个性化浏览中网页推荐的结构模型*)

苏贵洋　王永成　马颖华
（上海交通大学计算机系　上海200030）

Webpage Recommendation Model in Personal Surfing

SU Gui-Yang　WANG Yong-Cheng　MA Ying-Hua
(Dept. Computer Science, Shanghai Jiao Tong University, Shanghai, P. R. China, 200030)
su-gy@mail.cs.sjtu.edu.cn

Abstract　This paper first introduces the inevitability direction of Internet personal technique's development, and then analyzes the differences between Information Retrieval and Information Filtering. Further more, combining with the group user interests and individual user interests, using VSM and FCM algorithm, this paper proposes a webpage recommendation model in personal surfing. This model can be applied not only in single website, but also in enterprise information collecting system. It also brings forward a method to extend concept. The advantage of the model is automatic concepts learning in the processing of user's feedback and browsing. Contrasting to concept dictionary method, it is more agility and gains well retrieval results.

Keywords　Personal Webpage recommendation, Information accessing, Group interests clustering

1　前言

随着互联网建设的不断发展，网站和网页数目都已经很难用 Lawrence 在 Science[1]和 Nature[2]给出的结论来估算。每个在网上冲浪的人都有体会，迷失在 Internet 浩瀚的资料中是多么容易。世界上最大的搜索引擎 Google[3]已经宣称搜索并索引了2,073,418,204张网页，在这近21亿索引并分类的网页中搜索，用户同样会迷失在层层链接之中。

用户上网的主要目的，是获得自己需要的信息。面对烦杂的网络信息，人们正在寻求一种将用户感兴趣的信息主动推荐给用户的服务方式，这就是个性化的信息服务。新一代的 Internet 要求能够对用户进行更加个性化的服务，例如通过自动学习用户兴趣，推荐给用户可能感兴趣的站点和文章等等。

在信息检索领域中，信息过滤（Information Filtering--IF）技术和信息检索（Information Retrieval--IR）技术是目前用户进行信息获取的两大技术。信息过滤技术使用用户模型（User Profile）来描述用户兴趣，将新的文献与用户模型进行相似度计算，主动将相关度高的文献发送给该用户模型的注册用户，从而实现个性化的信息服务。而信息检索则是基于用户给定的检索命令在检索资源中查找符合要求的检索内容。Croft 指出了信息过滤与信息检索的不同[4]，文[5]也对此进行了分析[5]，综合起来信息过滤和信息检索的主要区别如下：

·信息过滤关注用户的长线需求（指在一段时间内，比较固定的信息需求）；而信息检索的用户查询则是随机的、易变的。

·信息过滤所面对的信息源是动态的；而信息检索所面对的信息流是相对稳定的。

虽然信息检索与信息过滤有以上区别，但信息检索和信息过滤的目的都是帮助用户寻找自己感兴趣的信息，它们有着极为密切的联系，信息过滤实质是建立在信息检索基础之上的。

广泛应用在网站建设中的信息获取技术是个性化网页推荐。一般来说，网站的个性化服务有两种：第一种服务是通过提供给用户自定义的类别定制。例如在 Yahoo!新闻中可以定义用户感兴趣的新闻类别，有些网站也提供基于类别的用户定制定期的 Email 通知。第二种服务借助用户既往浏览过的内容来提供。例如在 Yahoo!中用户可以把搜索引擎的分类类目定制到 Favorite 夹中，再次访问搜索引擎时，就可以通过定做的链接入口直接进入该分类网页。

基于类别的个性化服务局限性强，它取决于类别的设定，以及网页内容的分类。分类法一般来说较为粗糙，导致信息获取查全率（Recall）高而查准率（Precision）不高，对于大多数用户来说无法提供少而精的信息推荐。而基于既往浏览内容的个性化服务局限性更大，对于较为新鲜的信息内容，既往浏览的个性化服务通常会被忽略，但是实际上越是新颖的信息内容对于用户来说可能价值越大。

把信息检索和信息过滤的有关技术应用到个性化服务中，能有效地改善目前个性化服务的不足。本文试图构造一种有效的模型，应用部分信息检索和信息过滤技术实现在用户上网浏览过程中的个性化过程。

2　个性化浏览中网页推荐的结构模型

个性化浏览中，首先应该完成对用户兴趣的建模。以往的一些信息过滤系统，人们常常使用关键词来构造用户模型[6]。个性化服务中同样可以以关键词为特征进行个性化描述。关键词方法与类别方法相比有表达自由度高，模型维度高等特点。

*)本课题受国家自然科学基金资助（基金号：60082003）。苏贵洋　博士研究生，主要研究领域为网络信息智能处理。王永成　教授，博士生导师，主要研究方向为中文信息处理。马颖华　博士研究生，主要研究领域为自动标引及算法设计。

不过关键词方法也有其弊端，因为关键词往往并不能清晰地表达主题，这主要是由于词有一词多意、同义多词等现象，另外，关键词的专指度不如词组、句等更大的语言单位。完全以关键词为内容的建模方法，常常会造成不正确的匹配和漏检。目前信息检索技术在关键词的基础上，提出了"概念检索"(Concept Retrieval)。这种检索方法考虑了词与词之间的关系，采用自动或者手工建立的概念关联词典，对关键词检索项进行扩充和限定，试图解决一词多意、同义多词的问题。常见的技术有：潜在语义索引(Latent Semantic Indexing--LSI)[7~9]、概念词典法[10]等。

但 LSI 技术运算量大，以及它的奇异值分解(Singular Value Decomposition--SVD)降维运算无法进行语言学方面的解释等弱点，使得 LSI 的应用有一定的难度。概念词典的建立是一项非常庞大的工作，目前最大也是最常用的概念词典有 Wordnet，汉语有 Hownet 等，它们的词汇覆盖面仍然不尽人意，尤其是一些新创词汇不能及时收入词典。而 Internet 是信息流通最为活跃的地方，同样也是语言发展最为活跃的地方，每天都有新的词语被创建和使用，词典的方法对 Internet 应用来说不够灵活。

根据文[11]的研究结果：群体用户兴趣，可以对个体用户兴趣有较重要的影响作用[11]。这里，我们提出一种方法，根据个体用户兴趣聚类生成群体用户兴趣模型，通过个体对群体用户兴趣的继承来实现信息检索过程的概念扩充。本文讨论了在个性化浏览中应用该方法进行网页推荐的结构模型。该方法比词典法更灵活——因为它不需要任何词典，并且对新兴词语同样适用；与 LSI 方法相比，该方法计算量小，更易于理解和实现。模型示意如图1所示。

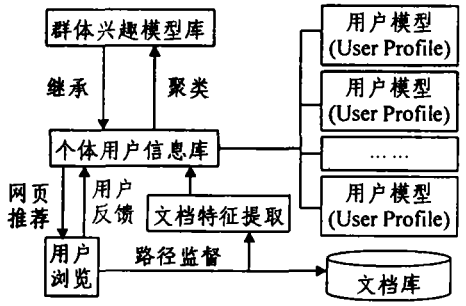

图1 个性化浏览中网页推荐的结构模型

在图1中，首先提出的是对群体用户进行建模。而每一个具体用户的兴趣，可以从群体用户兴趣中进行学习(继承)。在用户登陆后，将对用户的个性浏览(文档、时间、点击次数等)记录、分析和进行特征提取，从而进一步补充和完善用户个性信息，并反馈到群体用户信息库中，对群体用户兴趣库进行更新。

下面将对该模型中的核心算法和流程进行分析。

2.1 路径监督和特征提取

在用户浏览过程中要进行路径监督，这可以简单地采用记录到日志中的办法，然后再由程序去对日志文件进行分析。日志中可以记录用户浏览的路径，如对每一个链接的点击次数，浏览的文档，浏览时间等等，这些都可以表达用户的兴趣 U_I 以及兴趣的持久度 U_L 等信息。

特征提取即将用户浏览过的网页进行抽词，并转化为特征向量，从而得到学习的样本。鉴于 HTML 的特有特性，系统将首先对 HTML 文件进行预处理，去除格式 Tag，抽取文档主要内容，并对主要内容进行分析和关键词抽取。然后对文件进行关键词词频统计(tf/idf)，根据词频高低，建立特征向量。

2.2 个体用户模型的表示

向量空间模型 VSM(Vector Space Model)是60年代末 Gerard Salton 等人提出的。它是最早也是最出名的信息检索方面的数学模型。下面先介绍一下模型的基本概念。

文档 泛指一般的文献，此处指网络文档。

项 当文档的内容被简单地看成是它含有的基本语言单位(字、词、词组或短语等)所组成的集合时，这些基本的语言单位统称为项，即文档可以用项集(Term List)表示为 $D(T_1, T_2, \cdots, T_n)$，其中 T_k 是项，$1 \leqslant k \leqslant n$。

项的权重 对于含有 n 个项的文档 $D(T_1, T_2, \cdots, T_n)$，项 T_k 常常被赋予一定的权重 W_k，表示它在文档中的重要程度，即 $D=D(T_1, W_1; T_2, W_2; \cdots; T_n, W_n)$，简记为 $D=D(W_1, W_2; \cdots, W_n)$。项 T_k 的权重是 W_k，$1 \leqslant k \leqslant n$。

向量空间模型 给定一文档 $D=D(T_1, W_1; T_2, W_2; \cdots; T_n, W_n)$，简化起见，暂不考虑 T_k 在文档中的先后次序并要求 T_k 互异，这时可以把 $T_1, T_2, \cdots, T_n$ 看成一个 n 维的坐标系，而 $W_1, W_2; \cdots, W_n$ 为相应的坐标值，因而 $D(W_1, W_2; \cdots, W_n)$ 被看成是 n 维空间中的一个向量。称 $D(W_1, W_2; \cdots, W_n)$ 为文档 D 的向量表示。具体如图2所示。

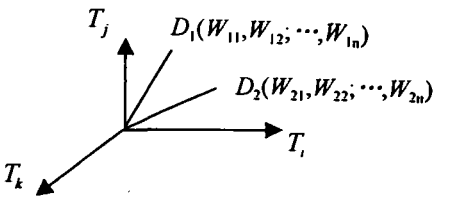

图2 文档的向量空间模型(VSM)及文档间的相似度

相似度 两个文档 D_1 和 D_2，可以借助于向量间的某种距离来表示它们之间的相似度。相似度常用向量之间的内积来计算，其定义相当于向量的夹角余弦，如下式(1)所示。

$$Sim(V_i, D_j)=\cos\theta=\sum_{k=1}^{m} v_{ik} \cdot d_{jk} / \left(\sqrt{\sum_{k=1}^{m} v_{ik}^2} \cdot \sqrt{\sum_{k=1}^{m} d_{jk}^2}\right) \quad (1)$$

采用向量空间模型表示文档的特征后，用户的兴趣就可以看成是一个文档，也就可以表示为一个向量 U。文档与用户兴趣的相似程度就可以用文档向量 V 与用户兴趣向量 U 的余弦相似度 $Sim(V, U)$ 来表示。

对文档向量各特征项加权时，可以简单地以文档特征项的词频作为其权值。对 HTML 网页文本，为了更好地体现文档的特征信息，可以对文档进行结构分析，提取文档结构属性(title，meta，strong，font size，H1，……)，增大特征项的权重。

对用户兴趣模型表示的设计，可以遵循面向对象结构化语言设计时类的构造思想和原则：层层分类，使得概念逐渐细化(群体兴趣—>个体兴趣)。

因此，应用向量空间模型，我们可以构造群体用户兴趣；同样，对用户兴趣的表达，首先要让用户继承某类别的群体用户兴趣，其次再学习生成个性的特征。

2.3 聚类以及群体兴趣生成

聚类时，该结构框架采用模糊 C-均值 (FCM)聚类算法。

FCM 算法是一个使目标函数 $J=\sum_{k=1}^{n}\sum_{i=1}^{C} u_{ik}^{m} \| x_k - v_i \|_A^2$ 最小化的迭代优化过程。其中，n 是 s 维样本集合中样本的个

数，C 是要划分的类数，x_k 是第 k 个数据样本，v_i 是第 i 个类的类中心矢量，u_{ik} 是第 k 个数据样本属于第 i 个类的隶属程度，m 是一个大于1的常数。给定类数 C 和初始隶属度矩阵 $\{u_{ik}\}_{n\times c}$（或初始中心矩阵 $\{v_i\}_{i\times c}$）。FCM 算法通过迭代式(2)、(3)，收敛到目标函数的一个局部极小点或一个鞍点。

$$u_{ik}=[\frac{1}{d_{ik}}]^{\frac{2}{(m-1)}}/\sum_{j=1}^{C}[\frac{1}{d_{jk}}]^{\frac{2}{(m-1)}},k=1,2,\cdots,n,i=1,2,\cdots,C \quad (2)$$

$$v_i=\sum_{k=1}^{n}u_{ik}^m x_k/\sum_{k=1}^{n}u_{ik}^m,i=1,2,\cdots,C \quad (3)$$

其中 $d_{ik}=\|x_k-v_i\|_A$ 是样本点到类中心的距离。

使用该聚类方法，可以将相似的个体用户兴趣归类，便于新的用户继承已有类别用户的兴趣特征。

聚类的过程将兴趣类似用户的兴趣进行了综合，扩充了个人兴趣表述中的向量空间的维度。个人兴趣继承群体兴趣后，个人兴趣描述向量将根据群体兴趣进行修订。通过这种聚类、继承的运算后，用户的兴趣相互补充，完成了概念扩充的过程，使得查全率增高。

2.4 网页推荐及用户反馈

设用户兴趣度为 I，定义一个阈值 I_k，对文档集中 $Sim(V,U)>I_k$ 的文档，则由系统自动推荐给用户。

用户可以对过滤后文档集中的文档按照与自己信息需求的相关程度打分，分数可以分为几个级别（例如可以设为3个级别，也可以分得更多更细），最低为0，最高为1。设文档为 D，兴趣度为 I，反馈信息形式可以表示为：$\{(D_1,I_1)(D_r,I_r)\cdots(D_m,I_m)\}$。根据用户的反馈，需要及时调整用户模型，必要时调整群体用户模型。

用户反馈是提高查准率的重要手段，该模型中能够很方便地实现这个反馈过程。这里用户反馈的兴趣使用关键词描述方法，因此反馈信息可以方便地用于修改关键词特征向量的参数。

结论 本文提出了利用向量空间模型(VSM)和 FCM 聚类算法聚类生成群体用户兴趣和个体用户兴趣在服务器端进行网页推荐的结构模型。在已经进行的小样本的试验中，初步地证明了该方法有一定的可行性，并且得到了较好的试验结果。这个方法能生成具有概念扩充能力的检索和过滤命令，从而更好地帮助用户完成其特定信息的获取。这个结构模型不仅可以应用在单个网站中，提供用户个性化浏览页面；也可以应用在企业的信息采集系统中，提供一种概念扩充的检索方法。这种结构模型的优点在于其在用户的反馈和浏览中进行概念的扩充和学习，比一般的概念词典扩充方法更灵活，检索效果更好。

参考文献

1 Setal L. Searching the World Wide Web. Science, 1998, 280 (5360):98～100

2 Lawrence S, Giles C L. Accessibility of information on the Web. Nature, 1999, 400:107～109

3 http://www.google.com

4 Belkin N J, Croft W B. Information filtering andinformation retrieval: Two sides of the same coin. Communication of ACM, 1992, 35(12):29～38

5 林鸿飞. 基于 Web 的信息过滤机制. 计算机工程与应用, 2002(2)

6 Pazzani M, Billsus D. L earning and revising userprofiles: The Identification of Interesting Web Sites. Machine Learning, 1997, 27:313～331

7 Letsche T A, Berry M W. Large-Scale Information Retrieval with Latent Semantic Indexing. Infomrmation Sciences, 1997

8 周水庚，关佶红，胡运发. 隐含语义索引及其在中文文本处理中的应用研究. 小型微型计算机系统, 2001, 22(2)

9 牛伟霞，张永奎. 潜在语义索引方法在信息过滤中的应用. 计算机工程与应用, 2001, 9

10 Moldovan D I, Mihalcea R. UsingWordNet and Lexical Operators to Improve Internet Searches. INTERNET COMPUTING, Jan.-Feb. 2000

11 Tu H-C, Hsiang J. An architecutre and category knowledge for intelligent information retrieval agents Decision Support Systems, 2000, 28:255～268

(上接第72页)

通过模拟实验我们发现：如果一个媒体在线时，能够确定其流行系数，将大大地改善整个流媒体系统的性能，它可以从完全缓存变成部分缓存，也可以从部分缓存变成完全缓存，由该媒体的访问度和访问的时间确定。λ、λ_C 或 λ_P 对流行性系数的影响很大，它们的取值与用户访问模式与分布的关系我们还要作深入的研究。

结束语 Proxy 缓存是一种由成本效益(cost-effective)改善互连网上流媒体服务系统性能的解决方案。通过在 Proxy 上缓存经常被访问的数据，将会极大地减少用户感知的延迟、服务器负载和网络流量。然而，传统的 Web Caching 技术不能很好地适用于流媒体数据，我们针对社区宽带网的特点，提出的一种基于 Server 和 Proxy 流行性的流媒体 Caching 策略，利用流媒体在流服务器的流行性系数 P_S 和在 Proxy 上的流行性系数 P_P，在 Proxy 端对流媒体实行部分缓存或完全缓存，改善了整个流媒体系统的性能。特别是在服务器端如果能准确地给出 P_S 的初值，就更能体现出该策略的优越性。由于 Proxy 与流服务器之间的传输路径可能要经过好几个不同的 ISP，我们下一步的工作就是探讨和解决流媒体在这段路径上的包丢失的恢复和质量调整给已缓存的数据量带来变化的问题。

参考文献

1 吴国勇，邱学刚，等. 网络视频 流媒体技术与应用. 北京邮电大学出版社, 2001

2 徐光怪，史元春译. 多媒体系统设计. 电子工业出版社, 1998

3 Jrexfod S S, Towsley D. Proxy Prefix Caching for Multimedia Streams. In: Proc. IEEE Infocom, March 1999

4 Rejaie R, et al. Multimedia Proxy Caching Mechanism for Quality Adaptive Streaming Applications in the Internet. In: Proc. IEEE Infocom, March 2000

5 Reisslein M, Hartanto F, Ross K W. Interactive Video Streaming With Proxy Servers. In: Proc. IEEE Infocom, April 2000

6 Gibson G A, Vitter J S, Wilkes J. Storage and I/O Issues in Large-Scale Computing. In: ACM Workshop on Strategic Directions in Computing Research, ACM Computing Surveys, 1996. http://www.medg.lcs.mit.edu/doyle/sdcr.

7 Wang Y, Zhang Z-L, Du D, Su D. A Network-Conscious Approach to End-to-End Video Delivery over Wide Area Networks Using Proxy Servers. In: Proc. IEEE Infocom, April 1998