基于内容过滤的个性化搜索算法*

曾 春+, 邢春晓, 周立柱

(清华大学 计算机科学与技术系,北京 100084)

A Personalized Search Algorithm by Using Content-Based Filtering

ZENG Chun+, XING Chun-Xiao, ZHOU Li-Zhu

(Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

+ Corresponding author: Phn: 86-10-62789150, Fax: 86-10-62771138, E-mail: bobofu00@mails.tsinghua.edu.cn

http://dbgroup.cs.tsinghua.edu.cn

Received 2002-10-21; Accepted 2002-12-04

Zeng C, Xing CX, Zhou LZ. A personalized search algorithm by using content-based filtering. *Journal of Software*, 2003,14(5):999~1004.

http://www.jos.org.cn/1000-9825/14/999.htm

Abstract: Traditional information retrieval technologies satisfy users' need to a great extent. However, for their all-purpose characteristics, they can not satisfy any query from the different background, with the different intention and at the different time. A personalized search algorithm by using content-based filtering is presented in this paper. The user model is represented as the probability distribution over the domain classification model. A method of computing similarity and a method of revising user model are provided. Compared with the vector space model, the probability model is more effective on describing a user's interests.

Key words: personalization; content-based filtering; search algorithm; user model; recommendation system

摘 要: 传统信息检索技术满足了人们一定的需要,但由于其通用的性质,仍然不能满足不同背景、不同目的和不同时期的查询请求.提出了一种基于内容过滤的个性化搜索算法.利用领域分类模型上的概率分布表达了用户的兴趣模型,然后给出了相似性计算和用户兴趣模型更新的方法.对比实验表明,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: 个性化;基于内容过滤;搜索算法;用户模型;推荐系统

中图法分类号: TP393 **文献标识码**: A

Web 已成为人们获取信息的一个重要途径,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自己需要的信息.搜索引擎是最普遍的辅助人们检索信息的工具,比如传统的搜索引擎 AltaVista(www.altavista.com),Yahoo!(www.yahoo.com)和新一代的搜索引擎 Google(www.google.com)等.信息检索技术满足了人们一定的需要,但由于其通用的性质,仍然不能满足不同背景、不同目的和不同时期的查询请

* Supported by the National Natural Science Foundation of China under Grant No.60221120146 (国家自然科学基金); the National Grand Fundamental Research 973 Program of China under Grant No.G1999032704 (国家重点基础研究发展规划(973))

第一作者简介: 曾春(1976－),男,江西萍乡人,博士生,主要研究领域为数字图书馆,个性化服务技术.

求.个性化服务技术就是针对这个问题而提出的,它为不同用户提供不同的服务, 以满足不同的需求.个性化服务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.个性化服务技术能够充分提高站点的服务质量和访问效率,以吸引更多的访问者.

目前存在着许多个性化服务系统[1,2],它们提出了各种思路来实现个性化服务.个性化服务系统根据其所采用的推荐技术可以分为两种:基于规则的系统和信息过滤系统.信息过滤系统又可分为基于内容过滤的系统和协作过滤系统.

基于规则的系统利用预定义的规则来过滤信息,其优点是简单、直接,缺点是规则质量很难保证,而且不能动态更新.此外,随着规则的数量增多,系统将变得越来越难以管理.基于内容过滤的系统利用资源和用户兴趣的相似性来过滤信息,它的关键问题是相似性计算,其优点是简单、有效,缺点是难以区分资源内容的品质和风格,而且不能为用户发现新的感兴趣的资源,只能发现和用户已有兴趣相似的资源.协作过滤系统利用用户之间的相似性来推荐信息,它能够为用户发现新的感兴趣的内容,其关键问题是用户聚类,其缺点是需要用户的参与.由于基于内容过滤和协作过滤各有其优缺点,所以有些系统同时采用了这两种技术.

本文提出了一种基于内容过滤的个性化搜索算法.基于内容过滤的基本问题包括用户兴趣的建模与更新以及相似性计算方法.本文利用领域分类模型上的概率分布表达了用户的兴趣模型,然后给出了相似性计算和用户兴趣模型更新的方法.对比实验表明,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.本文只关心文本资源,比如科技论文等,实际上,我们的方法还可以应用到其他领域.

本文第 1 节讨论文档和用户兴趣模型的表达.第 2 节讨论用户兴趣模型的更新.第 3 节描述相似性计算方法和基于该方法的个性化搜索算法.第 4 节描述实验系统并分析实验结果.第 5 节总结全文并进行展望.

1 文档和用户兴趣模型的表达

为了比较文档和用户兴趣,文档和用户兴趣模型的表达是一致的.文档的传统表达方式是矢量空间模型,其缺点是内容过滤时必须精确匹配文档,很难获得满意的结果.我们利用文档在不同领域中的概率分布来表达文档,其特点是避免文档间的精确匹配,从而极大地提高了搜索的精度.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[3].该方法需要一个训练的过程,首先从预定义好的主题词表中选取词来描述训练文档,为每个词都创建一个分类器,新文档将被每个分类器处理,对该文档有意义的词就赋予该文档.这样用户兴趣可以表示为一个主题词的矢量 $u=\langle kw_1,kw_2,\ldots,kw_n\rangle$,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,而且其覆盖的范围也有限,更简单的做法就是直接利用从文档中抽取的词来表达用户兴趣[4,5].该方法不局限于预定义好的主题词表,矢量的维数一般是不固定的,当然也可以指定一个固定的大小.这种方法因不能保证两个矢量之间存在很多相交的词,所以很难保证矢量相似性计算的精度.基于简单考虑,本文对比的就是这种方法.

1.2 概率模型

矢量空间模型只能表达用户感兴趣的主题词,不能很好地区别用户兴趣之间的差异.如果先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很好地体现用户兴趣的多样性,而且很容易实现.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搜索精度,因为用户在领域分类上更容易产生相似性.因此,概率模型比矢量空间模型能够更好地表达用户的兴趣和变化.

我们采用 Naïve Bayes 方法来进行分类模型的训练[6],这里我们讨论文档分类模型,用户兴趣和文档的表达是一致的.假定领域类型的集合为 $C=\{c_1,c_2,\ldots,c_n\}$,其中 n 为模型的大小,c_j 表示第 j 个领域,则文档 d 表示为一个

条件概率的矢量:$d=\langle p(c_1|d),p(c_2|d),\ldots,p(c_n|d)\rangle$,其中文档 d 对类型 c_j 的后验概率为

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

这里 $p(d)$表示为

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

$p(c_j)$用下式估计:

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

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

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

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,$|V|$表示文档集中全部不同特征的数目,则根据 Lidstone 连续定律(它克服了 Laplace 连续定律对数目较大的分类产生较大偏差的问题),对一正数λ(λ一般取 0.5,如果λ=1,则 Lidstone 定律与 Laplace 定律相同),$p(t|c_j)$的估计值可以表示为

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

2 用户兴趣模型的更新

用户兴趣模型建立以后,可以允许用户主动更新,也可以通过跟踪用户的行为进行动态更新.这里讨论的是后者,即根据用户当前的动作产生不同的更新.用户的动作可以是添加书签、下载文档、浏览摘要、忽略文档和删除书签等,这些动作体现用户不同的兴趣,所以具有不同的意义[7],见表 1.

Table 1 Meaning of user actions

表 1 用户动作的意义

User action	Meaning
Add a bookmark	Very high positive
Download a paper	High positive
View details of a paper	Moderate positive
Ignore a paper	Low negative or set to zero
Delete a bookmark	High negative

2.1 矢量空间模型

由于用户兴趣是用文档特征来表示的,所以当文档推荐给用户的时候,可以根据用户动作对应的文档来选取用户感兴趣的特征,并调整用户兴趣矢量中特征出现的次数或权重.假定用户 u 当前的动作为 a,其对应的意义为 w_a,用户动作对应的文档为 d,η是学习率,是一个小的常量,则利用下式来调整特征出现的次数或权重:

$$kw_i(u) \leftarrow kw_i(u)+\eta w_a kw_i(d). \tag{6}$$

2.2 概率模型

用户兴趣表示为领域分类模型上的概率分布,也就是一个条件概率的矢量,当文档推荐给用户的时候,可以根据用户动作对应的文档来修改矢量中对应每个分类的条件概率.首先计算文档 d 在分类模型上的概率分布,然后利用下式来修改用户兴趣矢量中对应每个分类的条件概率:

$$p(c_j \mid u) \leftarrow \frac{p(c_j \mid u)+\eta w_a p(c_j \mid d)}{1+\eta w_a}. \tag{7}$$

3 基于内容的过滤

在表示好文档和用户兴趣以后,可以利用文档和用户兴趣的相似性来过滤文档.本节介绍矢量空间模型和概率模型的相似性计算方法以及基于内容过滤的个性化搜索算法.

3.1 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(cosine similarity),用户 u 和文档 d 的相似性可以定义如下:

$$Sim(u,d)=\frac{u\cdot d}{\|u\|\cdot\|d\|}. \tag{8}$$

而对概率模型来说,直接计算矢量间的余弦相似度是不合适的,为了体现用户兴趣的多样性,我们提出了下面的命题[8].

命题 1. 假定用户 u 在给定分类模型 $C=\{c_1,c_2,\ldots,c_n\}$ 时条件独立于文档 d,则文档 d 推荐给用户 u 的概率可以表示为

$$p(u\mid d)=p(u)\sum_{j=1}^{n}\frac{p(c_j\mid u)p(c_j\mid d)}{p(c_j)}. \tag{9}$$

证明:由全概率公式可知,

$$p(u,d)=\sum_{j=1}^{n}p(u,d\mid c_j)p(c_j). \tag{10}$$

根据假定,用户 u 在给定分类模型 C 时条件独立于文档 d,所以有 $p(u|d,c_j)=p(u|c_j)$,进而得出 $p(u,d|c_j)=p(u|c_j)p(d|c_j)$,因此,式(10)可以变换为

$$p(u,d)=\sum_{j=1}^{n}p(u\mid c_j)p(d\mid c_j)p(c_j). \tag{11}$$

根据 $p(u|d)=p(u,d)/p(d)$,式(11)可以变换为

$$p(u\mid d)=\sum_{j=1}^{n}\frac{p(u\mid c_j)p(d\mid c_j)p(c_j)}{p(d)}. \tag{12}$$

由于 $p(u|c_j)p(c_j)=p(u)p(c_j|u)$,且 $p(d|c_j)/p(d)=p(c_j|d)/p(c_j)$,式(12)最后变换为式(9). □

根据命题 1 的结论,我们可以计算一篇文档推荐给用户的概率.其意义在于将概率模型的相似性计算问题转化为求条件概率的问题,体现了用户兴趣的多样性.

3.2 个性化搜索算法

根据命题 1 的结论,如果对一个搜索引擎产生的结果集按推荐概率进行重新排序,就能实现基于内容过滤的个性化搜索.值得注意的是,式(9)中的 $p(u)$ 是不用计算的,因为 $p(u)$ 不影响推荐概率之间的比较.下面是基于该方法的个性化搜索算法的详细描述.

算法 1. 基于内容过滤的个性化搜索算法.

输入:领域分类模型,用户兴趣模型,查询关键词,一个搜索引擎.

输出:个性化搜索的结果.

(1) 根据查询关键词,利用搜索引擎产生初步的搜索结果集 X.

(2) 置迭代次数 $i=0$.

(3) 对集合 X 中的第 i 篇文档,利用式(1)计算其在领域分类模型上的概率分布.

(4) 利用式(9)计算文档 i 推荐给当前用户的概率,加入列表 Y.

(5) 如果文档 i 是集合 X 中最后一篇文档,转(6);否则,置 $i=i+1$,返回(3).

(6) 根据列表 Y 中的概率由大到小排序文档并输出.

该算法实际上是基于另一个搜索引擎,所以对搜索结果中的每一篇文档都必须计算其在领域分类模型上的概率分布,这会极大地影响算法的性能.如果该搜索引擎能够预先计算好每一篇文档在领域分类模型上的概率分布,则算法的性能会得到很大的提高,从而满足实时处理的需要.

4 实验结果

本节介绍实验采用的数据集和实验评价标准,并进行实验分析.为了测试算法的性能,我们建立了一个个性化服务实验系统(http://dbgroup.cs.tsinghua.edu.cn/MyLibrary).在该系统中,我们实现了本文的算法,并通过跟踪用户的行为来学习用户的兴趣.

4.1 个性化服务实验系统

实验系统主要包括 4 个组成部分:浏览器插件、个人管理器、用户模型学习器和信息过滤器,如图 1 所示.浏览器插件主要是为用户提供一个便捷的工具,在用户配置好自己的登录信息之后,可以利用它直接实现个性化搜索而不必登录服务器.此外,浏览器插件还会主动收集用户信息并发送到服务器上.个人管理器的作用是为用户提供一个自我管理的平台,用户利用它可以管理自己的个人信息、个人兴趣和个人书签.用户模型学习器的作用是维护用户兴趣模型,它分析用户的信息和书签,并跟踪用户的行为来学习用户的兴趣.信息过滤器的作用是实现基于内容过滤的个性化搜索和推荐.图 2 是实验系统的一个快照.

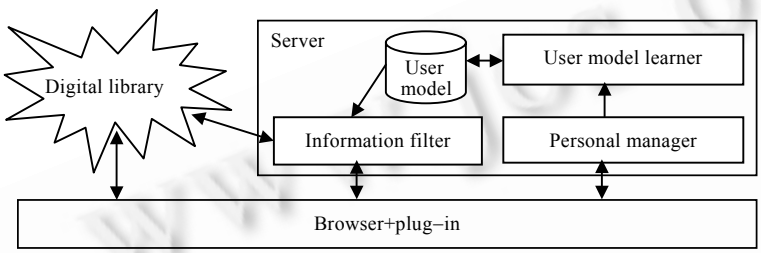

Fig.1 System architecture

图 1 系统体系结构

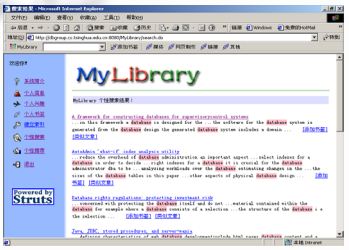

Fig.2 A snapshot of the system

图 2 实验系统的快照

与其他个性化服务系统的不同点在于:① 体系结构的不同.我们的系统分布在客户端和服务器端,可以跟踪用户在客户端的行为,也不影响用户的浏览和系统性能.② 用户兴趣模型的不同.我们用概率模型来表达用户的兴趣,并通过跟踪用户的行为来动态修改用户兴趣模型.

4.2 实验数据集

实验的数据集来自 INSPEC 科学文摘数据库,由于科学论文的关键词组和分类都比较明确,所以能获得比较清晰的结果.我们采用了 INSPEC 的分类体系,只选择计算机软件学科,分为 45 个类.我们从 INSPEC 数据库中选取了涉及计算机软件的 2 000 多篇论文摘要来训练领域分类模型,大小为 1.9MB.

在实验系统中,我们允许用户主动修改自己的兴趣,也通过跟踪用户的行为(比如添加书签、下载文档、浏览摘要、忽略文档和删除书签等)来动态修改用户的兴趣,然后根据用户的查询请求推荐与其兴趣相关的论文.

4.3 实验评价标准

我们采用信息检索领域广泛使用的查准率(precision)和召回率(recall)来评价实验结果.查准率和召回率的定义如下:

$$\text{Precision}=\frac{\text{搜索到的相关文档数}}{\text{搜索到的全部文档数}},\quad \text{Recall}=\frac{\text{搜索到的相关文档数}}{\text{系统全部相关文档数}}. \tag{13}$$

我们计算召回率为 0.2,0.4,0.6,0.8 和 1 时的查准率,平均精度定义为这 5 个点上的查准率的平均值.召回率为 0 时的精度是随意给定的,一般会稍微大于或等于召回率为 0.2 时的查准率.实验曲线类似于 ROC(receiver operating characteristic)曲线,曲线下的面积越大,说明算法的精度越高.

4.4 实验分析

我们对比了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.如图 3 所示,概率模型的

平均精度要远大于矢量空间模型的平均精度,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,所以会造成平均精度急剧下降.概率模型则避免了这个问题.它利用文档和用户兴趣在领域分类模型上的概率分布间接计算相似性,从而提高了搜索的平均精度.

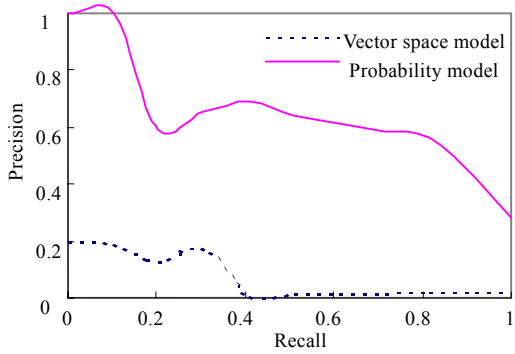

Fig.3 Comparison between two models

图 3 两种模型之间的比较

5 总结与展望

个性化服务是一种趋势,通用的检索系统不可能满足不同背景、不同目的和不同时期的查询请求.本文尝试解决了基于内容过滤技术中的一些问题,通过实验分析了影响基于内容过滤的因素,实验表明,算法的搜索精度有了很大的提高.

另外,我们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.但是,准确地描述用户兴趣是一个困难的问题,我们将继续跟踪这方面的技术,更准确地表达用户的兴趣和行为,进一步提高个性化搜索的平均精度.

References:

[1] Zeng C, Xing CX, Zhou LZ. A survey of personalization technology. Journal of Software, 2002,13(10):1952~1961 (in Chinese with English abstract).

[2] Pretschner A. Ontology based personalized search [MS. Thesis]. Lawrence, KS: University of Kansas, 1999.

[3] Dumais ST, Platt J, Heckerman D, Sahami M. Inductive learning algorithms and representations for text categorization. In: French J, Gardarin G, eds. Proceedings of the International Conference on Information and Knowledge Management. New York: ACM Press, 1998. 148~155.

[4] Witten IH, Paynter GW, Frank E, Gutwin C, Nevill-Manning CG. KEA: practical automatic keyphrase extraction. In: Fox EA, ed. Proceedings of the 4th ACM Conference on Digital Library. New York: ACM Press, 1999. 254~255.

[5] Turney PD. Learning algorithms for keyphrase extraction. Information Retrieval, 2000,2(4):303~336.

[6] Joachims T. A probabilistic analysis of the rocchio algorithm with TFIDF for text categorization. In: Fisher DH, ed. Proceedings of the 14th International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers, 1997. 143~151.

[7] Bollacker KD, Lawrence S, Giles CL. Discovering relevant scientific literature on the Web. IEEE Intelligent Systems, 2000,15(2):42~47.

[8] Hofmann T. Probabilistic latent semantic analysis. In: Laskey KB, Prade H, eds. Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers, 1999. 289~296.

附中文参考文献:

[1] 曾春,邢春晓,周立柱.个性化服务技术综述.软件学报,2002,13(10):1952~1961.