基于 VSM 的个性化信息过滤算法的研究

林美娜，苏 玉，张红艳

（烟台南山学院 计算机教学部，山东 烟台 265713）

摘 要：针对当前搜索引擎“所有用户，同一结果”模式的不足，分析了用户兴趣模型与文档的权值特征，在研究基于向量夹角余弦相关度排序算法的基础上，引入重要度因子，结合文档结构、查询请求及用户兴趣模型等信息，提出了一种基于 VSM 的个性化信息过滤算法，以实现个性化检索的目的，提高检索系统的查准率。

关键词：向量模型；个性化；信息过滤；相关度；信息检索

中图分类号：TP393 **文献标识码**：A **文章编号**：1674-7720(2012)21-0053-03

Research of personalized information filtering algorithm based on VSM

Lin Meina, Su Yu, Zhang Hongyan

(Department of Computer Teaching, Yantai Nanshan University, Yantai 265713, China)

Abstract: According to some weaknesses of "all users, the same results" in today's information retrieval, the paper analyzes the user interest profile and the document's weight feature. On the base of the study of the correlativity ranking algorithm based on vector angle cosine, it introduces the importance factor, and combines with the structure of the document, the query request and user interest model information, a new kind of personalized information filtering algorithm is proposed in this paper, which aims to realize personalized information retrieval and improves the system precision.

Key words: vector space model; personalization; information filtering; correlativity; information retrieval

由于 WWW 的不断丰富和日益庞大，用户通过搜索引擎获得的信息结果往往成千上万，而一般用户最多只浏览结果的前一百个 URL，后面的几乎可忽略不计。另外，目前基于关键字的搜索引擎技术，根据用户输入的查询请求关键字使用检索器进行检索，并将检索结果返回到客户浏览器，但没有考虑到具体用户的信息偏好，使得“所有用户，同一结果”。事实上，不同用户由于其领域背景和知识结构的不同，对文档相关性的判断是不同的。如何在信息检索中引入“个性化”机制，在结果返回浏览器之前，根据不同用户对信息结果进行相关度排序，过滤掉和用户相关度低的过量信息，检索出适量的、与用户兴趣相符的高质量的查询结果，并按相关性大小排序，提交给用户，是个性化搜索引擎要解决的一个重要问题[1]。

衡量文档的相关性有不同的方法和模型，当前最常用的模型有布尔模型、向量模型和概率模型[2]3 种。三者主要特点如表 1 所示。

表 1 3 种模型对比

	布尔模型	向量模型	概率模型
优点	基于集合理论和布尔代数的一种简单检索模型，形式化、简单化	可检索出与查询条件相近的信息，利用向量夹角余弦公式方便计算信息与查询关键词之间的相关度。	利用关键词间及其与文档间的概率相关性进行信息检索。从理论上，文档按照其相关概率的降序排列
缺点	文档要么匹配，要么不匹配，可能导致检索信息量过少，导致查全率低。	关键词被假设为相互独立的，忽略了各关键词之间及关键词与用户、关键词与文档之间的相关权值。	需要最初将文档分为相关和不相关的集合，对检索系统而言难实现。

基于表 1 的分析，在研究向量表示法的信息检索模型基础上，将检索到的文档和用户兴趣模型在同一空间里用向量表示，采用向量夹角余弦来计算相关度，并引入相关性权值概念，在一定程度上弥补了该模型关键词独立的不足。利用两向量间的夹角余弦来计算相关度，根据相关度大小对检索结果过滤、排序，以实现信息检

索的个性化，提高查准率。

1 个性化信息过滤算法

1.1 向量模型表示

向量模型要求将目标信息（用户兴趣及文档）在计算机内用空间向量来表示，即目标表示。它是指从目标信息中选择某些特征项，用这些特征项及在目标信息中的重要性（权值）来表示目标信息[3]。

（1）用户兴趣模型的 VSM 表示

设用户有 m 个兴趣，每个兴趣都用关键字来表示，二元组 (k_i,q_i) 表示用户对兴趣关键字 k_i 的兴趣权值为 q_i，则用户 p 的 VSM 模型可表示为：

$$\boldsymbol{Q}_p=\{(k_1,q_1),(k_2,q_2),\cdots,(k_m,q_m)\} \tag{1}$$

其中 m 为兴趣关键字集 $K=\{k_1,k_2,\cdots k_m\}$ 的大小。

（2）网页文档的 VSM 表示

WWW 缓存中的网页文档 D_j 可看作是一组关键字 $(k_1,k_2,\cdots k_n)$ 组成的集合，二元组 (k_i,t_{ij}) 表示关键字 k_i 在文档 d_j 中的出现频率为 t_{ij}。则其 VSM 模型可表示为：

$$\boldsymbol{D}_j=\{(k_1,t_{1j}),(k_2,t_{2j}),\cdots,(k_n,t_{nj})\} \tag{2}$$

n 为文档 d_j 的关键字集合 $K=\{k_1,k_2,\cdots k_n\}$ 的大小。所有 WWW 缓存中的网页文档向量可表示为：

$$\boldsymbol{D}=\{(k_1,d_1),(k_2,d_2),\cdots,(k_n,d_n)\}$$

1.2 基于 VSM 的相关度算法

在 VSM 模型中，相关度排序算法以两向量夹角余弦值来表示相关度[4]，由式（1）、（2）可得：

$$\mathrm{sim}(\boldsymbol{d}_j,\boldsymbol{Q})=\cos\theta=\frac{\sum_{i=1}^{n}t_{ij}\times q_i}{\sqrt{\left(\sum_{i=1}^{n}t_{ij}^2\right)\left(\sum_{i=1}^{m}q_i^2\right)}} \tag{3}$$

该算法只考虑了文档 D_j 与用户兴趣的相关度，忽略了查询请求中所有关键字在文档中的出现频率、出现位置（各级标题、是否为粗体、正文或其他位置）及关键字总长度占文档总长度的比例。查询请求关键字在文档 D_j 中出现的次数越多，位置越醒目占文档总长度的比例越大，说明文档 D_j 与用户的查询请求相关度越高。

通常文档信息包括各级标题、粗体文本、正文文本及其他信息，可以将 HTML 文档标记分为 Title、Head、Strong（粗体、斜体及下划线等）和 text（正文及其他信息）4 类，常规语义下标记不同，其重要度不同。为了表示不同位置关键词的重要程度，在算法中引入重要度因子向量 $\boldsymbol{CIV}(civ_1,civ_2,civ_3,civ_4,)$，设 $\boldsymbol{T}_{ij}$ 为关键字 k_i 在文档 D_j 中出现的频率向量 $(t_{ij1},t_{ij2},t_{ij3},t_{ij4})$，表示关键字在上述 4 类文档标记中的出现频率，则有：

$$\boldsymbol{T}_{ij}\cdot\boldsymbol{CIV}=\sum_{k=1}^{4}t_{ijk}\cdot civ_k \tag{4}$$

当 $\boldsymbol{CIV}=(1,1,1,1)$ 时，$\sum_{k=1}^{4}t_{ijk}\cdot civ_k=\sum_{k=1}^{4}t_{ijk}=t_{ij}$，一般取 $\boldsymbol{CIV}=(1,0.8,0.7,0.5)$[5]。

基于上述分析，相关度算法可修正为由两向量间夹角余弦值、查询请求关键字在文档 4 种类型中的出现频率、重要度因子、关键字长度与文档总长度之比三部分组成。

设文档 D_j 长度为 L_j，查询请求关键字集合 $S=(s_1,s_2,\cdots s_r)$，s_i 的长度为 L_i，在文档 D_j 的频率向量为 $\boldsymbol{ST}_{ij}(st_{ij1},st_{ij2},st_{ij3},st_{ij4})$，$st_{ij}=\sum_{k=1}^{4}st_{ijk}$，则所有查询请求关键字的出现权值频率之和为 $\sum_{i=1}^{r}\boldsymbol{ST}_{ij}\cdot\boldsymbol{CIV}=\sum_{i=1}^{r}\left(\sum_{k=1}^{4}st_{ijk}\cdot civ_k\right)$，总长度为 $\sum_{i=1}^{r}(st_{ij}\times L_i)$，得到改进的相关度算法 R_j 为：

$$R_{ij}=\frac{\sum_{i=1}^{n}t_{ij}\times q_i}{\sqrt{\left(\sum_{i=1}^{n}t_{ij}^2\right)\left(\sum_{i=1}^{m}q_i^2\right)}}\times\left(\sum_{i=1}^{r}\left(\sum_{k=1}^{4}st_{ijk}\cdot civ_k\right)\right)\times\frac{\sum_{k=1}^{r}(st_{kj}\times l_k)}{L_j} \tag{5}$$

在进行个性化信息过滤时，用户对某一关键字的兴趣度不变，即 $\sqrt{\left(\sum_{i=1}^{m}q_i^2\right)}$ 对最后排序不起作用，将其删除，式（5）化简为：

$$R_{ij}=\frac{\sum_{i=1}^{n}t_{ij}\times q_i}{\sqrt{\left(\sum_{i=1}^{n}t_{ij}^2\right)}}\times\left(\sum_{i=1}^{r}\left(\sum_{k=1}^{4}st_{ijk}\cdot civ_k\right)\right)\times\frac{\sum_{k=1}^{r}(st_{kj}\times l_k)}{L_j} \tag{6}$$

2 算法设计

根据改进的相关度算法式（6），基于 VSM 的个性化信息过滤器设计如图 1 所示。

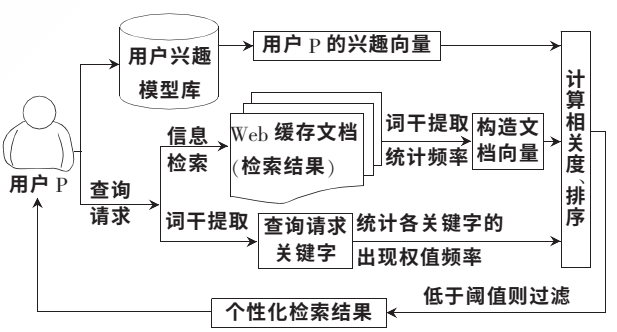

图 1 基于 VSM 的个性化信息过滤器结构图

（1）用户 P 的兴趣向量

为了减少计算量，可以根据用户 P 的兴趣向量进行兴趣提取。计算页面与用户查询请求相关度不需要考虑用户兴趣以外的关键词，并且只需提取某些权值较高的兴趣关键字，因为这些高权值关键字能够较稳定代表某用户的兴趣。选取数量 N 越大，系统检索速度越低。实

验验证，在正常的用户兴趣模型中，前 100 个关键字的权值之和一般能达到整个用户兴趣模型中全部关键词权值之和的 70%以上[6]。

(2)文档 D_j 向量

要实现将文档 D_j 进行向量化表示，首先需要利用分词技术对文档进行词干提取，得到文档 D_j 的关键字序列，针对(1)中所提取的用户兴趣模型中前 N 个兴趣关键字，统计出每个关键字在文档 D_j 中出现的频率 t_{ij}，构成文档 D_j 的 N 维向量$\{(k_1,t_{1j})、(k_2,t_{2j})\cdots(k_n,t_{nj})\}$。若某关键字未出现在文档 D_j 中，则 $t_{ij}=0(1\leqslant i\leqslant n)$。

(3)查询关键字的权值频率与总长度

对用户的查询请求进行词干提取，得到 m 个查询请求关键字，统计出该 m 个查询请求关键字的长度 l_k，基于重要度因子思想，统计在文档 D_j 中的不同位置(Title、Head、Strong、text)各查询关键词的出现频率 st_{ijk}，查询请求关键字 k 在文档 D_j 中的总长度为 $\sum_{k=1}^{4} st_{kj}\cdot l_k$。

(4)计算相关度、排序、过滤

在上述操作之后，利用式(6)计算出每个文档向量与某一用户查询请求的相关度，根据相关度由高到低进行排序，并与系统定义的阈值进行比较。若大于阈值，则保留该页面；否则判定为不相关，过滤此页。降低了信息量并提高查准率。

3 实验分析

已知用户 1 和用户 2 的社会角色分别为经济学家和房地产投资商，假设在此之前，系统已经通过学习建立起用户模型(为方便说明，取兴趣长度 $N=5$)，分别为：

Q_1={(经济，6.066548)、(发展，5.259445)、(产业，4.568227)、(房地产，2.358536)、(增长，1.158483)}

Q_2={(房地产，7.185453)、(销售，6.325842)、(开发，4.365851)、(投资，3.148545)、(增长，1.354228)}

图 1 和图 2 分别是对于同一查询请求“我国房地产现状”得到的检索结果中的两篇文章(节选)。针对文档 1、2 进行个性化信息过滤分析。

文档 1：我国房地产市场发展现状及其地位

(一)经济增长迅速刺激中国房地产市场的消费需求

改革开放以来，中国经济长期保持着两位数的高速增长，成为全球经济发展最为迅速的国家，在发展中国家中处于领先地位。经济快速增长，使得我国人民生活水平和社会购买力水平不断提高，社会对房地产的需求不断上升，原有的楼市供求矛盾加剧，最终给我国带来了更大的房地产市场消费需求。

(二)房地产业是经济发展的重要领域

经过长期的发展，我国房地产行业逐渐走向经济发展的前台，体现出产业化发展趋势。2006-2010 年连续五年的经济发展过程中，房地产吸引了众多行业领域的资本进入，这也是 2006-2010 年我国房地产市场价格不断攀升的重要资金缘由。在扣除外贸成分以后，房地产业是一个举足轻重的经济领域，是我国经济不可忽视的产业成分之一。

图 1 文档 1

文档 2：

2012 年 1-2 月份全国房地产开发和销售情况

一、房地产开发投资完成情况

2012 年 1-2 月份，全国房地产开发投资 5431 亿元，同比增长 27.8%；全国房地产开发企业房屋施工面积 394901 万平方米，同比增长 35.5%；全国房地产开发企业土地购置面积 4684 万平方米，同比下降 0.5%；土地成交价款 1044 亿元，增长 5.8%。

二、商品房销售和待售情况

2012 年 1-2 月份，全国商品房销售面积 7004 万平方米，同比下降 14.0%；其中，住宅销售面积下降 16.0%，办公楼销售面积下降 8.6%，商业营业用房销售面积增长 11.4%。商品房销售额 4145 亿元，下降 20.9%；其中，住宅销售额下降 24.7%，办公楼销售额下降 23.5%，商业营业用房销售额增长 17.2%。

2012 年 2 月末，全国商品房待售面积 30526 万平方米，比 2011 年年末增加 3332 万平方米。

图 2 文档 2

文档 1、2 的向量模型如表 2 所示。

表 2 文档 1、文档 2 的 VSM 表示

文档 1 的 VSM		文档 2 的 VSM	
关键词	权值	关键词	权值
经济	5.1	销售	5.8
房地产	5.6	下降	3.5
发展	5.3	房地产	3.3
市场	2.8	开发	4.0
产业	2.3	增长	2.5
增长	1.8	投资	1.3

根据式(6)，计算各自相关度数值得：

R_{11}=0.068 401　R_{12}=0.011 734

R_{21}=0.034 498　R_{22}=0.090 414

R_{ij} 为用户 i 针对该查询请求与文档 j 的相关度

从结果可以看出，对于“我国房地产现状”这一查询请求，用户 1 与文档 1 的相关度大于与文档 2 的相关度，因此文档 1 排在前，而对用户 2，与文档 2 的相关度大于与文档 1 的相关度，文档 2 排在前，较好地满足了个性化的需求。

基于 VSM 的个性化信息过滤算法，在分析了各种模型表示方法之后，采用 VSM 表示法将用户兴趣与文档统一表示，在基于向量夹角余弦的相关度排序算法基础上引入重要度因子，并进行相应改进，综合考虑文档结构、查询请求与文档的相关权值等因素，实现了为用户检索出真正关心的信息的目的，大大提高了系统查准率。

参考文献

[1] 王开选，张永奎.信息过滤中用户模型的表示方法[J].计算机工程，2006，32(5)：205-206.

[2] He Weihong, Cao Yi. An E-commerce recommended system based on content-based filtering [J]. Wuhan University Journal of Natural Sciences, 2006, 11(5):1091-1096.

[3] 曾春，邢晓春，周立柱.个性化服务技术综述[J].软件学

报,2002,13(10):1952-1961.

[4] 尚冬娟,张敏.信息过滤系统中的混合式过滤算法[J].重庆工学院学报(自然科学版),2008,22(1):118-121.

[5] 汪琴,安贺意,秦颖.网络信息过滤和个性化信息服务[J].情报科学,2007,6(25):858-863.

[6] 王春红,张敏,杨秀荣.基于 Web 的信息过滤系统的设计与实现[J].电子科技大学学报.2009,11(38):79-82.

(收稿日期:2012-04-12)

作者简介

林美娜,女,1979 年生,硕士,教师,主要研究方向:人工智能。

苏玉,女,1982 年生,硕士,教师,主要研究方向:计算机网络。

张红艳,女,1980 年生,硕士,教师,主要研究方向:现代教育技术。