基于模糊近似度的Web文本过滤模型*)

The Feature Acquiring Algorithm on The Web Text

刘明吉　饶一梅　王秀峰　黄亚楼

（南开大学信息技术科学学院　天津300071）

Abstract The booming growth of the Internet provides us a great deal of information resource. In this paper, we create a text filtering model based on VSM. In this model, Web text mining is an efficient technique, which discoveres valuable and potential knowledge from those unstructured texts. In this paper, we use VSM as the description of Web text and give a feature subset algorithm which is based on the Genetic Algorthm. This algorithm can greatly improve the efficiency of dealing with Web texts and give much better way to classify and cluster the texts. Our experiments show that this method is active well in feature dimension reduction.

Keywords VSM, Text filtering, Genetic algorithm, Text mining, KDD

从1991年诞生以来，WWW（World Wide Web）得到了迅猛的发展，它已经成为拥有约3亿用户、400万站点的巨大分布式信息空间。它包含了技术资料、商业信息、新闻报道、娱乐信息等多种类别和形式的信息，资源分布很分散，且没有统一的管理和结构，如何快速、准确地从浩瀚的信息资源中提取用户所需要的信息已经成为一个新的研究课题。WWW上最多的就是文本信息，因此Web信息处理的核心就是如何处理这些Web文档。

数据挖掘和知识发现（Data Mining and Knowledge Discovery，DMKD）可以帮助人们从大量原始数据中挖掘出隐含的、有用的尚未发现的信息和知识，有效地解决信息丰富知识贫乏问题。因此，基于Web文本信息的挖掘作为数据挖掘的一个新主题，引起了人们的极大兴趣。Web文本信息的挖掘就是在大量训练样本的基础上，得到文本数据间的内在特征，并以此为依据在网络资源中进行有目的的信息提取。

在本文中，我们首先介绍了Web文本信息的向量空间表示模型（VSM），并在此模型的基础上提出了一个文本过滤模型，并就该模型中的特征向量获取和模糊向量匹配等一些关键问题做了详细的描述和讨论，最后通过实验验证了该过滤模型的有效性和准确性。

1　文本的表示

Internet上的信息绝大部分以Web文档的形式出现。如何表示这种半结构化和无结构的文本数据类型，使其易于被计算机处理，是Web挖掘必须面临的最基本的前期工作。目前这方面的研究工作已经取得了一定的进展。60年代末由Gerard Salton等人提出的向量空间模型（Vector Space Model，VSM）是近几年来应用较多且效果较好的方法之一。下面简单介绍一下VSM模型的基本概念。

定义1（文档，Document）　泛指一般的文献或文献中的片段（段落、句子组或句子），一般指一篇文章。

定义2（项，Term）　当文档的内容被简单地看成是它含有的基本语言单位（字、词、词组、或短语等）所组成的集合时，这些基本的语言单位统称为项，即文档可以用项集（Term List）表示为$D(T_1,T_2,\cdots,T_n)$，其中T_k是项，$1\leqslant k\leqslant n$。

定义3（项的权重，Term Weight）　对于含有n个项的文档$D(T_1,T_2,\cdots,T_n)$，项T_k常常被赋予一定的权重W_k，表示它们在文档中的重要程度，即$D(T_1,W_1;T_2,W_2;\cdots;T_n,W_n)$，简记为$D=D(W_1,W_2,\cdots,W_n)$。

定义4（向量空间模型，VSM）　给定一文档$D=D(T_1,W_1;T_2,W_2;\cdots;T_n,W_n)$，由于$T_k$在文档中既可以重复出现又应该有先后次序的关系，分析起来仍有一定的难度。为了简化分析，可以暂不考虑T_k在文档中的先后顺序并要求T_k互异（即没有重复）。这时可以把

*）天津自然科学基金项目（003700111）和（993600811）资助。**刘明吉**　博士研究生，研究领域为金融信息系统、数据仓库和数据挖掘。**饶一梅**　博士研究生，研究领域为计算智能技术。**王秀峰**　博导，研究领域为遗传算法、计算智能技术。**黄亚楼**　博导，研究领域为机器人、数据挖掘。

$T_1, T_2, \cdots, T_n$ 看成一个 n 维的坐标系，而 $W_1, W_2, \cdots, W_n$ 为相应的坐标值，因而 $D(W_1, W_2, \cdots, W_n)$ 被看成 n 维空间中一个向量（如图1中的 D_1 和 D_2）。我们称 $D(W_1, W_2, \cdots, W_n)$ 为文档 D 的向量表示。

定义5（相似度，similarity） 两个文档 D_1 和 D_2 之间的（内容）相关程度常常用它们之间的相似度 $Sim(D_1, D_2)$ 来度量。当文档表示为 VSM，我们可以借助于向量之间的某种距离来表示文档间的相似度，常用向量之间的内积来计算：

$$Sim(D_1, D_2)=\sum_{k=1}^{n} w_{1k}\times w_{2k}$$

或用夹角余弦来表示：

$$Sim(D_1, D_2)=\cos(\theta)=\left(\sum_{k=1}^{n} w_{1k}\times w_{2k}\right)\Big/\sqrt{\sum_{k=1}^{n} w_{1k}^2}\times\sqrt{\sum_{k=1}^{n} w_{2k}^2}$$

其中 w_{1k}, w_{2k} 为向量 D_1, D_2 中的元素。

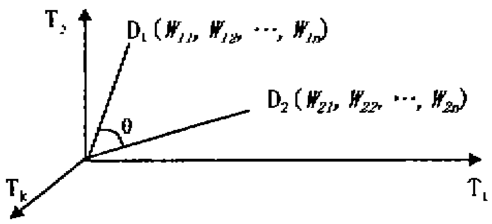

图1 文档的向量空间模型（VSM）及文档间的相似度 $Sim(D_1, D_2)$

在该模型中，空间文档被看作由一组正交词条所张成的矢量空间，每个文档 d 可以由一些规范化矢量 $V(d)=(t_1, w_1(d); \cdots; t_i w_i(d); \cdots; t_n w_n(d))$ 来表示，其中 t_i 为词条项，$w_i(d)$ 为 t_i 在 d 中的权值。而文本的一切特性则由它的特征向量来表示，文档之间的比较，实际上转变为特征向量的比较。

2 基于示例学习的文本过滤模型

我们的文本过滤模型是建立在文本的向量空间模型基础上的。我们对文本的学习和相似度计算就可以转变成对文本的特征向量的处理。文本过滤的处理任务基本上可以归结为以下两个部分：

（1）示例文本的学习和用户模板的获取

首先由用户提供一定数量的示例文本，作为系统学习的对象。系统根据这些示例文本获得用户的个性化查询要求和需要，并把这些特征词或词组构成反应用户需求的特征向量（即用户模板），这些特征向量就是进行文本过滤时的判别标准。

（2）实例文本的匹配和过滤

进行实例文本的匹配过程中，首先获取反映实例文本的特征向量，然后把该向量与用户兴趣模板进行匹配，满足一定阈值条件的文本就是符合用户需要的文本，而对于不满足一定阈值条件的文本就过滤掉。

文本过滤模型的主要设计思想是首先根据用户提交的示例文本，通过机器学习建立用户需求模型，然后在相应的文本集中搜索符合用户需求模型的文本。其模型如图2所示。

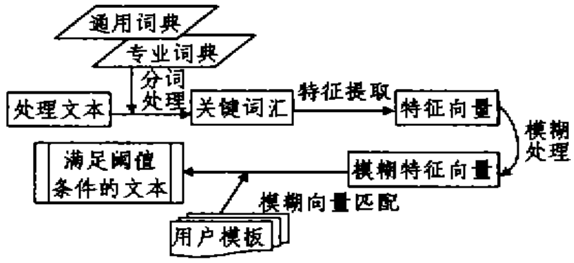

图2 基于示例学习的文本过滤模型

3 过滤模型中的几个关键技术

（1）文本特征向量的获取

文本过滤模型的核心问题就是反映文本内容的特征向量的获取。作为文本的特征向量应该具有彻底性（Exhaustivity）和专门性（Speciality）。其中彻底性是指文本所讨论的内容被特征向量覆盖的程度；专门性是指特征向量必须能反应文本的具体内容。

目前文本特征矢量的获取一般是通过一些分词算法和词频统计方法，从文档中选出尽可能多的词、词组和短语，由它们来构成文档矢量。但是由这种方法来表示文档，矢量的维数非常巨大。这种未经处理的文档矢量给后续的处理工作带来巨大的计算开销，使整个处理过程的效率非常低下。因此，我们必须对文档矢量做进一步精化处理，在保证原文含义的基础上，找出最能反映文本内容，又比较简洁的特征矢量。

我们认为，Web 文本实际上可以看作是由众多的特征词条构成的多维信息空间。特征矢量的选择实际上就是多维信息空间中的寻优过程。因此处理特征矢量的选择问题很自然地就想到使用高效的寻优算法。作为拟自然的一种通用搜索方法，遗传算法在复杂空间问题领域中表现出来的良好性能使它很自然地被用到了文本特征矢量的学习领域。我们在遗传算法的基础上，提出了一个基于文本矢量空间的特征矢量启发式抽取算法。

遗传算法（Genetic Algorithm，GA）是一种集效率与效果于一身的优化搜索方法。它利用结构化的随机信息交换技术组合群体中各个结构中最好的生存因素，从而复制出最佳代码串，并使之一代一代地进化，

最终获得满意的优化结果。在这里,我们把 Web 文本特征向量的获取问题转化成为 Web 文本空间的寻优问题。首先对 Web 文本空间进行遗传编码,以文本向量构成染色体,通过选择、交叉、变异等遗传操作,不断搜索问题域空间,使其不断得到进化,逐步得到 Web 文本的最优特征向量。由于本问题的特殊性,我们对传统的遗传算法作了大量的改进操作,使其更符合问题和实际情况。详细的遗传编码和实现已有专文论述,请参考文[1]。

(2)语义的模糊处理

文本的内容实际是人类自然语言的书面表达形式,它反映了为类的思想、认识和活动等各个方面。自然是属于人的认识范畴,其对象就是模糊的、不明确的。语言中的词、句、段之间的相互联系也是模糊的。句子或者存在模糊性或者存在歧义性。所谓语言信息处理,是指用计算机对自然语言的形、音、义等信息进行处理,即对字、词、句、篇章的输入、输出、识别、分析、理解、生成等进行操作和加工。由于人们对事物的描述有多种方式,这主要来自看待事物的角度不同,即同一概念具有多种表达形式。另外,由于在文章撰写时修辞的缘故,对于同一概念,为了避免用词重复,常常出现同义替代现象。这些处理必须建立在模糊语义的基础上。我们这里要讨论的文本过滤问题,也是在对语义篇章理解的基础上,根据文本的各项属性,来判断与判别标准是否一致或者说存在较大的相似度,因此在文本过滤处理过程中,必须处理好文本特征与分类标准之间的模糊匹配问题,而用精确的理论和技术来处理包含模糊特性的事物,是不会得到所期望的效果的。

在本过滤模型中,我们采用分类多层关键词对照表来解决模糊信息的处理问题。我们建立了二个词典:中心关键词词典和同义词词典,其中主词典中的词条要求在含义上保持尽可能的相互独立,其结构如图3所示。

中心关键词词典

词条名	近义词数目	开始搜索标号
……		
同情	2	18
喜欢	5	20
……		

同义词词典

标号	词条名	近似度度量
	……	
18	怜悯	0.7
19	可怜	0.5
	……	

图3　分类多层词典结构

(3)模糊向量的匹配技术

在传统的模糊处理过程中,通常采用基于同一标准的模糊测度处理方式,这种方法处理比较简便,效率也比较高。但是,由于在文本向量模型的处理过程中,关键词的权重决定了关键词汇的主次地位。这样,对同类关键词再基于同一中心词进行模糊测度处理就会产生一定的偏差。因此,我们在这里采用中心词移动的策略来解决这个问题。

例如,我们在对获得的特征向量 $A(a_1,a_2,\cdots,a_i,\cdots,a_n)$和 $B(b_1,b_2,\cdots,b_i,\cdots,b_n)$进行匹配时,由于 a_i 和 b_i 是相对同一个中心词的模糊测度距离,如图4所示。

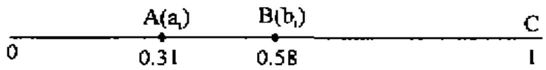

图4

此时,如果按照原来的办法,则我们将把 $a_i=0.31$ 和 $b_i=0.58$代入相似度计算公式。在本模型中,我们采用中心词移动策略来计算相对距离,也就是取 A(或 B)作为新的中心,修改 $a_i=0$(或 $b_i=0$)$b_i=|a_i-b_i|$(或 $b_i=|a_i-b_i|$),从而计算特征向量 $A'(a_1,a_2,\cdots,0,\cdots,a_n)$和 $B'(b_1,b_2,\cdots,|a_i-b_i|,\cdots,b_n)$的相似度,此时将获得更加精确的计算结果。

4　实验与评价

根据本文所提出的基于示例学习的 VSM 过滤模型,建立起 Web 文本过滤系统。该过滤系统在 Windows 98环境下用 Delphi 5.0实现。该系统分为训练学习和过滤处理两个处理阶段。在训练学习阶段,根据用户提供的示例学习文本,获取反应用户兴趣的用户模板。在过滤阶段根据待处理文本的特征向量与用户模板之间的模糊匹配程度,对文本进行过滤处理。

分词处理都采用基于通用词典和专业词典相结合的分词处理方法,生成关键词表,把文本表示成空间向量形式。然后通过基于 GA 的特征向量获取算法进行向量维数的压缩和约简,获取最能反映文本的特征向量。

在测试过程中,我们先提供一些示例文本,然后再把一些相关文本和不相关文本作为测试文本,让系统进行过滤处理。测试标准为查全率和准确率:

$$查全率=\frac{查出的相关文本数}{实际相关文本数}$$

$$准确率=\frac{查出的相关文本数}{实际测试文本数}$$

从实验结果可以看出,随着示例文本的不断增加,过滤模型获得了越来越准确的用户兴趣模型,从而使得过滤处理的效果越来越好。

测试情况	示例文本数	测试文本		准确率	查全率
		相关文本	不相关文本		
测试一	8	10	3	78%	86%
测试二	10	8	6	83%	85%
测试三	20	15	15	81%	88%
测试四	30	10	20	92%	85%
测试五	30	50	30	86%	89%
测试六	10	50	50	82%	85%
测试七	50	30	10	91%	93%
测试八	50	50	50	90%	91%

结束语 随着 Internet 的迅速普及，WWW 上的信息资源越来越丰富，更多的用户要求从网上获取信息。但是网络信息的快速膨胀，使得人们越来越难以获得符合自己要求的数据信息。因此，广大网民迫切需要一个具有自学习功能的信息过滤器，提高信息获取的效率和信息的质量。

在文章中，我们提出了一个基于示例学习的文本过滤器模型。在该模型中，我们采用改进遗传算法来获取 Web 文本的特征向量，从而有效地解决了传统文本处理过程中，文本向量维数过大的问题，提高了系统处理的效率。另外，我们基于语义的模糊特性，采用了模糊技术来表示特征向量，进一步提高了过滤模型的处理精度。从实验结果可以看出，该模型基本满足文本过滤处理的要求，具有令人满意的精度。

参考文献

1 Gudivada V N. Information retrieval on the World Wide Web. IEEE Internet Computing, 1997, 1(5): 58～68

2 Mladenic D. Machine Learning on non-homogeneous, distributed text data, doctoral dissertation. University of Ljubjana, 1998

3 Mladenic D, Grobelnik M. Feature selection for classification based on text hierarcy. Working Notes of learning from Text and the Web. Conf. on Automated learning and Discovery CONALD-98, 1998

4 张晓辉. WWW 上的信息发现与搜索引擎技术. 小型微型计算机系统, 1998, 19(6): 66～71

5 王实, 高文, 李锦涛. Web 数据挖掘. 计算机科学, 2000, 27(4): 28～31

6 张月杰, 姚天顺. 基于特征相关性的汉语文本自动分类模型的研究. 小型微型计算机系统, 1998, 19(8): 49～55

7 王伟强, 高文, 段立娟. Internet 上的文本数据挖掘. 计算机科学. 2000, 27(4): 32～36

8 吴立德, 等. 大规模中文文本处理. 复旦大学出版社, 1997

9 王继成, 潘金贵, 张福炎. Web 文本挖掘技术研究. 计算机研究与发展, 2000, 37(5): 513～520

10 邹涛, 王继成, 等. WWW 上的信息挖掘技术及实现. 计算机研究与发展, 2000, 36(8): 1019～1024

11 陈恩红. 基于遗传算法的机器学习方法及应用研究: [博士论文]. 中国科学技术大学, 1995

12 Liu Mingji, Wang Xiufeng. A Knowledge Discovery Algorithm Based on Genetic Algorithm. The Third World Congress on Intelligent Control and Automation, IEEE WCICA'2000

13 Victor Ciesie lski, Zhnag Memgjie. Using Genetic Algorithms to Improve the Accuracy of Object Detection. PAKDD-99, 19～24

14 Rudolph G. Convergence Properties of Canonical Genetic Algorithms. IEEE Trans on Neural Networks, 1994, 5(1): 96～101

(上接第39页)

部，教师})＝R({系主任})＝0.1，R(处长，系主任)＝0.2。

对任意取值{处长}，其广义上、下近似分别为{处长}和{{处长}∪{处长，系主任}}，故度量区间是(0.1，0.3)。其余可同理推出，不再赘述。

对于大型数据库，作用更加明显。

结论 用环境映射知识形成背景知识的过程本身又是一个系统学习的过程。通过对新环境的假设和对不同假定之间矛盾的处理，系统会提高归纳、类比等抽取知识的能力，并在运行过程中调整已有的关于环境间映射的知识，这对如何把握其它相似环境乃至更复杂环境提供了认知基础。

客观世界本质上的非全息性使得表示和处理都必须面对不确定性问题，本文给出了一种如何解决数据库中存在不精确信息问题的方法。

参考文献

1 Dubois D, Prade H. Possibility Theory: An Approach to Computerized Processing. New York: Plenum Press, 1988

2 Blum A L, Langley P. Selection of relevant features and examples in machine learning. Artificial Intelligence, 1997, 97: 245～271

3 洪家荣. 归纳学习. 科学出版社, 1997

4 Medin D L, Goldstone R L, Gentner D. Respects for Similarity. Psychol. Rev., 1993, 100: 254～278

5 孙吉贵, 刘瑞胜, 陈荣. 不完全信息下的溯因诊断. 吉林大学自然科学学报, '98, 10(4): 34～38

6 Chen A L P, Tseng F S C. Evaluating aggregate operations over imprecise data. IEEE Transactions on Knowledge and Data Engineering, 1996, 8: 273～284

7 Slowinski R, Vanderpooten D. A Generalized Definition of Rough Approximations Based on Similarity. IEEE Transactions on Knowledge and Data Engineering, 2000 12: 331～336

8 Bonikowski Z, et al. Extensions and intentions in the rough set theory. Information Sciences, 1998, 107: 149～167