基于隐性知识的信息检索多维匹配模型

阳小华，　马家宇，　刘志明，　刘振宇，　刘　杰，　蒋　辉，　李晓昀

(南华大学 计算机科学与技术学院　湖南 衡阳 421001)

摘要：将隐性知识管理理论引入信息检索，分析用户查询过程和作者写作活动中存在的隐性知识，在 3 个层面上予以定义，建立基于隐性知识的信息检索多维匹配模型. 从背景知识和行为模式等方面实现信息检索隐性维度的匹配，以解决用户查询意图和作者思想表达不充分的问题，从而提高信息检索的质量.

关键词：信息检索；隐性知识；多维匹配模型

中图分类号：TP 391　　**文献标识码**：A　　**文章编号**：1671－6841(2010)02－0014－04

0　引言

信息检索是指利用一定的检索算法，借助于特定的检索工具，针对用户的检索需求，从非结构化的文档集中找出与用户需求相关的信息[1]. 提高信息检索质量的关键在于 2 个方面：第一，如何尽可能准确地表达信息提供者的思想；第二，如何尽可能准确地获取信息需求者的意图，进而尽可能准确地在检索条件与文档索引之间实现匹配.

目前应对问题主要从知识化和个性化入手，一方面通过概念扩展、本体模型、语义 Web、自然语言理解和潜在语义分析等多种技术，揭示用户和文档表达的语义，全面理解用户的实际信息需求以及文档表达的实际内涵；另一方面采用个性化信息检索，根据用户的知识结构、信息需求、行为方式和心理倾向等，有的放矢地为具体用户实现信息检索服务，提高了半模糊查询和模糊查询的检索质量[2-3]. 经过多年研究，个性化信息检索发展出了多种个性化技术，在用户上下文、日志、Cookie 历史记录、用户提交查询和协同行为等方面提取信息，分别在不同场合取得了一定的效果[4-11]. 但是迄今为止，基于知识的语义分析技术仍然存在语义理解准确度不高、系统消耗过大等问题. 知识检索只关注用户和作者表达出来的信息，将其处理成带有语义的知识，即只处理显性知识. 可以预见，这种语义理解即便发展到完美程度，由于不涉及隐性知识，不能深入挖掘用户意图和作者思想，就不能从根本上提高信息检索的质量. 部分个性化检索技术零星地利用了用户背景、行为等隐性知识，但是由于现有查询结构是一维的，只能处理关键词等显性知识，用户背景、行为等隐性知识要转化为显性知识加入到关键词中才能得到利用. 同时，现有个性化技术缺乏对用户信息行为的全方位、全过程跟踪和理解，缺乏理论上的系统指导. 本文将隐性知识管理理论引入信息检索，建立基于隐性知识的信息检索多维匹配模型，以解决用户查询意图和作者思想表达不充分的问题，从而提高信息检索的质量.

1　用户查询过程和作者写作活动中的隐性知识

隐性知识是指那种难以言传的知识. 由于存在隐性知识，用户的查询请求不能充分地表达用户的查询意图，文档也不能完整表达作者的思想. 用户查询过程中的隐性知识可以分为 4 类：①用户提交的查询条件以关键词的形式表示，其描述的语义是显性知识. 研究表明，用户提交查询时较多使用 1 到 2 个查询词，用户可能“有意识或下意识地欠表达”他的意图，没有清楚地表达其信息需求，这是第一类隐性知识；②通过用户在使用数字系统时遗留的痕迹(数字系统内部知识发生的变化)，可以推测出用户行为的功用(目标)，该功用为

收稿日期：2009－12－01

基金项目：湖南省科技厅 2006 年计划基金资助项目，编号 2006GK3086.

作者简介：阳小华(1963－)，男，教授，博士生导师，主要从事智能信息检索与知识管理、软件分析与建模及教育信息化工程研究. E-mail：xiaohua1963@yahoo.com.cn.

信息检索所包含，使信息检索系统可以对用户行为所起的作用给出“自己的见解”，构成关于用户行为模式的知识，这是第二类隐性知识；③用户的查询请求未包含背景－视界的知识，不同的知识有不同的背景－视界，查询中“认为是理所当然的东西”没有表达出来，这是第三类隐性知识；④有些无法表达的特殊的用户智力模式则是第四类隐性知识，是用户智力行为的倾向性，由于不能编码，不存在于数字系统中．其中，第二类、第三类隐性知识利用的困难是由查询表达能力决定的，当前单维查询结构的信息检索只利用关键词代表的显性知识，查询的背景、目标、行为模式等难以被表达，可以说查询条件中只有 what，没有 why，how，background，environment 等．

再从隐性知识角度看待作者写作活动，由于存在隐性知识，文档作者的思想并不能通过文档 100％表达，文档作者的许多隐含语义不能显现．应用隐性知识理论，写作活动存在如下隐性知识分类：①作者在写作文档时，由于保密、隐私、简单省力等原因，“有意识或下意识地欠表达”他的意图，这是第一类隐性知识；②作者写作时的各种行为以及长期行为习惯产生个性化的功用，其痕迹保存在文档及相关资料内，体现出对作者行为所起的作用给出“自己的见解”，构成作者行为模式方面的知识，这是第二类隐性知识；③作者提交的文档未包含背景－视界的知识，“认为是理所当然的东西”没有表达出来，这是第三类隐性知识；④存在人类拥有的智能，但完全不能编码，这种智力倾向性则是第四类隐性知识．除第四类隐性知识以外，第一、第二、第三类隐性知识都存在于信息检索活动中．信息检索的隐性知识虽然没有被文档模型显式地表达，但用户在使用数字系统时，其遗留的行为信息和文档资料已经记录在电脑或网络中，形成了赛博空间中的隐性知识．隐性知识蕴含在用户和作者使用数字系统时遗留的痕迹中，如日志、行为记录、用户反馈、工作场景、PKM（个人知识管理）环境等，与查询和写作相关的部分构成了信息检索的隐性知识．

总之，就信息检索服务的提供者和消费者而言，信息检索系统存在 3 种类型、3 个层次的隐性知识：①“有意识或下意识欠表达”的隐性知识．文档作者和用户出于保护隐私、书写简单或其他原因，能够提供却未提供的知识；②背景－视界的隐性知识．与文档内容或查询内容有关的没有表达出来的背景和情境方面的知识；③行为模式的隐性知识．作者和用户在操作数字系统、运用知识时未明确表达的与智能相关的知识．

从理论上说，表达所有的隐性知识是不可能的，特别是第四类隐性知识，完全不能编码，不能存在于数字系统中．但是其他几类隐性知识，可能通过用户或作者的数字化活动，部分遗留在其个人信息空间中，其中包含对信息检索有用的知识．这些知识虽然不从属于信息检索系统，不能被其直接使用，但与信息检索系统具有隐性关系，构成了信息检索系统的隐性知识．对它们的挖掘利用有利于信息检索质量的提高．

2　基于隐性知识的信息检索多维匹配模型

现有信息检索的单维查询结构中，查询扩展只能沿显性知识一个维度进行，查询背景、目标、行为模式等隐性知识要转化到显性维度才能利用．在服务器端，某些技术根据用户对文档点击的情况对文档再索引，以用户提交的查询词来表征选中的文档，实际上是通过用户的查询、阅读行为发掘作者文档的隐性知识，进一步彰显文档的语义，这些做法都在一定程度上体现了对信息检索隐性知识的利用．

从宏观上看，信息检索系统的隐性知识广泛存在于个人信息空间中．个人信息管理从整体上对用户的数字信息进行了大量研究，描绘用户个人信息空间（personal information space，PIS）的特征，管理构建用户个性化的知识体系．对信息检索而言，PIS 积累了大量用户个人信息，对用户查询意图和作者文档思想的揭示具有一定价值，对信息检索质量的提高有重要作用．

目前信息检索领域只重在研究显性维度的匹配，在用户提交的查询请求到作者的文档之间实现显到显的匹配．即便采用用户反馈、用户模型、工作场景信息等个性化方法，一定程度地利用了部分隐性知识，也是先将隐性知识显性化，以文档模型表示的方式加入到查询中（查询扩展），再与文档匹配，最终在信息检索的显性维度上完成匹配．在引入隐性知识理论以后，可以发现除了这种文档模型表示的查询到文档之间显到显的匹配外，还存在查询背景和文档背景、查询模式和写作模式之间等隐性维度的直接匹配，完全可以在查询和文档显性匹配维度外，增加背景知识匹配和检索目标匹配等隐性维度，建立基于隐性知识的多维信息检索模型，见图 1．

如图 1 所示，信息检索可以从背景、目标和行为模式等多个维度进行匹配，其中背景知识匹配主要利用

第二类背景－视界的隐性知识,目标、行为模式匹配主要依赖与行为模式相关的第三类隐性知识,根据行为模式推测用户检索动机.按内容、类型等因子对PIS进行划分,对划分的PIS子空间进行排序,从可用性最大的检索子空间到同类子空间,再到工作场景、用户领域知识、长期行为模式等个人信息子空间,给出统计意义的排序.然后衡量搜索个人信息子空间的代价,计算基于子空间隐性知识的信息检索收益代价比.具体做法如下:第一,可以按个人信息子空间在信息检索中加入隐性知识,测量搜索引擎系统的查全率和查准率指标变化,标定各个子空间隐性知识对信息检索质量的贡献,并赋予测评维度权重;第二,采用显性问卷反馈,通过设计调查表和改造搜索引擎接口的方法使用户和作者显式地提交背景－视界和行为模式知识,对获得的查询和文档以及用户和作者的知识进行显式反馈;第三,进行模糊综合测评,由专家确定信息检索隐性知识范围和权重,根据查询和文档的语义以及用户和作者的行为表现对隐性知识情况进行评价,给出策略性方法和指标标准.

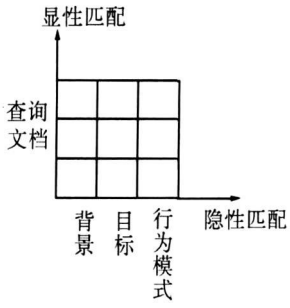

图1 信息检索的多维匹配模型

Fig·1 Multi-dimensional matching model of information retrieval

在明确信息检索系统的隐性知识的测度与可用性之后,根据信息检索多维匹配模型,计算隐性维度直接匹配对信息检索质量提高的贡献.首先,在查询用户模型之外建立文档作者模型,通过对用户提交查询返回文档的作者进行二次查询,建立文档作者模型,使关于作者的知识与用户模型中关于用户的知识进行匹配.其次,建立写作背景和查询背景模型,在爬虫程序抓取文档时搜索与文档相关的背景信息;或显式地要求作者在提交文档时,同时提交写作背景信息,对此背景－视界信息建模,分别建立与查询和文档内容相关的短期知识背景模型和与用户及作者领域知识相关的长期知识背景模型.第三,分析对比多维匹配(关键词＋背景＋目标等)与单一显性匹配(关键词查询扩展)的差别,研究搜索PIS中具有可用性的隐性知识以及表示形式转化为信息检索文档模型表示的显性知识的代价,计算检索效率提高和检索质量的变化,从而得到多维匹配结构的信息检索系统(图2).

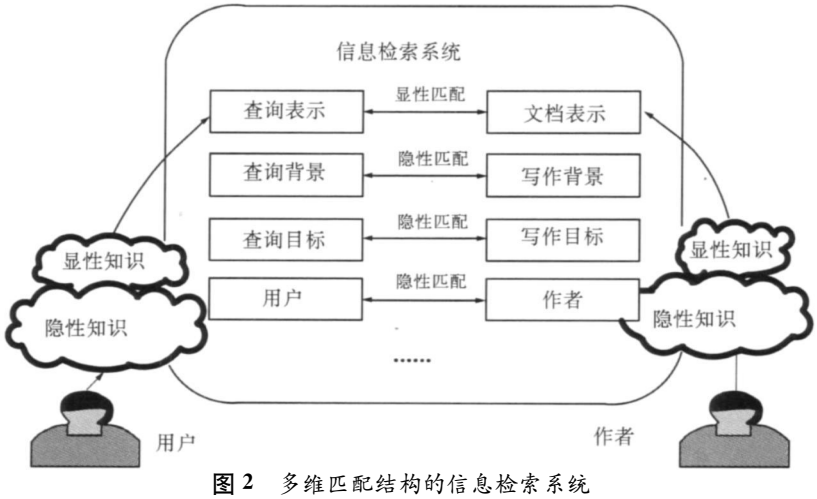

图2 多维匹配结构的信息检索系统

Fig·2 Information retrieval system with multi-dimensional matching structure

当前,单维匹配结构的信息检索系统在检索质量上存在一些无法克服的困难.比如,一名中学生学习牛顿定律时,在关键词中加入“中学”、“牛顿定律”,他的中学知识背景只能在查询表示这一显性维度提交,无法充分表达,其根本原因在于“中学”和“牛顿定律”并不在同一维度,单维匹配结构的信息检索系统不具备表达能力,查询结果中自然出现了许多“大学物理学”的条目(google),降低了检索质量.

3 小结

目前,信息检索的相关研究虽然没有明确提出对隐性知识的利用,但是个性化检索方面已经有了一些应用隐性知识提高信息检索质量的做法.由于缺乏理论指导,这些研究较为零星分散,个性化技术存在盲目性

和片面性，缺乏基于知识管理高度的、全面系统的研究．本文从知识管理的角度入手，把隐性知识理论引入信息检索，以隐性知识理论作为指导，运用隐性知识管理领域取得的理论和技术成果，解决用户查询意图和作者思想表达不充分的问题，对开展基于隐性知识的信息检索的理论研究和实际应用，具有重要的指导意义．

参考文献：

[1] 赵欣欣，索红光，刘玉树．基于改进汉宁窗的信息检索模型[J]．广西师范大学学报：自然科学版，2006，24(4)：191-194．

[2] Chirita P A，Firan C S，Nejdl W．Summarizing local context to personalize global Web search[C]// Proceedings of the 15th ACM International Conference on Information and Knowledge Management．New York：ACM Press，2006．

[3] Chirita P A，Nejdl W，Pair R，et al．Using ODP metadata to personalize search [C]//Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval．New York：ACM Press，2005．

[4] Joachims T．Optimizing search engines using clickthrough data [C]//Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining．New York：ACM Press，2002．

[5] Sugiyama K，Hatano K，Yoshikawa M．Adaptive Web search based on user profile constructed without any effort from users[C]//Proceedings of the 13th International Conference on World Wide Web．New York：ACM Press，2004．

[6] Shen Xuehua，Tan Bin，Zhai Chengxiang．Context-sensitive information retrieval using implicit feedback[C]//Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval．New York：ACM Press，2005．

[7] Teevan J，Dumais S T，Horvitz E．Beyond the commons：investigating the value of personalizing Web search[C]//Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval．New York：ACM Press，2005．

[8] Kawashige T，Oyama S，Ohshima H，et al．Context matcher：improved Web search using query term context in source document and in search results [C]// Frontiers of WWW Research and Development-APWeb．Berlin：Springer-Verlag，2006．

[9] Oliver N，Smith G，Thakkar C，et al．SWISH：semantic analysis of window titles and switching history[C]// Proceedings of the 11th International Conference on Intelligent User Interfaces．New York：ACM Press，2006．

[10] Siersdorfer S，Sizov S．Meta methods for model sharing in personal information systems[J]．ACM Transactions on Information Systems，2008，26(4)：1-34．

[11] Goncalves D，Jorge J A．In search of personal information：narrative-based interfaces[C]//Proceedings of the 13th International Conference on Intelligent User Interfaces．New York：ACM Press，2008．

Multi-dimensional Matching Model of Information Retrieval Based on Tacit Knowledge

YANG Xiao-hua，　MA Jia-yu，　LIU Zhi-ming，　LIU Zhen-yu，　LIU Jie，
JIANG Hui，　LI Xiao-yun

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

Abstract：Tacit knowledge management theory is introduced to information retrieval．The tacit knowledge existing in users' query and authors' writing activity is analyzed which is to be defined on three levels，and then multi-dimensional matching model of information retrieval based on tacit knowledge is built．The background knowledge and behavior patterns etc．belonging to implict dimension is matched to address the users' query intention and writers' ideas，and thereby the quality of information retrieval is improved．

Key words：information retrieval；tacit knowledge；multi-dimensional matching model