

文章编号:1673-0062(2012)01-0054-05

一种基于桌面信息的个性化推荐方法

李晓昀,余 颖,阳小华,万亚平,马家宇,刘志明,蒋 辉

(南华大学 计算机科学与技术学院,湖南 衡阳 421001)

摘 要:针对当前信息检索服务中存在的固有缺陷,提出了一种基于用户桌面信息抽取的个性化推荐方法.详细介绍了通过用户桌面资源信息抽取建立长期用户模型,以及通过工作场景信息抽取建立短期用户模型的算法.长期用户模型提供了完整全面的用户兴趣偏好信息,短期用户模型则为预测用户当前信息需求提供了依据.实验结果表明,基于用户桌面信息抽取的个性化推荐服务能较好地预测用户当前需求、具有良好的推荐效果.

关键词:桌面;个性化推荐;信息抽取;个人信息空间

中图分类号:TP393;TP311 **文献标识码:**B

An Approach to Personalize Recommendation Based on Desktop Information Extraction

LI Xiao-yun, YU Ying, YANG Xiao-hua, WAN Ya-ping, MA Jia-yu,
LIU Zhi-ming, JIANG Hui

(School of Computer Science and Technology, University of South China, Hengyang, Hunan 421001, China)

Abstract: In terms of the inherent detects in present information retrieval service, this paper proposed an approach to personalize recommendation based on desktop information extraction and introduced the algorithm to build the long-term user model based on desktop resources extraction, which provides information about the comprehensive interest preference of a user, and to establish the short-term user model based on working scenario information extraction, which serves as the basis to predict the user's current information need. The experiment results showed that this approach could efficiently and effectively predict the user's current information need and achieved pretty good recommendation results.

key words: desktop; information extraction; personalized recommendation; personal information space

收稿日期:2011-10-08

基金项目:湖南省自然科学基金资助项目(10JJ6097);湖南省教育厅科研基金资助项目(11C1098);湖南省科技计划基金资助项目(2010GK3011);南华大学校级教改课题基金资助项目(2010ZZ030)

作者简介:李晓昀(1979-),女,江苏苏州人,南华大学计算机科学与技术学院讲师,硕士.主要研究方向:智能信息系统与知识管理,个性化服务,信息检索.

0 引 言

数字化资源的不断增长,使得信息超载和信息匮乏问题日益严重. 无论是在互联网还是在用户的本地硬盘上,人们越来越难迅速地找到自己所需的信息. 虽然各类搜索引擎为人们进行信息检索提供了一个较好的平台,但这种盲目搜索的方式依然受到以下三方面因素的制约,无法为用户提供更好的信息服务:1) 人们的检索目的表达不明确、不充分;2) 信息提供者的表达不明确、不充分;3) 人们不可能浏览所有的返回链接.

针对信息检索系统所固有的缺陷,推荐技术的发展虽然在一定程度上帮助用户获得所需信息,但目前的推荐研究大多只是基于用户访问某个 Web 站点时所获得的个人信息,包括收集如浏览时间、鼠标单击、移动、滚动情况等隐性反馈信息^[1-2],通过分析用户的浏览历史来获取用户隐性反馈信息^[3]、建立用户点击模型^[4]等. 事实上,用户的个人电脑中保存了大量真实的、高度相关的个人信息,目前对这些桌面信息的研究相对集中在提高信息检索质量^[5-6]、个人信息管理^[7]等方面. 为更好地满足用户的个性化

需求,本文提出一种基于桌面信息抽取建立用户模型,进而为用户提供个性化推荐的服务的方法. 实验结果表明,用户对为所推荐的项目信息具有一定兴趣,这说明基于用户桌面信息建立起来的用户兴趣模型,能够较好的反映用户的兴趣偏好,预测用户需求,从而改善推荐系统中用户的个性化服务效果.

1 桌面信息抽取模型

由于用户个人电脑的私用性,其中必然存储了大量与用户的学习、工作、生活及个人兴趣爱好等密切相关的资料,如文本文档、影音文件及图片等(以下均将这些文件称为桌面文档). 除了这些文档的内容之外,在用户使用这些文档的过程中,也遗留了大量与用户个人兴趣偏好紧密相关的珍贵信息. 通过对这些文档进行分析,并进行相关信息抽取,能够获得与该用户密切相关的、较真实且全面的长期用户模型. 然而一般地,在一个特定时间之内,用户使用个人电脑都是针对某一个特定的任务,如撰写论文、查找信息、浏览新闻等. 因此,针对用户使用个人电脑的这两种特性,用户桌面信息抽取可分为两个主要部分:桌面资源分析及工作场景分析,如图 1 所示.

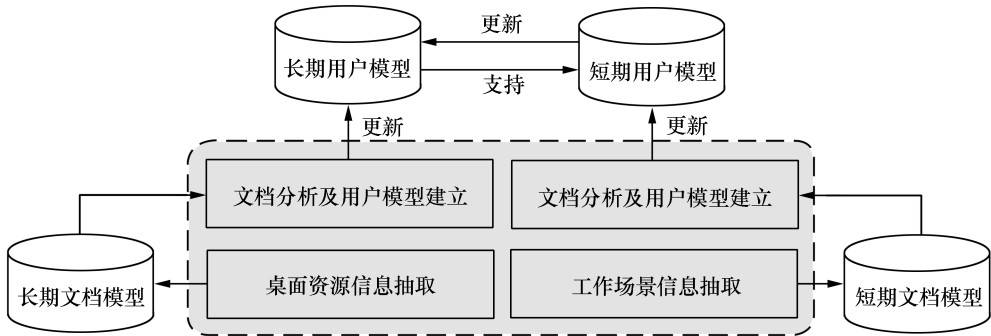


图 1 桌面信息抽取模型

Fig.1 Model of desktop information extraction

桌面资源信息抽取与工作场景信息抽取是相互独立却又相互关联的两部分. 用户的本地硬盘中长期存储的各类文档及相应操作历史数据较全面地反应了用户兴趣偏好,为获得较为完整的用户个人信息空间描述、建立长期用户模型提供了依据. 相对而言,用户工作场景则为判断用户当前特定的工作任务类别、预测其当前信息需求提供了珍贵信息.

2 桌面信息抽取算法

2.1 桌面文档资源抽取

用户将文档存储在其个人电脑中,在一定程度上表示该用户认可了这些文档的价值,即与其需求相关联. 然而,在用户长期使用电脑的过程中,这些文档并非具有同等重要性. 在文档的操作历史中,隐藏了一类重要信息,这类信息反应用户

对各个文档的需求程度. 因此, 为建立全面且真实的长期用户模型, 将桌面文档信息分为两类进行抽取: 文档操作历史及文档内容, 如图 2 所示.

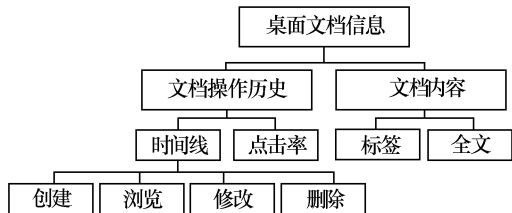


图 2 桌面文档信息

Fig. 2 Desktop document information

1) 桌面文档操作历史信息抽取

文档的操作历史信息分为两部分: 文档的时间线及点击次数. 受文献[8]的启发, 由于一个用户的兴趣可能随着时间变化而偏移, 所以引入时间敏感度作为表示某一文档的重要程度的因素之一. 时间线包含一个文档的创建时间、上次浏览时间、最近修改时间以及删除时间. 由此获得一个文档的时间敏感度 $timesst$, 如公式(1)所示.

$$timesst = del. (\alpha \cdot creat + \beta \cdot view + \gamma \cdot modify) \quad (1)$$

其中, α, β 和 γ 是参数. $creat, view, modify$ 分别表示经模糊语义分析^[3]后一个文档的创建时间敏感度、上次浏览时间敏感度, 以及最近修改时间敏感度. 其中, 被删除文档的相关信息在一定时间阈值 θ 内被保存. 记录文档的删除时间能提供一类重要信息, 如果一个文档被用户删除, 可认为该用户视该文档价值不大或使用完不再需要. 因此, 在个性化推荐中, 被删除文档的信息记录为去除那些与该文档相关、却对该用户无用的信息提供了重要依据. 实验结果表明, 用户极少点击那些他们曾经浏览过且已删除了的文档, 这意味着删除时间的记录为及时掌握用户兴趣偏移、改善用户体验提供了重要信息.

除了桌面文档的相关时间信息外, 文档的点击次数 $hits$ 亦是一类反映用户偏好的信息. 例如, 一些文档虽然新, 但用户短暂浏览后发现无价值, 却不删除该文档, 亦不再打开, 由此可判断该文档的价值相对较低; 而另一些文档, 虽然存储在用户硬盘中时间较长, 用户却经常浏览或进行相关的操作, 这就表明该用户依然认为这些文档的价值相对较高.

一个文档的权值信息 doc_j 由公式 3 计算

获得.

$$doc_j(timesst, hits) = A \cdot timesst + B \cdot hits + C \quad (2)$$

其中, A, B, C 为经验值.

2) 桌面文档内容信息抽取

相关度是个性化服务的关键. 存储在用户个人电脑中的文档在为用户提供个性化服务时, 相对信息世界的其他文档, 具有更高价值. 针对桌面文件类型的多样性特点, 除了抽取如影音文件中的标签信息外, 主要对文本文件则进行词频分析, 提取 $Top-N$ 关键词, 进一步采用 TF-IDF 权重计算方法, 结合文档权值 doc_j , 可获得用户模型中一个关键词 i 的相应权值 $weight$, 如公式(3)所示.

$$weight_i = doc_j \cdot \frac{n_{i,j}}{\sum n_{k,j}} \cdot \log\left(\frac{N}{n}\right) \quad (3)$$

其中, $n_{i,j}$ 为关键词 i 在文档 j 中的出现次数, 分母为该文档中所有出现的关键词个数; $\log(N/n)$ 为逆向文件频率, N 为文档集中文档总数, n 为该文档集中出现关键词 i 的文档数.

2.2 工作场景信息抽取

虽然桌面资源分析为个性化服务提供了全面、完整的用户个人信息, 但事实上, 在某一段特定时间内, 用户的当前任务相对一致. 以撰写学术论文为例, 用户可能浏览本地已有的相关文档或使用因特网检索其他相关文献等. 由此有理由认为在任务栏上已打开的、或打开时间超过一定阈值的文档的内容是用户视为与其当前研究任务密切相关的. 因此, 当前的工作场景信息, 相对于桌面其他文档信息, 对预测当前用户需求更具有价值.

工作场景文档集包含当前打开的本地文档及正在浏览的网页. 前者的相关信息由长期文档模型中获取, 而网页信息主要包含三部分内容:

1) 网页内容信息, 由分词获取 $Top-N$ 关键词, 并以此来描述该网页内容.

2) 网页的浏览时间, 时间信息经过模糊语义分析^[3], 可反应出用户对该网页的相应关注程度.

3) 用户的特殊行为, 包括打印、收藏、保存等, 这些动作暗示了用户认可该网页的价值. 抽取工作场景文档集信息的, 即可获取短期用户模型, 在长期用户模型的支持下, 即可对用户当前任务进行判断, 并预测其需求.

3 个性化推荐实验系统构建

实验以学术研究为实验场景, 辅助检索系统

以石鼓搜索(www.usc.edu.cn)为基础.实验要求参与者在个人电脑上安装客户端,设置实时监控的磁盘或文件夹,从而建立起长期用户模型.实验过程中,分析用户工作场景建立短期用户模型,预测用户的当前信息需求,同时启动石鼓搜索在后台自动为用户进行相关信息检索,并结合长期用户模型,对检索结果进行精炼,最后将最相关的Top-N 信息项目以推荐窗口的形式提交到用户桌面.用户对推荐信息的行为反馈将作为系统各模块参数调整的依据.系统结构如图 3 所示.

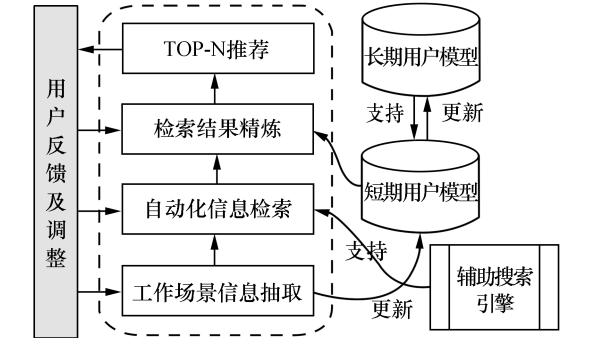


图 3 个性化推荐系统结构图

Fig.3 Structure of Personalized Recommendation System

4 实验结果分析

用户对被推荐项目的反馈体现在点击行为、浏览时间及其他相关操作(如打印、保存和收藏等)中,这些反馈信息在一定程度上反映了推荐效果及稳定性.因此,采用文献[2]中所述方法,通过分析用户对被推荐项目信息的反馈信息,包括用户对被推荐项目的点击行为以及相应浏览时间,计算用户对被推荐项目的兴趣度,以此度量推荐效果.用户对项目的兴趣度分为 1~5 五个级别,分别表示极不感兴趣、不感兴趣、一般、感兴趣和极感兴趣.参与测试者共 25 人,通过用户日志分析计算用户的平均兴趣度,结果如表 1 所示,其分析图表如图 4 所示.

表 1 用户平均兴趣度
Table 1 User average interest

用户数	5	10	15	20	25
平均兴趣度	4.27	4.19	4.23	4.31	4.28

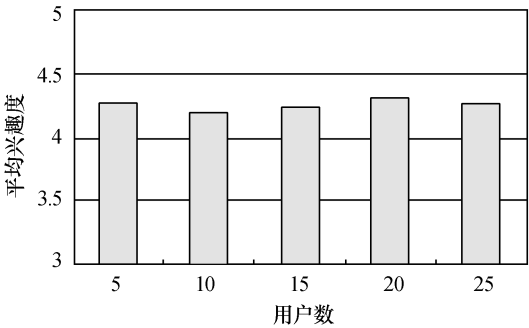


图 4 平均兴趣度分析图表

Fig.4 Analysis chart of average interest degree

平均兴趣度分析图表显示出参与测试者对系统所推荐的项目信息的平均兴趣度介于“有兴趣”与“极有兴趣”之间,这说明系统的推荐效果良好,基本迎合了用户信息需求.另一方面,从随机抽取的用户兴趣度平均值分析来看,其值始终保持在一定范围内,表明该系统的推荐效果基本稳定.

5 总 结

基于用户桌面信息抽取的个性化推荐方法通过抽取用户桌面文档信息,建立了真实且全面的长期用户模型,通过工作场景文档集信息抽取建立短期用户模型,较准确地预测用户当前信息需求,并将所推荐项目信息以窗口的形式呈现在用户桌面.实验证明,这种积极主动的桌面推荐方式具有较好的推荐效果,能为用户提供高效的个性化信息服务.然而,桌面信息抽取涉及的个人隐私及安全问题,以及工作场景信息的多样性、抽取效率和系统反应速度较慢等问题仍需进一步研究.

参考文献:

[1] Zeyuan Allen Zhu, Chen Weizhu, et al. A novel click model and its applications to online advertising[C]// Proc. 3rd WSDM. New York, USA. 2010;321-330.
[2] Xiaoyun Li, Ying Yu. Research on self-adaptive recommendation system based on implicit feedback[J]. Computer Engineering, 2010, 36(16): 270-272.
[3] Nunberg G. As google goes, so goes the nation[N]. New York Times, 2003 May.
[4] Olivier Chapelle, Ya Zhang. A dynamic bayesian network click model for web search ranking[C]// Proc. 18th WWW. Madrid, Spain, 2009;1-10.

(上接第 57 页)

- [5] Jaime Teevan, Susan T Dumais, Eric Horvitz. Potential for Personalization [C] // Proc. TOCHI, March 2010.
- [6] Paul-Alexandru Chirita, Claudiu S Firan, Wolfgang Nejdl. Summarizing local context to personalize global web search [C] // Proc. 15th CIKM. Arlington, Virginia, USA. Nov. 2006.
- [7] Christopher Peery, Wei Wang, et al. Multi-dimensional search for personal information management systems

[C] // Proc. 11th international conference on Extending database technology: Advances in database technology. Nantes, France. March, 2008.

- [8] Jaime Teevan, Susan T. Dumais, Eric Horvitz. Personalizing Search via Automated Analysis of Interests and Activities [C] // Proc. 28th SIGIR. Salvador, Brazil. August 2005.