文章编号：1000-5013(2016)02-0175-04 **doi**：10.11830/ISSN.1000-5013.2016.02.0175

采用向量空间模型的个性化信息检索方法

许建豪

（南宁职业技术学院 信息工程学院，广西 南宁 530008）

摘要： 为了提升检索结果与用户个性化需求的符合程度，依托向量空间模型提出一种新的检索方法. 将用户查询关键词和语料库内的文本信息都映射为向量，从而把检索过程转化为向量相似性的比对. 在比对过程中，通过关键词权重突出用户个性化需求，通过余弦相似度判断符合程度. 实验结果表明：文中方法的检索结果与用户需求的符合程度明显提高.

关键词： 信息检索；向量空间模型；个性化需求；语料库

中图分类号： TP 181 **文献标志码**： A

目前，中国的互联网用户已近 7 亿，占全国人口的 50%[1]. 人们对互联网日益依赖，需要从互联网上浏览和搜索各类信息. 如何使信息搜索结果尽可能臻善，已经成为各大互联网信息搜索引擎密切关注的重要课题[2]. 从目前的搜索引擎设置看，网络用户在信息搜索时，一般只能输入几个关键词. 但这些关键词并不一定能够准确地反映用户的兴趣和需求，加之很多搜索引擎就是通过词语匹配完成查找，更削弱了关键词丰富的自然语言特征，使检索到的信息结果差强人意[3]. 为此，信息检索领域的学者，致力于使搜索过程尽可能地符合用户的兴趣和需求，按照用户的个性化要求实现信息检索[4]. 国外学者在信息检索领域开展的研究工作较早，已具有比较丰富的研究成果[5-11]. 本文构建一个向量空间模型表达用户的个性化需求，并通过实验验证此方法的检索性能.

1 个性化检索方法设计

在信息检索方法设计中，为了使检索结果更符合用户的个性化需求，要求抽象的检索模型对用户检索需求有足够的理解能力. 基于此，文中选择向量空间模型作为构建个性化检索方法的基础模型.

1.1 向量空间模型

向量空间模型(VSM)将要检索的文本信息表征为向量空间上的向量，将文本检索的过程映射为向量运算，进而通过向量空间上的待检索文本向量和模板文本向量的相似性匹配获得最终的检索结果. 向量空间模型在文本信息检索中的应用，涉及到关键词、文件、相似性距离、向量空间模型等概念.

设向量空间是 m 维的，关键词是整个向量空间上的一部分，待检索文本信息用向量表示为 $\boldsymbol{T}_i(t_{i,1}, t_{i,2}, \cdots, t_{i,m})$，$t_{i,j}$ 为第 j 个词语的权重. $\boldsymbol{K}$ 为待查询的内容，其在向量空间的表示为 $\boldsymbol{K}(k_1, k_2, \cdots, k_m)$，$k_j$ 为查询中第 j 个词语的权重.

对于查询向量和文本信息向量之间的相似度计算，可以采取很多种方法. 文中采用两个向量之间的余弦夹角进行判断，即用余弦相似度方法判断两者之间的相似程度，即

$$\mathrm{Sim}(\boldsymbol{T}_i, \boldsymbol{K}) = \cos(\boldsymbol{T}_i, \boldsymbol{K}) = \sum_{j=1}^{m} t_{i,j} k_j \Big/ \sqrt{\sum_{j=1}^{m} t_{i,j}^2 \cdot \sum_{j=1}^{m} k_j^2}. \tag{1}$$

用 $f_{t,k}$ 表示 t_i 中关键词 k_i 出现的次数，则 k_i 在整个文本信息中出现的概率为

收稿日期： 2015-12-25

通信作者： 许建豪(1977-)，男，副教授，主要从事网络技术及信息检索的研究. E-mail：jianhaoxu@yeah.net.

基金项目： 广西高校科研基金资助项目(YB2014495)

$$p_{i,k}=\frac{f_{t,k}}{\sum_{j=1}^{m}f_{i,j}}. \tag{2}$$

为了便于对词频概率的使用,一般需要执行归一化处理,即

$$\hat{p}_{i,k}=\frac{p_{i,k}}{\sqrt{\sum_{i}p_{i,j}^{2}}}. \tag{3}$$

在向量空间模型中,还要考虑每个词汇在多少个文本中出现,其反映了一个词汇的区分度.区分度越低,表明这个词汇被使用的越广泛.对于这个特征,描述方法为

$$u=\log\frac{\mathrm{Num}}{f_{\mathrm{t}}}. \tag{4}$$

式(4)中:Num为此次查询中文本的总数;f_{t}为词汇出现的文本频率.

至此,可以根据空间向量的常见方法,计算关键词的权重,即

$$\theta_{i,j}=\frac{(\log\ p_{i,j}+1)-\log(\frac{\mathrm{Num}}{f_{\mathrm{t}}})}{\sqrt{\sum_{j=1}^{m}[(\log\ p_{i,j}+1)-\log(\frac{\mathrm{Num}}{f_{\mathrm{t}}})]^{2}}}. \tag{5}$$

式(5)中:$\theta_{i,j}$为关键词的权重;$p_{i,j}$为每个词语出现的词频;Num为此次查询中文本的总数;f_{t}为词汇出现的文本频率.

向量空间模型不仅可以实现查询要求和文本信息之间的匹配,还从词频、文频的角度增强关联性分析,具有反馈能力和一定的自然语言理解能力.

1.2 检索方法设计

为了使检索到的信息结果尽可能地满足用户的个性化需求,需要和用户进行反馈.基于向量空间模型的经典反馈查找最佳结果的方法为

$$\bar{\boldsymbol{R}}=\arg\max[\mathrm{sim}(\boldsymbol{R},T_{\mathrm{g}})-\mathrm{sim}(\boldsymbol{R},T_{\mathrm{n,g}})]. \tag{6}$$

式(6)中:$\bar{\boldsymbol{R}}$为最佳的查询结果;T_{g}为和用户个性化需求相关的文本集合;$T_{\mathrm{n,g}}$为和用户个性化需求不相关的文本集合.

式(6)为理论上的向量空间模型反馈查询方法,为了简化其在实际中的运用,改写为

$$\bar{\boldsymbol{R}}=\lambda_1\boldsymbol{R}_0+\lambda_2\frac{1}{|T_{\mathrm{g}}|}\sum_{t_j\in T_{\mathrm{g}}}\bar{t}_j-\lambda_3\frac{1}{|T_{\mathrm{n,g}}|}\sum_{t_j\in T_{\mathrm{n,g}}}\bar{t}_j. \tag{7}$$

式(7)中:$\boldsymbol{R}_0$为用户初始设置的个性化查询向量;λ_1,λ_2,λ_3分别为3个控制参数,以调整3部分之间的平衡,例如,经过反馈发现和用户检索需求不相关的文本数量更多,需要增大λ_2以维持平衡.

2 实验结果与分析

计算机硬件配置:酷睿双核、主频2.0 GHz的CPU,8 GB内存,500 GB硬盘.软件配置:Windows 7操作系统,Matlab程序设计语言及编译平台,Eviews统计分析软件.采用的文本信息检索对象为英国国家语料库(BNC).该语料库包含各种类型的文本信息子集,如经济领域、政治领域、军事领域、科技领域、生活领域等.

在文本信息检索的实验中,根据提出的基于向量空间模型的个性化检索方法,在BNC预料库中按照用户输入的关键词进行检索.因为很多关键词具有不同的领域特征,所以分别在一个领域和多个领域内搜索文本信息.科技领域内搜索文本信息的实验结果,如表1所示.表1中:A20,A30,A50,A1 000分别为检索结果中前20项,前30项,前50项和前1 000项的个性化符合程度的文献数.由表1可知:当$\lambda_1=30\%$,$\lambda_2=70\%$,$\lambda_3=1\%$的配置情况时,文中方法的检索效果达到最佳.

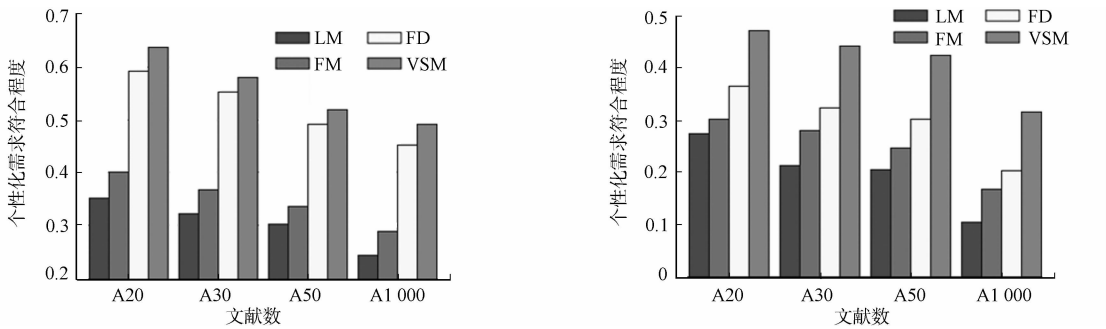
选择局部匹配检索法(LM)、全局匹配检索法(FM)、反馈检索法(FD)作为比较方法,在BNC预料库上开展个性化检索实验,4种方法的对比结果,如图1(a)所示.由图1(a)可知:LM方法检索结果和用户个性化需求的符合程度最低,文中方法检索结果和用户个性化需求的符合程度最高.

表 1　科技领域内文本信息的个性化检索结果

Tab. 1　Personalized search results of text information in the field of science and technology

序号	调整参数设置	个性化需求符合程度			
		A20	A30	A50	A1 000
1	$\lambda_1=10\%,\lambda_2=90\%,\lambda_3=1\%$	0.572 6	0.562 9	0.498 3	0.402 8
2	$\lambda_1=20\%,\lambda_2=80\%,\lambda_3=1\%$	0.573 9	0.568 0	0.501 6	0.407 3
3	$\lambda_1=20\%,\lambda_2=80\%,\lambda_3=5\%$	0.582 1	0.571 3	0.505 5	0.409 9
4	$\lambda_1=20\%,\lambda_2=80\%,\lambda_3=15\%$	0.596 1	0.572 2	0.508 3	0.412 6
5	$\lambda_1=20\%,\lambda_2=80\%,\lambda_3=25\%$	0.602 4	0.574 8	0.513 2	0.414 4
6	$\lambda_1=30\%,\lambda_2=70\%,\lambda_3=1\%$	0.637 7	0.580 1	0.518 9	0.490 8
7	$\lambda_1=40\%,\lambda_2=60\%,\lambda_3=1\%$	0.621 3	0.574 3	0.516 9	0.417 8
8	$\lambda_1=50\%,\lambda_2=50\%,\lambda_3=1\%$	0.600 9	0.573 6	0.508 8	0.413 9
9	$\lambda_1=60\%,\lambda_2=40\%,\lambda_3=1\%$	0.595 5	0.570 2	0.501 8	0.402 4
10	$\lambda_1=70\%,\lambda_2=30\%,\lambda_3=1\%$	0.583 7	0.565 4	0.493 7	0.401 6
11	$\lambda_1=80\%,\lambda_2=20\%,\lambda_3=1\%$	0.561 2	0.561 1	0.490 1	0.392 5
12	$\lambda_1=90\%,\lambda_2=10\%,\lambda_3=1\%$	0.554 5	0.552 4	0.485 5	0.388 4

在多个领域内搜索文本信息的结果，如表 2 所示. 由表 2 可知：当 $\lambda_1=20\%,\lambda_2=80\%,\lambda_3=25\%$ 的配置情况时，文中方法的检索效果达到最佳. 不同方法检索结果的比较，如图 1(b)所示. 由图 1(b)可知：文中方法检索结果和用户个性化需求的符合程度最高，且在多领域条件下，这种优势更加明显.



(a) 单一领域　　(b) 多个领域

图 1　4 种方法的对比结果

Fig. 1　Comparison results of 4 methods

表 2　多个领域内文本信息的个性化检索结果

Tab. 2　Personalized retrieval of text information in multiple domains

序号	调整参数设置	个性化需求符合程度			
		A20	A30	A50	A1 000
1	$\lambda_1=10\%,\lambda_2=90\%,\lambda_3=1\%$	0.453 6	0.432 4	0.417 0	0.302 9
2	$\lambda_1=20\%,\lambda_2=80\%,\lambda_3=1\%$	0.458 8	0.433 9	0.418 3	0.304 4
3	$\lambda_1=20\%,\lambda_2=80\%,\lambda_3=5\%$	0.462 9	0.435 1	0.419 2	0.307 8
4	$\lambda_1=20\%,\lambda_2=80\%,\lambda_3=15\%$	0.467 4	0.438 6	0.420 5	0.310 2
5	$\lambda_1=20\%,\lambda_2=80\%,\lambda_3=25\%$	0.470 9	0.442 6	0.423 3	0.314 3
6	$\lambda_1=30\%,\lambda_2=70\%,\lambda_3=1\%$	0.465 3	0.440 7	0.422 4	0.312 5
7	$\lambda_1=40\%,\lambda_2=60\%,\lambda_3=1\%$	0.461 2	0.435 4	0.417 8	0.310 6
8	$\lambda_1=50\%,\lambda_2=50\%,\lambda_3=1\%$	0.453 8	0.432 9	0.416 6	0.308 3
9	$\lambda_1=60\%,\lambda_2=40\%,\lambda_3=1\%$	0.451 1	0.428 1	0.411 7	0.302 5
10	$\lambda_1=70\%,\lambda_2=30\%,\lambda_3=1\%$	0.448 2	0.426 6	0.409 9	0.295 8
11	$\lambda_1=80\%,\lambda_2=20\%,\lambda_3=1\%$	0.445 3	0.421 3	0.405 8	0.294 3
12	$\lambda_1=90\%,\lambda_2=10\%,\lambda_3=1\%$	0.441 4	0.417 2	0.399 6	0.291 0

3　结束语

引入向量空间模型，将用户的个性化搜索需求抽象为向量，并结合关键词权重计算区分用户在不同个性化需求方向上的强弱，采取余弦相似度判别方法执行检索工作，再根据反馈查找思想提升检索结果

与用户检索需求的符合程度.在BNC预料库下的实验结果表明:无论是单一领域限制下的检索,还是多领域下的检索,文中方法的检索结果都更符合用户的个性化需求,明显优于LM,GM,FD等方法.

参考文献:

[1] 邹聪.浅析网络免费学术资源在医学信息检索教学中的有效应用[J].内蒙古科技与经济,2014,316(18):74-76.

[2] MARS B,HERON J,BIDDLE L,et al. Exposure to, and searching for, information about suicide and self-harm on the Internet: Prevalence and predictors in a population based cohort of young adults[J]. Journal of Affective Disorders,2015,185:239-245.

[3] 陈叶旺,余金山.一种改进的朴素贝叶斯文本分类方法[J].华侨大学学报(自然科学版),2011,32(4):401-404.

[4] DARABAD V P,VAKILIAN M,BLACKBURN T R. An efficient PD data mining method for power transformer defect models using SOM technique[J]. International Journal of Electrical Power and Energy Systems,2015,71(4):373-382.

[5] MADISON A,BUETTI S,LLEARS A. Singleton search performance predicts performance on heterogeneous displays: Evidence in support of the information theory of vision[J]. Journal of Vision,2015,15(12):12-14.

[6] MONCHAUX S,AMADIEU F,CHEVALIER A. Query strategies during information searching: Effects of prior domain knowledge and complexity of the information problems to be solved[J]. Information Processing and Management,2015,51(5):557-569.

[7] TANG Yuzhe,LIU Ling. Privacy preserving multi-keyword search in information networks[J]. IEEE Transactions on Knowledge and Data Engineering,2015,27(9):2424-2437.

[8] 邹向坤.基于Delphi的病历卡片信息检索系统的设计与实现[J].河北北方学院学报(自然科学版),2015,31(4):113-115.

[9] 陈秀丽.基于信息需求下电子商务档案信息检索的智能化研究[J].档案天地,2015(10):19-21.

[10] 甘丽新,万常选,王明文.基于层次依赖的Markov网络信息检索扩展模型[J].计算机科学与探索,2014,8(12):1485-1493.

[11] KUMAR A V,ALI R F M,CAO Yu. Application of data mining tools for classification of protein structural class from residue based averaged NMR chemical shifts[J]. Biochimica Et Biophysica Acta,2015,1854(10):1545-1552.

Research on Personalized Information Retrieval Method Using Vector Space Model

XU Jianhao

(School of Information Engineering, Nanning College for Vocational Technology, Nanning 530008, China)

Abstract: In order to improve matching degree between the retrieval results and of user's personalized needs, a new method based on vector space model is proposed in this paper. Maps the user query keywords and the text information in the database to the many vectors, and then transforms the retrieval process to the comparison of the vector similarity. In the process, the user's personalized needs are highlighted by the keyword weight, and the matching degree is determined by the cosine similarity. Experimental results show that the retrieval results of this method are significantly improved with the user's requirements.

Keywords: information retrieval; vector space model; personalized needs; corpus

(**责任编辑**:钱筠 **英文审校**:吴逢铁)