基于上下文的个性化信息检索技术研究

王威

指导教师 林坤辉 教授

厦门大学

学校编码：10384　　分类号＿＿＿密级＿＿＿

学号：24320061152649　　UDC ＿＿＿

厦门大学

硕　士　学　位　论　文

基于上下文的个性化信息检索技术研究

Research On Context Sensitive

Personalized Information Retrieval Technology

王 威

指导教师姓名：林 坤 辉　　教授

专　业　名　称：计算机软件与理论

论文提交日期：2009 年 4 月

论文答辩时间：2009 年 6 月

学位授予日期：2009 年　月

答辩委员会主席：＿＿＿＿＿＿

评　　阅　　人：＿＿＿＿＿＿

2009 年　　月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下，独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

摘　要

搜索引擎是互联网用户最常用的信息查询工具。目前主流的搜索引擎并没有明确区分不同用户的查询意图，而不同用户即使输入相同的查询词，其查询需求也是有差别的。个性化信息检索技术就是针对以上问题提出的。个性化信息检索通过收集和分析个人信息和查询的上下文，而不是仅仅依靠检索词来判断用户的真实需求，因而能够根据用户的不同需求而返回个性化的检索结果以提高检索精度

本文在对个性化信息检索相关技术进行了较为全面、深入的分析基础上，分别研究短期上下文和长期上下文的个性化信息，以及如何根据基于上下文的个性化信息进行个性化模型建模，从而改善信息检索系统查询性能。最后搭建了一个基于上下文的个性化检索原型系统。

在研究短期上下文的个性化信息方面，为了改善信息检索系统对 Ad hoc 请求只针对查询词的缺点，首先给出了基于上下文的个性化检索的形式化描述，其次设计了短期上下文的个性化检索算法。该算法以单元统计语言模型为基础，结合隐性相关反馈技术，克服了用户 Ad hoc 请求时仅依靠单独查询词的局限性。通过实验证明该算法使查询精度平均提高 50%。

在研究长期上下文的个性化信息方面，为了克服在传统信息检索系统中，无法根据个人长期行为特点进行响应的缺点，本文以北大网络中心对天网搜索引擎的用户行为分析为基础，设计了长期上下文的个性化检索算法。该算法通过分析用户行为日志，建立起长期个性化模型，对当前查询起到改善作用。通过实验证明了该方法的有效性和较好的抗噪声性。

本文设计并实现了一个基于火狐浏览器的个性化检索系统。该系统以基于短期上下文的个性化检索算法为算法基础，利用 Lemur 语言模型工具，以搜狗实验室公开的全网新闻数据作为语料集。同时利用该系统作为今后研究的实验平台，搜集真实用户行为日志，为今后研究工作奠定基础。

关键词：个性化检索；相关反馈；语言模型

Abstract

Search Engine is a popular tool for information retrieval (IR) nowadays. However, the traditional search engine dose not do a job in identifying the individual's unique search goal, while it only returns the retrieval results associating with the query user provides. In order to overcome the problem, there have been many attempts to improve retrieval accuracy based on personalized information retrieval technology.

Based on in-depth survey on the existing studies about personalized IR, this paper discusses separately on the short-term context and the long-term context personalized information. The thesis did research on how to use the context-based personalized information to improve the retrieval accuracy. Finally, a prototype system of client-side personalized search agent was built. The main work of the thesis includes:

In terms of the short-term context personalized information, first we proposed a formal description of the personalized information retrieval. Second, based on the statistical language model and the implicit feedback technology, we proposed an approach to iteratively update the query model according to the short-term context such as query histories and clicked documents. This method was proved to be satisfied with the user ad hoc query intention. By the experiment, it found that the the mean average precision is increased by about 50%.

In terms of the long-term context personalized information, first we discussed about long-term context, which is based on the research of the Peking University Network Center on the Skynet search engine. Second we improved the approach of building long-term user profile by interpolating the query history language models. Finally by the experiment, we found that this improved method has the good anti-noisy performance.

Finally, a prototype system of personalized IR system was built, which used the context-sensitive personalized information. This system was built as a real experimental environment for the future research.

Keywords: personalized information retrieval; context-sensitive information retrieval; statistical language model

目录

Contents

第一章 绪论

随着互联网信息日益丰富，在日常工作生活中人们越来越依赖信息检索系统来查找所需的信息。信息检索（Information Retrieval）这一术语最早是由 Calvin N. Mooers 在 1950 年的 Zator Technical Bulletin (No.48)中公开提出的。信息检索最初主要是应用于图书馆中的文献检索，1954 年美国海军兵器中心（NOTS）图书馆在 IBM701 型号计算机上成功建立了世界上第一个计算机文献检索系统。随着计算机技术与互联网的发展，信息检索系统也从批处理方式的文件检索发展到七十年代后的联机情报检索，乃至现在的大规模互联网信息检索和数字图书馆文献检索。可以说，信息检索技术已经融入我们每天的工作和生活。

1.1 研究背景

搜索引擎作为网络信息服务最基本的手段，在一定程度上可以满足用户对互联网上信息检索的要求，但由于其通用的性质，或称作为商品化软件的要求，这些通用的搜索引擎所表现的数据信息覆盖领域广、信息量大、数据不稳定、冗余度大等特性，导致用户查询的精度非常低，其效果难以满足不同背景、不同目的和不同时期的用户查询请求。缺陷主要表现在以下几个方面

(1) 适应用户兴趣变化的能力较差

现有大部分信息检索系统采用关键词输入方式进行检索。对任何用户都采用同一种模式，很容易让用户感到迷茫，有时用户也无法准确地表述自己的兴趣。尽管搜索引擎对每个用户输入的查询条件都能够返回一个按相关度排序的结果，但是由于没有考虑单个用户的查询需求，把查询条件有关的所有检索结果都返回给用户[1]，不能区分不同用户的查询意图，导致了用户对查询结果满意度的降低。

(2) 用户与检索系统的交互方式比较单调

在系统响应上，传统的搜索引擎是将有序的结果文档集合分页显示的方式进行结果反馈的，这样的响应方式一定程度上限制了用户与检索系统的交互。针对不同需求的用户，提供不同的输入方式是目前现有系统所缺少的。因此用户对检索系统的使用上无法进行个性化的操作，导致了系统对用户的查询意图理解模糊

而只能采取统一的方式进行结果反馈。

个性化信息检索技术就是针对以上问题提出的。个性化信息检索通过收集和分析个性化信息，而不是仅仅依靠检索词来判断用户的真实需求，因而能够根据用户的不同需求而返回个性化的检索结果以提高检索精度[2]。

相关反馈（Relevance Feedback）技术是通过查询后处理来实现个性化检索最常采用的方法[3]。相关反馈的提出是基于这样的经验：很少有用户能够构造出理想的查询词，也就是说用户无法用几个简单的查询词来描述自己的需要，但是如果系统把文档呈现给用户，显然用户是有能力判断其相关性的。相关反馈技术已经被证明可以有效地提高检索精度[3-4]。但是，相关反馈依赖于用户来对文档进行相关性评价，比如明确指出哪些文档含有相关信息等，根据研究[5]表明，用户往往不愿意花费时间精力来进行这样的相关性判定。

隐式反馈（Implicit Feedback）是以一种用隐式的（用户几乎察觉不到的）方式获得用户的反馈信息的方法，也就是通过用户与系统的正常交互行为来推测用户的兴趣偏好，不需要用户额外花力气去做相关性评价。研究[6]表明，隐式反馈技术虽然不如显式反馈精确，但在交互式环境中可以成为显式反馈的有效替代。实际上，最近的研究[1,7]表明如果充分利用客户端丰富的用户行为作为隐式反馈信息，甚至能比利用显式反馈取得更好的效果。因此，基于隐式反馈信息的个性化信息检索受到研究者的广泛关注。

1.2 研究动因

如上所述，基于隐式反馈的个性化信息检索是提高检索系统性能的有效技术之一。但是根据我们的了解，目前已有的相关研究中，尚有下列问题没有得到很好的解决：

1. 短期上下文的隐式反馈信息可以帮助系统实时更新用户描述，能够反映用户的 ad hoc 信息需求。这种 ad hoc 信息需求也许只是用户的一时兴起，一旦得到满足，就对这种信息再也没有兴趣了。然而，目前尚没有一个框架能够以统一的方式来开发利用短期上下文的隐式反馈信息。

2. 根据对用户行为习惯的分析，由于用户兴趣爱好和职业背景的相对稳定性，在使用信息检索系统时通常呈现出：查询词在一个比较长的时间内趋于稳定

的数据特征。因此如何结合长期上下文的个性化信息，提升用户对现有信息检索系统查询的满意度，目前还没有一个很好的解决方法。

3. 基于上下文的个性化信息检索技术研究，缺乏有效的、标准的评测数据，很难说明一个模型的优劣。如何在基于标准的数据集上进行算法的评测，目前这方面的资料还比较缺乏。

1.3 论文主要工作

本文对目前的个性化信息检索相关技术进行了较为全面、深入的阐述。根据个性化信息检索的实现方式不同，对目前的个性化信息检索研究工作进行了分类和讨论，并对一些代表性工作进行了介绍和分析。通过对现状的分析，总结了现阶段研究存在的一些不足，本文主要工作可以分为如下几点：

1. 为了克服传统信息检索系统在进行 ad hoc 查询时单独依赖于检索词的不足，本文结合统计语言模型和相关性反馈的特点，设计了一种短期上下文环境中的个性化检索算法。对该算法在 TREC AP88-90 标准语料集上做了全面的测试和对比分析，证明了该方法可以更好地描述用户的需求。

2. 传统的信息检索系统在根据长期上下文信息（例如个人的兴趣爱好、职业背景等），实现个性化信息检索上存在着一定的不足。本文在北京大学网络实验室对于天网搜索引擎用户行为数据的分析基础上，分析了长期上下文信息对提高个性化信息检索性能的可行性。本文设计了在长期上下文环境中的个性化检索算法，同时对该方法进行了实验评测。

3. 在短期上下文环境中个性化检索算法基础上，本文利用搜狗实验室提供的全网新闻数据为语料集，实现了个性化信息检索系统原型，证明了该算法在中文环境下的有效性。同时，系统客户端是基于火狐浏览器的，负责记录存储用户在使用系统的长期行为日志，为今后的研究奠定数据基础。

本文的创新点主要体现在以下几点：

1. 本文设计了短期上下文的个性化检索算法。该算法以单元统计语言模型为基础，利用查询会话中的查询历史和隐性反馈信息，采用改进了的相关反馈技术，进行短期用户模型建模。通过实验证明该算法使查询精度平均提高了 50%。

2. 本文设计了在长期上下文环境下的个性化检索算法。该算法通过分析用

户长期行为日志，采用对逐个历史查询进行单元模型建模的方法，建立起长期的个性化模型。通过实验证明在查询精度上，该算法较基准算法提高了 2.5%，也说明了该算法有较好的抗噪声性。

3. 本文设计并实现了一个基于火狐浏览器的个性化检索原型系统，证明了在中文环境下，本文设计的短期上下文的个性化检索算法的有效性。

1.4 论文组织结构

第一章 分析本文的研究背景，给出了研究动机，归纳本文的主要研究工作并介绍论文的组织结构。

第二章 对目前的个性化信息检索的相关技术进行了较为全面、深入的分析。根据所采用的个性化信息种类以及个性化检索的实现方式，对目前的个性化信息检索研究工作进行了分类和探讨，并对一些有代表性的工作进行了分析。

第三章 首先给出基于上下文的个性化检索的抽象模型和形式化描述，接着设计了短期上下文环境中的个性化检索算法。该算法以统计语言模型作为基础，结合了隐式反馈和相关反馈的技术。最后对该算法在 TREC AP88-90 的标准语料集合上进行测试，同时对算法结果进行比较和分析。

第四章 首先分析用户的查询行为习惯，接着设计了长期上下文环境中的个性化检索算法，即利用长期行为日志进行个性化模型建模。最后对该模型进行实验评测，同时对评测的结果进行分析。

第五章 按照第三章中介绍的短期上下文环境的个性化检索方法，设计了一个基于火狐浏览器的个性化检索系统原型。该系统将搜狗实验室公布的全网新闻数据作为语料集合，同时负责记录下真实的、长期的用户行为日志，为今后研究提供数据支持。

第六章 论文的总结和进一步研究方向。