基于搜索结果的个性化推荐系统研究

卫 琳
(郑州大学 升达经贸管理学院,河南 郑州 451191)

摘 要:搜索引擎返回的信息太多且不能根据用户的兴趣提供检索结果,使得用户使用搜索引擎难以用简便的方式找到感兴趣的文档。个性化推荐是一种旨在减轻用户在信息检索方面负担的有效方法。文中把内容过滤技术和文档聚类技术相结合,实现了一个基于搜索结果的个性化推荐系统,以聚类的方法自动组织搜索结果,主动推荐用户感兴趣的文档。通过建立用户概率兴趣模型,对搜索结果STC聚类的基础上进行内容过滤。实验表明,概率模型比矢量空间模型更好地表达了用户的兴趣和变化。

关键词:搜索结果;聚类;个性化推荐;概率模型
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2007)09-0065-03

A Study of Personalization Recommendation System Based on Search Result

WEI Lin
(Shengda Economics Trade and Management College, Zhengzhou University, Zhengzhou 451191, China)

Abstract: It is difficult for the users to find interested documents in a simple and effective way by using search engines, since the returned information from search engines is too large and the engines are commonly hard to provide the users with required results based on their interests. Personalization recommendation is a valid method for lightening the user's burden on information retrieval. The paper presents a personalization recommendation system based on search result by combining content-based filtering technology and document-clustering technology, trying to recommend interested documents on their own initiative for the users by organizing search results automatically with the clustering method. By the founding of probability interest model of the users, the system realizes the content-based filtering of search results based on STC clustering. Experiments indicate that probability model can outperform VSM in expressing interests and changes of the users.

Key words: search result; clustering; personalization recommendation; probability model

0 引 言

Internet已成为人们获取信息的重要途径,由于Web信息的日益增长,人们不得不花费大量的时间去搜索浏览需要的信息。搜索引擎是最普遍的辅助人们检索信息的工具,但仍然不能满足不同背景、不同目的和不同时期的查询请求。搜索引擎搜索出来的内容真正被用户使用的,可能只有最前面很少的一部分,而用户真正感兴趣的内容却不能被找到。该问题的解决方案可以采用个性化服务技术[1]。通过收集和分析用户信息来学习用户的兴趣和行为,从而实现主动推荐的目的[2]。

对于搜索引擎返回的结果和它们本身所代表的原文档分别进行分类,结果发现误差为20%[3]。对于追求速度的搜索引擎来说,这是可以忍受的。但是由于搜索结果包含的信息较少,这对于个性化推荐是不利的,并且用户在使用系统时等待太长的时间是不合适的。文献[3]提出的后缀树聚类算法能够把搜索结果分离为具有较好内聚性的聚类,且该算法具有线性时间复杂度。文中将基于内容过滤的个性化推荐技术和后缀树算法相结合,实现了一个基于搜索结果的个性化推荐系统。利用领域分类模型上的概率分布表达了用户的兴趣模型,对搜索结果的聚类结果进行内容过滤,给出个性化推荐的文档。

1 文档和用户兴趣模型的表达

文档的传统表达方式是矢量空间模型,其缺点是

收稿日期:2006-11-29
基金项目:国家自然科学基金(60472044);河南省信息网络重点开放实验室基金(2006)
作者简介:卫 琳(1968-),女,河南郑州人,硕士,讲师,研究方向为Web挖掘、个性化推荐技术。

内容过滤时必须精确匹配文档,很难获得满意的结果。文中利用文档在不同领域中的概率分布来表达,其特点是避免文档间的精确匹配,提高了搜索的精度。同样地,可以利用用户兴趣在不同领域中的概率分布来表达用户兴趣模型。

1.1 矢量空间模型

表达文档和用户兴趣比较直接的做法是利用文档特征。用户兴趣是多方面的,可以根据其浏览过的文档选取合适的主题词来表达用户兴趣。用户兴趣可以表示为主题词的矢量 $\boldsymbol{u}=\{kw_1,kw_2,\cdots,kw_n\}$,其中 kw_i 表示第 i 个主题词出现的次数或权重。矢量的维数 n 一般是固定的,这样就保证了文档和用户兴趣之间相似性计算的精度。不过,预先定义好主题词表需要做大量的工作,其覆盖的范围有限。更简单的做法是直接利用从文档中抽取的词来表达用户兴趣[4]。

1.2 概率模型

建立一个领域分类模型,然后计算所有文档和用户兴趣在这个分类模型上的概率分布,用该概率分布来表达文档和用户兴趣就可以较好地体现用户兴趣的多样性。由于分类模型的类型个数远小于主题词的个数,这样,一方面提高了算法的运算速度,另一方面也提高了算法的推荐精度。

文中采用 Naive Bayes 方法来进行分类模型的训练[5]。假定领域类型的集合为 $C=\{c_1,c_2,\cdots,c_n\}$,其中 n 为模型的大小,c_j 表示第 j 个领域,则文档 $\boldsymbol{d}$ 表示为一个条件概率的矢量:$\boldsymbol{d}=\{p(c_1\mid\boldsymbol{d}),p(c_2\mid\boldsymbol{d}),\cdots,p(c_n\mid\boldsymbol{d})\}$,其中文档 $\boldsymbol{d}$ 对类型 c_j 的后验概率则为:

$$p(c_j\mid\boldsymbol{d})=\frac{p(\boldsymbol{d}\mid c_j)p(c_j)}{p(\boldsymbol{d})} \tag{1}$$

这里 $p(\boldsymbol{d})$ 表示为:

$$p(\boldsymbol{d})=\sum_{j=1}^{n}p(\boldsymbol{d}\mid c_j)p(c_j) \tag{2}$$

$p(c_j)$ 利用下式估算:

$$p(c_j)=\frac{c_j\text{ 中的文档数}}{\text{文档集中全部文档数}} \tag{3}$$

假定文档的所有特征都独立出现,则 $p(\boldsymbol{d}\mid c_j)$ 可以表示为文档所有特征条件概率的乘积:

$$p(\boldsymbol{d}\mid c_j)=\prod_{t\in\boldsymbol{d}}p(t\mid c_j) \tag{4}$$

假定 $n(c_j,t)$ 表示特征 t 在类 c_j 中出现的次数,$n(c_j)$ 为 c_j 中全部特征出现的次数之和,$\mid V\mid$ 表示文档集中全部不同特征的数目,则根据 Lidstone 连续定律,对一正数 λ(λ 一般取 0.5),$p(t\mid c_j)$ 的估计值可以表示为:

$$p(t\mid c_j)=\frac{n(c_j,t)+\lambda}{n(c_j)+\lambda\mid V\mid} \tag{5}$$

2 用户兴趣模型的更新

用户兴趣模型建立以后,可以通过跟踪用户的行为进行动态更新。用户的动作可以是添加书签、下载文档和删除书签等,这些动作体现用户不同的兴趣。文中采用机器学习中广泛使用的相关反馈方法来进行用户兴趣模型的更新。用户的信息根据下面的公式做相应刷新:

$$\boldsymbol{Q}_{i+1}=\alpha\boldsymbol{Q}_i+\beta\sum_{d\in R}\boldsymbol{v}_d-\gamma\sum_{d\in NR}\boldsymbol{v}_d \tag{6}$$

其中 $\boldsymbol{Q}_i$ 是初始的用户向量,而 $\boldsymbol{Q}_{i+1}$ 是经过修改的用户向量,$\boldsymbol{v}_d$ 是文档 $\boldsymbol{d}$ 的向量表示。α,β 和 γ 是已经设置好的反馈参数。R 和 NR 分别代表相关和不相关的文档集合。Rocchio 相关反馈方法是使用最广泛的,它的参数是这样设定的:$\alpha=1$, $\beta=2$, $\gamma=0.5$。

3 内容过滤和个性化推荐算法

3.1 搜索结果聚类

文献[6]中提出的 STC 聚类算法是搜索结果聚类的高效算法,它包括三个步骤:(1)文档清洗;(2)建立后缀树,获得基本聚类;(3)合并基本聚类。文中用到 STC 算法的前两步,当对搜索结果进行处理获得基本聚类后,把一个基本聚类中的所有搜索结果整体看作为一个文档,在这个基础上进行内容过滤。

3.2 相似性计算方法

对矢量空间模型来说,相似性计算的传统做法是计算矢量间的余弦相似度,用户 $\boldsymbol{u}$ 和文档 $\boldsymbol{d}$ 的相似性可以定义如下:

$$\mathrm{Sim}(\boldsymbol{u},\boldsymbol{d})=\frac{\boldsymbol{u}\cdot\boldsymbol{d}}{\|\boldsymbol{u}\|\cdot\|\boldsymbol{d}\|} \tag{7}$$

采用贝叶斯方法将相似度计算问题转化为求条件概率的问题。特定文档对于特定用户的条件概率越大表明它们的相近度也就越大,反之则越小。假定用户 $\boldsymbol{u}$ 在给定分类模型 $C=\{c_1,c_2,\cdots,c_n\}$ 时条件独立于文档 $\boldsymbol{d}$,则文档 $\boldsymbol{d}$ 推荐给用户 $\boldsymbol{u}$ 的概率可以表示为:

$$p(\boldsymbol{u}\mid\boldsymbol{d})=p(\boldsymbol{u})\sum_{j=1}^{n}\frac{p(c_j\mid\boldsymbol{u})p(c_j\mid\boldsymbol{d})}{p(c_j)} \tag{8}$$

3.3 个性化推荐算法

对搜索引擎产生的聚类结果按推荐概率进行排序,就能实现基于内容过滤的个性化推荐。

算法 1 基于内容过滤的个性化推荐算法。

Input:领域分类模型 P,用户兴趣模型 $\boldsymbol{u}$,查询关键词 q,一个搜索引擎 G

Output:个性化推荐的结果

Step1:根据查询关键词 q,利用搜索引擎 G 产生搜索结果集 X;

Step2:利用 STC 算法处理 X,获得基本聚类 Base

C,并把每个基本聚类 BaseC_j 转化为对应的文档$\boldsymbol{d}_i$,结果记为 X';

Step3:for (对每一个文档 $\boldsymbol{d}_i$){

利用式(1)计算其在领域分类模型 P 上的概率分布 $p(c_j \mid \boldsymbol{d}_i)$;

利用式(8)计算文档 $\boldsymbol{d}_i$ 推荐给当前用户的概率 $p(\boldsymbol{u} \mid \boldsymbol{d}_i)$;}

Step4:定义 M_i 为聚类 C_i 的质心,其值为 C_i 中文档的推荐概率的算术平均值;定义文档 $\boldsymbol{d}_j$ 关于聚类 C_i 的相关度 $S_{ji} = \mid p(\boldsymbol{u} \mid \boldsymbol{d}_j) - M_i \mid$, S_{ji} 值越小,表示 $\boldsymbol{d}_j$ 与 C_i 越相关;

Step5:用第一个文档 $\boldsymbol{d}_1$ 作为聚类 C_1;

Step6:在文档 $\boldsymbol{d}_j$ 和每一个聚类 C_i 之间计算相关度 S_{ji},并标记相关度 S_{ji} 的最小值 $S_{\min}$;

Step7:假如当前 $S_{\min}$ 小于极限值 $S_{\text{threshold}}$,把文档 $\boldsymbol{d}_j$ 加入到最相关的聚类;否则新建一个聚类,把文档 $\boldsymbol{d}_j$ 加入;

Step8:假如还有文档没有处理,重复执行 Step6 ~ Step8;

Step9:按聚类 C_i 的质心 M_i 值的大小从大到小排序;

Step10:把每一个聚类 C_i 中的全部文档对应的基本聚类进行合并,作为聚类 C_i 的最终结果;

Step11:输出第一个聚类的全部或部分文档作为个性化推荐结果。

4　实验结果及分析

基于以上算法,实现个性化推荐系统 SRPRS。SRPRS能够根据用户个人兴趣推荐聚类后的搜索结果,首先根据 Web 缓存模型和训练得到的领域分类模型,获得用户兴趣;其次对搜索结果用文中介绍的个性化推荐算法进行推荐。如图 1 所示。

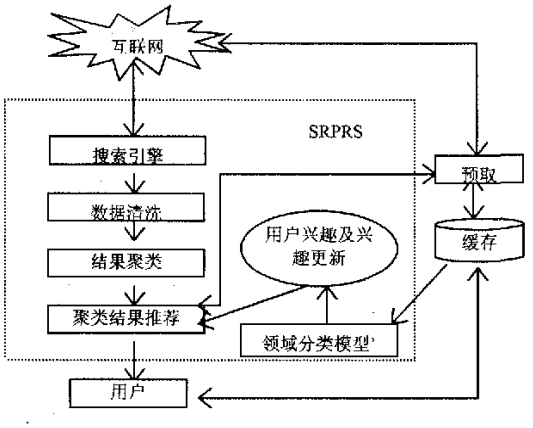

图 1　个性化推荐实验系统 SRPRS 体系结构

4.1　实验数据集

实验采用娱乐、财经、教育、军事、体育和科技的文档各 100 篇文章。搜索引擎采用 Google,实现了一个采用向量空间模型(VSM)来表达用户兴趣的个性化推荐系统。参与实验的人员分别在没用使用推荐系统、SRPRS 系统、VSM 推荐系统中各进行查询,并标记出其感兴趣的文档。

4.2　实验结果

(1)时间性能如图 2 所示。

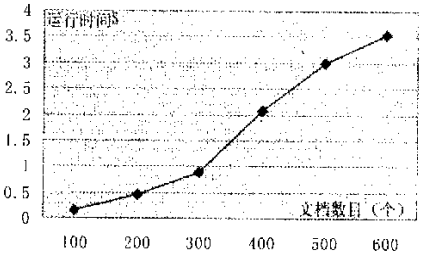

图 2　个性化推荐算法的时间性能

(2)查准率(精度)如图 3 所示。

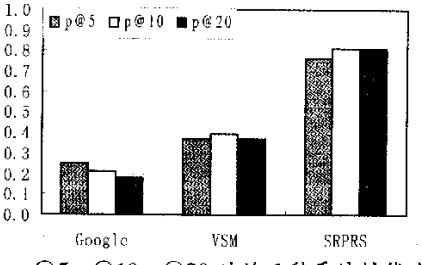

图 3　p@5,p@10,p@20 时的三种系统性能比较

(3)查全率(Recall)如图 4 所示。

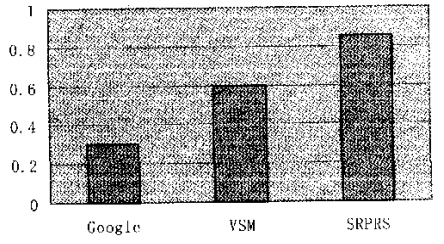

图 4　三种系统的查全率

图 2 中,个性化推荐的过程基本上满足线性增长。图 3,4 中,无论是查准率还是查全率,个性化推荐系统都具有较大的优越性。

对比矢量空间模型和概率模型所表达的用户兴趣模型对搜索算法的影响,概率模型的精度要大于矢量空间模型的精度,主要原因在于基于矢量空间模型的内容过滤需要进行精确匹配,而文档和用户兴趣之间相同关键词的个数一般都很少,所以会造成精度较低。概率模型则避免了这个问题。

(下转第 70 页)

元素形状的原因，会留下噪声区域的边缘，梯度图像边缘具有明显的方块效应，边缘较粗糙、定位不够准确、细节信息丢失较多(见图 1(b))。而改进算法的效果是明显的，相比之下，很好地抑制了噪声，同时也更多地保留了边缘信息，边缘精细(如眼睛、羽毛等细节处)，定位较准确(见图 1(d))。

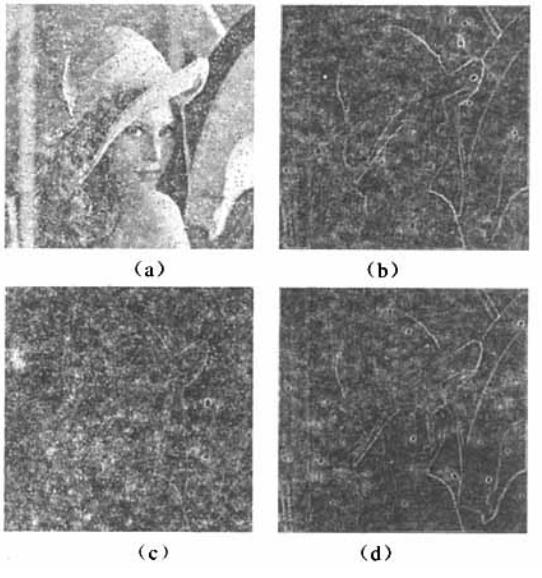

图 1 噪声图像及其各种形态学边缘检测图像

图中，(a)为加 5% 脉冲噪声的 lena；(b)为结构元素为 $\begin{bmatrix}1&1&1\\1&1&1\\1&1&1\end{bmatrix}$、抗噪型形态学差分算子(式 6)的检测(a)图的结果；(c)为 3×3 四个方向的多方位结构元素、传统形态学差分算子($f-f\ominus b$)的检测(a)图的结果；(d)为文中改进算法检测(a)图的梯度图像。

3 结 论

通过分析和实验，可以看到，抗噪型差分算子可以较好地在提取边缘的同时抑制噪声，但同时较多地丢失了明、暗的边缘信息；多方位结构元素通过构造多个方向的结构元素，增强了提取边缘的准确性和完整性，但对噪声更为敏感。改进算法将两种算法的优势相结合，同时使用小尺度结构元素对残留的噪声点去除，很好地解决了抑制脉冲噪声和精细提取边缘的矛盾，具有一定的实用性。

参考文献：

[1] 孙即祥. 图像分析[M]. 北京：科学出版社，2005.

[2] Schulze M A, Wu Qing X. Noise reduction in synthetic aperture radar imagery using a morphology－based nonlinear filter [C]//Proceedings of DICTA95, Digata Image Computing: Techniques and Applications. Brisbane, Australia: [s. n.], 1995:661－666.

[3] Wang Qin, Li Ziqin, Li Qi, et al. Anedge detection algorithm for imaging ladar[J]. Chinese optics letters, 2003, 1(5):272－274.

[4] 龚 炜，石青云，程明德. 数字空间中的数学形态学——理论以及应用[M]. 北京：科学出版社，1997.

[5] 李向吉，丁润淘. 脉冲噪声污染图像中的数学形态学边缘检测器[J]. 中国图像图形学报，1998，3(11)：903－906.

[6] 梁 勇，李天牧. 多方位形态学结构元素在图像边缘检测中的应用[J]. 云南大学学报：自然科学版，1999，21(5)：392－394.

[7] 付永庆，王咏胜. 一种基于数学形态学的灰度图像边缘检测算法[J]. 哈尔滨工程大学学报，2005，26(5)：685－687.

[8] Gonzalez R C, Wodds R E. 数字图像处理[M]. 阮秋琦等译. 北京：电子工业出版社，2003.

(上接第 67 页)

5 结 语

通用的搜索引擎不能满足不同背景、不同目的和不同时期的查询请求，文中提出了利用领域分类模型上的概率分布来表达用户的兴趣模型，与文档聚类技术相结合，主动推荐用户感兴趣的文档。与矢量空间模型相比，概率模型更好地表达了用户的兴趣和变化。

参考文献：

[1] Zeng C, Xing C X, Zhou L Z. A survey of personalization technology [J]. Journal of Software, 2002, 13(10):1952－1961.

[2] Shi Lei, Han Yingjie, Ding Xiaoguang et al. An SPN based Integrated Model for Web Prefetching and Caching[J]. Journal of Computer Science and Technology, 2006, 21(4):482－489.

[3] Zamir O, Etzioni O. Web Document Clustering: A Feasibility Demonstration[C]//In: Proc. of SIGIR'98. New York: ACM Press, 1998:46－54.

[4] Witten I H, Paynter G W, Frank E, et al. KEA: practical automatic keyphrase extraction[C]//In: Proc. of the 4th ACM Conf. on Digital Library. New York: ACM Press, 1999:254－255.

[5] Joachims T. A probabilistic analysis of the rocchio algorithm with TFIDF for text categorization[C]//In Proc. of the 14th Int. Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers, 1997:143－151.

[6] Zamir O, Etzioni O. Grouper: A dynamic clustering interface to web search results[J]. Computer Networks: The Int. Journal of Computer and Telecommunications Networking, 1999, 31(11－16):1361－1374.