 $\varphi(u, x_i)$。

步 3：将满足 $\varphi(u, x_i) \geq \Phi_0$ 的所有文档组成集合 X_i，计算 $\rho(\tau_{ij})$。

步 4：如果 $\rho(\tau_{ij}) \geq \alpha$，则在 R 的相应的位置上面置 1，否则置 0。

第四步：推导任一篇未知文档的决策属性

对于任一未知文档，根据上面第二步计算出向量 D_p，再根据公式（1）分别计算该文档与标识为 1 的文档相似度，如果与某个文档的相似度大于等于 Φ_0，则说明此文档是我们感兴趣的，否则就是我们不感兴趣的。

下面我们用一个实验将该方法与传统的方法进行比较。

5、实验过程及结果分析

我们选择了新闻稿进行实验，选择体育、财经、军事三类新闻稿各约 1100 多篇（其中体育类 1197 篇、财经 1121 篇、军事 1190 篇共 3580 篇）作为实验数据（所有的新闻稿均取自千龙新闻网 www.21dnn.com）。将这些数据取 60%作为训练集，40%作为测试集。首先由

训练集得到训练规则，通过测试集进行测试，最后对 VSM、LSI、和粗糙集三种方法进行了比较，并对结果进行了分析。

5.1. 关键词的选取

我们首先搜集了所有的二元词共 98558 个，用部分新闻稿计算词频，去掉其中的虚词、高频词和低频词。从中选出最具有代表性的词 170 个。

5.2 建立词语-文档矩阵 M

计算选取的关键词在文档中出现的频率，从而构成词语-文档矩阵 $M=(m_{ij})$ ，m_{ij} 表示第 j 个词语在第 i 个文档中出现的频率。

5.3 生成规则库

根据上述方法中的第二步对 M 进行奇异值分解，取 k=20 构成新矩阵 M_k，并按上述方法中的第三步对分解好的矩阵进行规则提取，生成规则库。

5.4 实验结果比较

我们用三种方法：VSM、LSI、以及粗糙集方法进行多次实验，得到了图 1 所示的 P-R 图。

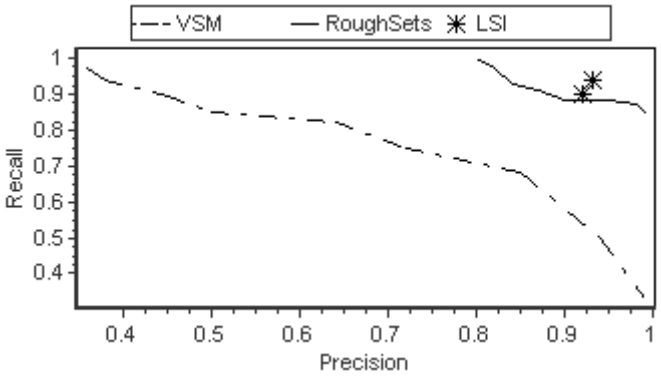


图 1 三种算法的 P-R 图

从图中可以看到，我们的方法在准确率和查全率上面比 VSM 都要好一些。准确率上也比 LSI 方法平均要好一些，尽管在查全率方面比 LSI 稍微差一点。

三种算法的时间复杂度与空间复杂度描述如下：

算法	训练算法	过滤过程	空间存储
VSM	O (m)	无	中间向量 V
LSI	$O(m\times r^2)$	$O(m^2)$	U_k,V_k,S_k
粗糙集方法	$O(m^2)+O(m\times r^2)$	O (m)	U_k,V_k,S_k,标识向量 R

表 1 三种方法的复杂度比较

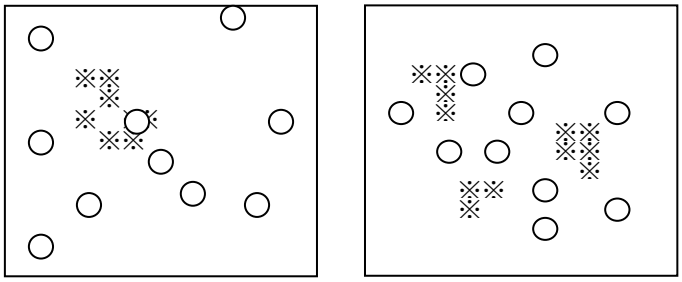
其中奇异值分解的复杂度是 $O(m\times r^2)$。从表 1 中我们可以看到：粗糙集方法虽然在训练过程中的时间复杂度比 LSI 要高，但是在过滤过程中却比 LSI 方法低了一个数量级。而且在空间存储方面也只比 LSI 方法多存储一个 1×M 的标识向量，并没有造成多少存储负担。所以在复杂度方面粗糙集方法还是优于 LSI 方法的。

5.5 原因分析

一个文档可以由一个向量来表示，图 2 表示了我们需要的和不需要的文档的向量的两种不同的分布情况，※是我们需要的文档向量，而〇是我们不需要的，我们要把※过滤出来。在 a 图的情况下,VSM 可以很容易地把※过滤出来,而且准确率会比较高,但是在 b 图的情况下,VSM 无论是准确率和查全率都不会很高，但是用粗糙集的方法或是 LSI 即使在这种情况下，也能达到很高的准确率。

由于 LSI 方法在进行过滤时只是进行相似性比较和排序，并没有过滤掉一些准确率不高

的向量，所以我们的结果比 LSI 在准确率方面要好一些。但是同样是这个原因，LSI 在查全率方面要优于我们的方法。



a b

图 2 常出现的两种情况

6、结束语

该文所提到的实验系统是在 Visual C++ 6.0 和 Delphi6.0 下,并借助 Matcom4.5 初步实现。我们初步探讨了奇异值分解和粗糙集理论相结合的一种信息过滤方法。运用代数的方法来重新调整词语-文档矩阵，然后运用粗糙集理论中的规则推理方法建立规则库，通过这种方法进行信息过滤，更能表达文档的内容，避免了传统的向量空间方法对信息过滤的盲目性，这无疑是对信息过滤的一种有益的尝试。而且在实验中我们验证了该方法无论在复杂度还是在准确度方面都是可以接受的，是一种切实可行的方法。

参考文献：

1、Pawlak. Z, Rough Sets, International Journal of Computer Sciences, 1982. 11, pp341-356.

2、Krzysztof J.Cios, Witold Pedrycz, Roman W. Swiniarski, Data Mining Methods for Knowledge Discovery, Kluwer Academic Publishers, 1998.

3、Azar.Y, Fiat. A, Karlin. A, etal, Spectral Analysis for Data Mining, Proceedings of the Thirty-Third Annual ACM Symposium on Theory of Computing, 2001, pp 619-626.

4、Christos H. Papadimitriou, Prabhakar Raghavan, Hisao Tamaki, etal, Latent Semantic Indexing: A Probabilistic Analysis, In Proceedings of ACM Symposium on Principles of Database Systems, 1997

A New Method for Information Filter Based on SVD and Rough Sets

Chen Cai-yun Li Zhi-guo

Center for Combinatorics of Nankai University, Tianjin 300071

Abstract

This paper proposed a new method for information filter based on Rough Sets theory and SVD. We have changed the importance of terms in corresponding documents by singular value decompose (SVD). Then we generated the rules which are useful to us base on the decision tables of Rough Set theory. When an unknown document was inputted, we just matched approximately the condition property of the document to these rules and remained useful information. The experiment proved that the method was better than traditional VSM and LSI in precision.

key words: SVD Rough Sets Information Filter Rule Generation

作者简介：

陈彩云，女，1975 年 11 月生，南开大学组合数学研究中心 2001 级博士生，研究方向 组合数学与数据挖掘，chencaiyun@eyou.com；

李治国，男，1977 年 9 月生，南开大学组合数学研究中心 2000 级硕士生，研究方向 组合数学与应用，sbickle@eyou.com；