Internet网络安全的信息过滤模型分析

张选芳

(中国民航飞行学院计算机与信息工程系 四川 广汉 618307)

【摘要】当前，由于Internet应用的逐渐普及，WWW已经成为一个巨大的分布式信息空间，为用户提供了一个极具价值的信息源。如何快速、准确得到确定领域中的信息，使信息过滤成为Internet网中关注的热点。基于此，该文分析了信息过滤中常用的向量空间模型、布尔逻辑模型和潜在语义索引三种信息过滤模型，以及存在的问题，在此基础上推出了一种准确度更好的模糊集合的信息过滤模型。

关 键 词 Internet网; 网络安全; 信息过滤; 模糊集合

中图分类号 TP393 文献标识码 A

Analysis of Information Filtrating Model of Internet Security

Zhang Xuanfang

(Department of Computer and Information Engineering, China Aviation Flight College Sichuan Guanghan 618307)

Abstract Nowadays, Internet is more and more widely used. WWW has been a largely distributed information space and provided a valuable information resource for users. How get the information rapidly and accurately makes information filtrating become the focus of Internet. Based on these, the paper analysis the vector space model, Boolean logic model and the latent semantic index model in common use in the information filtrating and problems in existence. And the paper presents an accurate and better information filtrating model based on the blur muster.

Key words Internet network; network security; information filtrating; blur muster

随着网络通信技术的发展和普及，在Internet网上涌现出的各种数据为各种用户提供了一个极具价值的信息源。但是，基于Internet网所固有的开放性、动态性和异构性，使用户很难准确快捷地从WWW上获取所需信息。这就需要根据用户的个性兴趣，在浩如烟海的动态信息中过滤掉无用信息，把所得到的不相关信息减至最小。目前，信息过滤成为当前重要的研究课题，其基本思想是从动态信息源中过滤掉比较固定的非需求信息，方法是通过代理服务器加入内容过滤功能，对内可以消除通过网页机密造成的泄漏；对外可以过滤掉网页中的无用信息。信息过滤要求过滤的内容性和实时性[1]，这两性指明了评价信息过滤模型优劣的标准是过滤精度和过滤速度。匹配算法的速度是决定信息过滤速度的因素之一，它通常由基本的过滤模型所决定，目前常用的基本信息过滤模型主要有：向量空间模型[2]，布尔逻辑模型[3]，潜在语义索引模型[4]。这三种模型的出现使信息过滤的查准率、查全率及效率都大有提高。然而，它们之间存在着各自的缺陷。为此，本文将在对三种模型比较的基础上讨论模糊集信息过滤模型。

1 三种信息过滤模型分析

1.1 向量空间模型

在向量空间模型构造的信息过滤系统中，习惯使用字项来标识文档。如：一个包含不健康信息的文档D，

收稿日期：2004 – 03 – 03
作者简介：张选芳(1950 –)，女，副教授，主要从事数据库应用和网络通信等方面的研究.

用一个m维向量来表示，其中m是能够用来表示文档内容的字项的总数。给每一个字项赋予一个权值，用来表明它的重要程度。该文档$\boldsymbol{D}$的向量表示为

$$\boldsymbol{D}=(w_1,w_2,\cdots,w_m) \tag{1}$$

式中　w_i表示第i个字项的权植。在进行信息过滤的过程中，首先对请求的页面数据进行加工，将其看成是一个由n个词组成的向量$\boldsymbol{P}$，然后比较向量$\boldsymbol{P}$和向量$\boldsymbol{D}$的相似程度

$$\mathrm{sim}(\boldsymbol{C},\boldsymbol{D})=\cos\boldsymbol{q}=\boldsymbol{C}\cdot\boldsymbol{D}\Big/\|\boldsymbol{C}\|\cdot\|\boldsymbol{D}\|=\sum_{i=1}^{n}u_i w_i\Big/\sqrt{\sum_{i=1}^{n}u_i^{\,2}\sum_{i=1}^{n}w_i^{\,2}} \tag{2}$$

由上式知，若两向量的夹角变小，则相似程度随余弦值而变大，待过滤文本符合过滤需求的可能性增加。设定过滤阀值$\boldsymbol{y}$，当$\mathrm{sim}(\boldsymbol{C},\boldsymbol{D})\geq\boldsymbol{y}$，其$\boldsymbol{D}$所对应的内容符合过滤需求，应禁止在网络中传输和扩散。

1.2　布尔逻辑模型

布尔逻辑模型是一种相对简单的模型，原理是给定一系列的具有二值逻辑的特征变量，其来源是从文档中抽取，用它们描述文档的特征。例如，在过滤过程中，它以文献中是否包含关键词来作为取舍的标准。最简单的关键词表可以设计成只有三个字段：关键词、包括关键词的文献号及关键词在相应文献中出现的次数。过滤时，提取请求的页面的关键词提交给系统，系统通过交集运算来判断是否要过滤掉该页面。

1.3　潜在语义索引模型

潜在语义索引模型已被广泛地应用到信息检索领域中，它是利用字项与文档对象之间的内在关系形成信息的语义结构。这种语义结构反映了数据间最主要的联系模式，忽略了个体文档对词的不同的使用风格。这是挖掘文档的潜在的语义内容，而不仅仅是使用关键字的匹配，是对字项文档矩阵使用奇异值分解(Singular-Value Decomposition，SVD)方法来实现的，把小的奇异值去掉。文献[5,6]中使用LSI技术对Netnews上的文档进行信息过滤，并就使用LSI技术与使用关键字匹配进行信息过滤的性能进行了比较。

对于奇异值分解来实现信息过滤的原理是给定一个字项文档矩阵$\boldsymbol{X}$，$\boldsymbol{X}$有r(表示文档集中关键字项的个数)行c(文档集中文档的数量)列。对$\boldsymbol{X}$进行奇异值分解得

$$\boldsymbol{X}=\boldsymbol{T}_0\boldsymbol{S}_0\boldsymbol{D}_0^{\,T} \tag{3}$$

式中　$\boldsymbol{T}_0$是$r\times m$矩阵，称其标准正交列为左奇异向量；$\boldsymbol{S}_0$是$m\times m$的对角阵，$\boldsymbol{S}_0$中的正奇异值是以递减的顺序排列的；$\boldsymbol{D}_0$是$m\times c$矩阵，$\boldsymbol{D}_0$的标准正交列可称为右奇异向量；m是矩阵$\boldsymbol{S}$的秩。对矩阵$\boldsymbol{T}_0$，$\boldsymbol{S}_0$和$\boldsymbol{D}_0$的处理是$\boldsymbol{X}$矩阵被重构。LSI技术的关键在于只取矩阵$\boldsymbol{S}_0$的k个奇异值，其他值置零。值k是一个设置参数，一般情况下经常设置在100～200之间。原始矩阵$\boldsymbol{X}$可近似表示为 $\hat{\boldsymbol{X}}=\boldsymbol{TSD}^T$，其中$\boldsymbol{T}$是具有标准正交列的$r\times k$矩阵，$\boldsymbol{S}$是一个$k\times k$的对角阵，$\boldsymbol{D}$也是具有标准正交列的$c\times k$矩阵。

无论是在LSI还是在关键字向量匹配方法中，文档都是以多维向量来表示的。关键字向量中的值表示字在文档中出现的频率。LSI向量中的值是通过SVD分解得到的缩减了的值。内容相近文档的向量也是相近的。这就是信息过滤中的本质所在。

1.4　分析

上述三种模型中布尔模型是通过对关键字集进行各种逻辑关系运算形成布尔表达式，根据所得到的布尔表达式的值1或0进行检索。这样检索的结果集经常是海量的，没有主次之分。向量空间模型是把每一个文本用一个字项权重的向量来表示，它只是一个数学描述，没有考虑到各种用户的实际情况和信息需求乃至文本的语义。潜在语义索引模型虽然较好地处理了信息的相关性，比较善于处理信息空间中具有大量因数的情况，但是并没有很好地解决信息分类这个问题。鉴于此，采用模糊集方法来进行信息过滤，可以大大地提高信息过滤系统的性能。

2　基于模糊集的信息过滤模型

自从模糊集合这一概念被提出之后，模糊集合在实际中得到广泛的应用。在经典集合概念中，每个元素相对于确定的集合而言可描述成“非此即彼”，而对于模糊集合，每个元素都有对应于该集合的一个隶属度，该元素与其集合的关系可用隶属度来表示。将模糊集合的概念应用到信息过滤中，更接近于对信息的抽象的理解。本文中，每个领域被定义为一个模糊集，领域关键字作为集合的元素，其隶属函数表示与该

领域的相关程度。同一领域关键字可跟不同的领域有或强或弱的相关性。

若一个模糊集合A代表某一领域，定义

$$A=\{(k_1,\ f_A(k_1)),(k_2\ f_A(k_2)),\cdots,(k_n,\ f_A(k_n))\} \tag{4}$$

式中 k_i代表一个关键字，$f_A(k_i)$代表k_i关键字对于该集合A的隶属函数。则模糊集合之间的运算为

$$\begin{cases} f_A\cap_B(k_i)=\min(f_A(k_i),\ f_A(k_i)) \\ f_A\cup_B(k_i)=\max(f_A(k_i),\ f_A(k_i)) \\ f'_A(k_i)=1-f_A(k_i) \end{cases} \tag{5}$$

式中 $\cap$即“与”，是集合间取同一元素的隶属函数的最小值；$\cup$即“或”，是集合间取同一元素的隶属函数的最大值；′即“非”，取1与原隶属函数的差值。

这里的信息关键词是指那些最能代表文本限制信息的词汇。其最大特点是在限制的信息范畴内频繁出现而很少出现在其他文本中。由于限制的信息通常只在其相关信息范畴内大量出现，而在信息范畴出现的概率很小，因此其检测也是利用这种爆发性。若一个词的爆发性较强，则是理想的信息关键词。

文档的主题信息存在于关键词分布矢量中，多个信息关键词在文章中的联合分布更包含了极为可靠的范畴信息，可以认为一篇文档与某一个主题的关联程度取决于文章的信息关键词分布矢量与该主题的核矢量之间的相似程度，即当两矢量完全重合时说明文章的内容完全符合该主题，如两矢量垂直，说明文档的内容与该主题无关，其相似程度取决于文档的信息关键词分布矢量与该主题的核矢量之间夹角的余弦值

$$F_j(k_i)=v(k_i)\cdot K(T_j)\Big/\left|v(k_i)\right|*\left|K(T_j)\right| \tag{6}$$

式中 $v(k_i)$表示关键词k_i的分布矢量；$K(T_j)$表示范畴为a_j的核矢量；$F_j(k_i)$是关键字k_i在代表范畴为a_j的模糊集合中隶属函数。过滤时根据所求得的余弦值是否大于事先所设定的门槛值，从而决定是否过滤掉该页面。

3 结 束 语

信息过滤系统的性能，关键在于过滤模型的完备程度。通过新文档的向量与过滤模型某一分文档的相近程度来判断词文档是否符合用户的兴趣。传统方法对语料库的标注是二值的，即一篇文档要么属于该范畴，要么不属于该范畴，而无法定量地表示文档与主题之间的联系强弱。而基于模糊集的方法用隶属函数描述文档与主题之间的联系强弱，与前三种信息过滤模型比较，则大大提高了信息过滤的查准率。

参 考 文 献

[1] 何 静，刘海燕，宫云战. TIS的WWW代理服务器中实现基于网络安全的内容过滤[J]. 计算机工程与应用, 2003, 39(20): 25-26

[2] Salton G, Edward A F, Wu H. Extended boolean information Retrieval[J]. Communications of the ACM, 1983, 26: 1 022-1 036

[3] Deerwester S, Dumais S T. Indexing by latent semantic analysis[J]. Jouranl of the American Society for Information Science, 1990, 41(6): 391-470

[4] Salton. Automatic text processing: the transformation, analysis and retrieval of information by computer[M]. Addison-Westey: Reading, Mass, 1989

[5] Foltz P W. Using latent semantic indexing for information filtering[C]. In: R B Allen Ed. Proceeding of the Conference on Office Information Systems, 2003 Cambridge, 2003MA: 40-47

[6] Maltz D A. Distributing information for collaborative filtering on usenet net news[EB/OL]. http://www.cs.cmu.edu/, 2001

编 辑 徐安玉