上下文信息检索研究综述

田 萱 李冬梅
(北京林业大学信息学院 北京 100083)

摘 要 上下文信息检索强调把有关用户、资源和查询的上下文与信息检索技术统一组织在一个整体框架内,以向用户提供最适合用户需求的检索信息。全面介绍了上下文信息检索的研究现状,概括了国内外研究者对上下文信息检索过程中涉及的上下文因素及其分类,并从用户上下文、文档上下文和系统上下文3个角度对国内外有关上下文信息检索技术的研究作了概述。最后从5个方面探讨了上下文信息检索领域存在的挑战,指出对用户检索背后需求的探究、基于语义的理解和融合上下文的信息检索模型等将是该领域目前亟需解决的问题。

关键词 上下文信息检索,上下文,个性化检索,检索模型,语义理解
中图法分类号 TP391.1 **文献标识码** A

Survey on Contextual Information Retrieval

TIAN Xuan LI Dong-mei
(School of Information Science and Technology, Beijing Forestry University, Beijing 100083, China)

Abstract With the development of IR techniques, contextual information retrieval (CIR) has been identified to be a promising direction for improving search. In CIR, the retrieval of information depends on the time and place of submitting query, history of interaction, task in hand, and many other factors that are not given explicitly but lie implicitly in the interaction and surroundings of searching, namely the context. In this paper, a survey on research work of CIR was given, and the contextual elements of CIR were summarized. In addition, contextual elements were summarized into user context, document context, and system context and the related research work where introduced. At last, the key challenges in CIR were discussed from five aspects. Exploitation of real need behind user's query, search based on semantic understanding, contextual information retrieval model and et al were pointed out as main problems needed to be resolved.

Keywords Contextual information retrieval, Context, Personalized search, Retrieval model, Semantic understanding

1 引言

信息检索的目标是"所得即所需"(What You Get Is What You Want)。一个具体的体现就是:不同的用户在使用同样查询的时候可能获得不同的结果;更进一步,同一个用户在不同时间或者不同地点发出同样的查询可能获得不同的结果。例如,同一个用户对"java"信息的需求:在工作时间希望得到有关编程语言java的相关文档,在休息时间希望得到有关java岛的旅游信息。为了达到这样的目的,检索系统需要充分地理解并掌握检索活动的主体(用户)和客体(资源)。

面对这样的挑战,人们一方面在信息资源端做工作,提出了语义网(Semantic Web,也称为语义Web)的概念[1,2],使得检索系统能够更好地理解内容,从而使检索结果更符合检索的条件;另一方面是在用户端做工作,通过各种手段获得用户的特征信息并进行用户建模,使用用户个性化信息来修正查询条件,从而改善检索结果。这两个方面的研究对达到"所得即所需"的目标起到了很大的推动作用。

尽管语义网和用户建模技术极大地提高了检索系统的智能化、个性化水平[3,4],但是,人们也已经意识到,将资源和用户分开来考虑,难以达到"所得即所需"的目标。必须用系统的观点来看待信息检索活动,也就是说,用户检索的结果应该是特定"环境"下的结果,这个环境就是检索过程的上下文(Context)。考虑了上下文的检索称为上下文信息检索(Contextual Information Retrieval,CIR)。

2 上下文信息检索的概念

WordNet是Princeton大学的心理学家、语言学家和计算机工程师联合设计的一种基于认知语言学的英语词典[5]。在WordNet2.1中,上下文(context)被定义如下:

1) 语言学上下文,即在一个语言单位附近的片断,用以帮助解释该语言单位。

2) 环境,即一种情形或事件发生于其中的环境和背景。

信息检索领域中,上下文最初是指"自然语言处理中的文档片段",专门用于自然语言学中指代短语或句子在实际应

到稿日期:2011-03-20 返修日期:2011-05-26 本文受北京林业大学新进教师科研启动基金(BLX2w8019),中央高校基本科研业务费专项资金(YX2011-30)资助。

田 萱(1976-),女,博士,讲师,主要研究方向为上下文信息检索、知识工程,E-mail:tianxuan@bjfu.edu.cn;李冬梅(1973-),女,博士生,讲师,主要研究方向为知识工程、信息系统。

用中的语言环境。它在自然语言处理中的价值体现在两个方面：一方面，在自然语言知识获取的过程中，上下文是知识获取的来源，在相应推理机制下，上下文本身就是知识；另一方面，在自然语言处理的应用问题解决过程中，上下文扮演着解决问题所需信息和资源提供者的重要角色。

从20世纪中期开始发展的信息检索系统，基本上是千人一面(one size fits all)，不同用户提出同一查询，得到的答案完全相同。这种模式带来的最大问题就是不够人性化，难以准确地满足不同用户的个性化需求。所以，人们最先关注的是和用户有关的上下文，即把用户有关的信息引入检索系统以满足用户的"所得即所需"。文献[6]于2000年总结的Web搜索中的上下文信息主要包括和用户查询意图以及用户查询表达相关的信息。文献[7]也指出上下文和个性化检索紧密相关，用来帮助提高用户检索体验，需要理解每一个用户查找信息的模式习惯、用户目标，以及信息本身。

然而，对信息检索系统而言，可利用的上下文并不仅限于此。2002年9月在Massachusetts Amherst大学智能信息检索中心(the Center for Intelligent Information Retrieval)召开的关于智能信息检索未来研究方向和发展的研讨会上，许多信息检索领域顶级研究者经过讨论给出了上下文信息检索定义[8]，即：

定义1(上下文信息检索，Contextual Information Retrieval，CIR) CIR就是把有关用户、查询的上下文知识和信息检索技术融合在一起，统一组织在一个整体框架内，以向用户提供最适合用户需求的检索信息。

随着人们对CIR的关注，2003年第12届TREC(Text Retrieval Conference)国际会议第一次增加了HARD评测(High Accuracy Retrieval from Documents track)。HARD评测的目的是考察用户及其相关信息对检索过程和检索结果评估的影响，即考察信息检索过程中上下文(如用户地域特点、文档风格等上下文信息)对信息检索性能的影响。

2004年第1届IRiX(Information Retrieval in Context)研讨会在第27届SIGIR上举行，并一举成为SIGIR2004上参会人数最多、最受人关注的研讨会 。该研讨会的总目标是如何在信息检索过程中考虑上下文因素以提高用户信息需求满意度。在该研讨会上，信息检索领域中的上下文定义如下。

定义2(上下文，Context) 信息检索中的上下文包括一切与检索查询相关的任务信息、交互历史信息、用户信息等明确给出或隐含在检索交互环境中的相关信息。

从定义2中可以看出，只要和用户检索过程相关的一切隐含或明确的信息都将是智能个性化信息检索的上下文，都可能用于优化检索系统，提高检索性能。因此，智能信息检索的上下文实际上是无所不在，无处不在。

事实上，从20世纪90年代后期以来，围绕信息检索、信息推荐等信息服务系统的上下文的研究就层出不穷，有许多研究成果已经成功运用在实际系统中来帮助提高效率和性能，如针对用户兴趣的相关反馈技术[9-13]、针对Web文档链接内容的PageRank技术[14,15]、针对用户访问历史记录的Web日志分析技术[16,17]等，并取得了一系列重要成果。这为人们进一步挖掘可用上下文以帮助提高检索效果树立了信心，指明了方向。在2007年欧洲信息检索大会上(European Conference on Information Retrieval，ECIR)，Yahoo公司新兴搜索技术(Emerging Search Technology)部门的Andrei Broder指出上下文信息不仅在当前第三代搜索引擎实现满足"查询背后的需求"(the need behind the query)目标中起着关键作用，更在未来第四代搜索引擎实现"上下文驱动的信息推送"(context driven information supply)目标中占据主导地位[18]。

3 CIR中的上下文因素及其分类

3.1 学术界的观点

Peter Ingwersen等人把信息检索系统中涉及的上下文因素抽象概括为六大因素，表示为一个上下文分层嵌套模型(Nested model of context stratification for IR)[19]，如图1所示。该模型作者认为，传统信息检索技术更多的是关注检索对象本身以及检索对象之间的特征，如词语、段落以及文档内容的超级链接等；如今，信息检索系统的上下文技术开始转向用户检索对话过程中(session-time)可获取的上下文信息，如鼠标移动、打印保存等操作，即转向从交互式过程中获取用户的上下文信息。

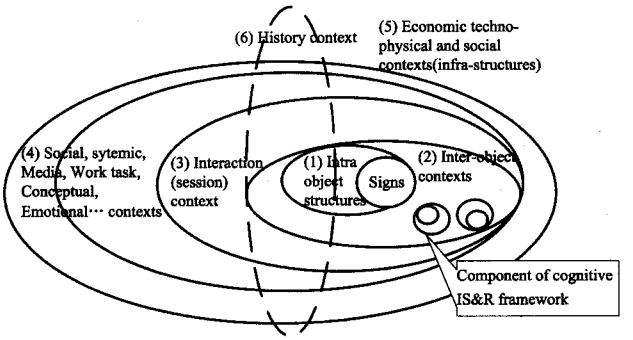

图1 Peter Ingwersen等人提出的分层嵌套的上下文模型

2005年IRiX(Information Retrieval in Context)研讨会上研究者则把信息检索中的上下文看作是包含了信息检索过程中涉及的各种因素的超类，把各种因素不同取值之间的组合看作情景(Situation)，把每种因素的取值可能性看作是任务(Task)。上下文中包含的因素主要是3个方面的，分别是系统、用户和环境。其中每一方面的因素又包含多种因素，如用户方面包括动机(Motivation)、知识(Knowledge)、历史(History)和个体差异(Individual differences)等，系统方面包括资源(Resource)、检索模型(Retrieval Model)、设备(Device)、接口(Interface)等方面。

3.2 产业界的观点

Andrei Broder在2007年欧洲信息检索大会上强调了当前和未来上下文信息在信息检索过程中的重要性，指出当前第三代检索技术是依赖上下文信息满足"查询背后的需求"(the need behind the query)，并提出未来第四代检索技术需要实现"上下文驱动的信息推送"(context driven information supply)[18]。同时，他指出第三代搜索引擎中上下文中的决定因素(Context Determination)包括空间信息(如user location/target location)、查询信息(如previous queries)、个人信息(如user profile)、明确信息(如user choice of a vertical search)以及潜在信息(如use Google from China，use google. cn)等5种。

除了上面学术界给出的阐述外，产业界给出了更为实用

的基于上下文的信息检索的说明[20,21]。他们把基于上下文的信息检索看作是由信息检索领域3种技术构成的三维空间上的一个平面。这3种技术保障了对上下文信息的获取和挖掘，如图2所示。这3种技术分别是：

1)智能的文本挖掘和数据挖掘，通过自动文本概念标注、模式发现和实体知识识别等技术发现各种可用的信息；

2)灵活的内容构建技术，能从结构化或半结构化的数据源中发现独立的XML模式和相关关联；

3)高性能的检索技术，面对超大规模的数据能进行迅速和可扩展的内容处理和检索。

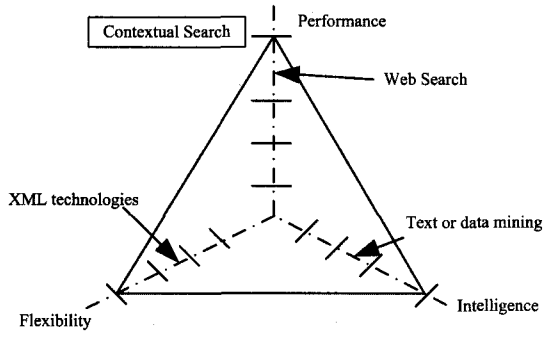

图2 产业界关于上下文检索的一种观点

3.3 本文的分类观点

纵观上面的讨论，Peter Ingwersen等人的看法层次分明，抽象意义明显；2005年IRiX研讨会上给出的结论比较系统、清楚自然，更便于在实现过程中区分和理解；Andrei Broder给出的上下文更符合Internet上搜索引擎环境下的应用；而产业界则在技术层面上给出了挖掘应用上下文因素的相关分析。

结合以上讨论和上下文信息在信息检索领域已有的研究成果，本文把人们当前比较关注的上下文因素按照信息检索的逻辑流程分为3个类别，如图3所示，分别是用户上下文、文档上下文和系统上下文。

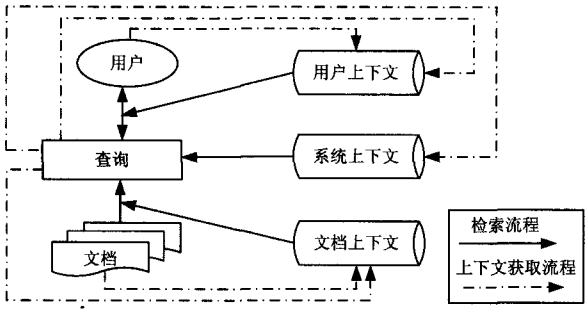

图3 本文对检索系统中上下文因素的分类

1)用户上下文：检索系统中围绕用户的上下文信息，如用户的兴趣、爱好等，以及用户的查询日志、检索反馈行为等。用户上下文提供了理解用户需求的信息，是现在实现个性化检索及未来实现上下文驱动的信息推送服务的必要条件之一。为了便于详细解析和用户有关的上下文信息，本文把用户上下文分为用户静态上下文和检索任务上下文两个类别。用户静态上下文和用户的专业背景、工作内容、爱好、经验、生活习惯、理解水平等因素密切相关；检索任务上下文包括检索任务的内容、特点、发生时间、作用范围、发生背景等方面，一方面反映用户本次信息需求的内容，另一方面反映用户检索需求的变化和迁移。用户静态上下文描述了用户检索需求的一种长期特点，是用户长期检索兴趣的反映；检索任务上下文描述了用户检索需求的一种短期特点，是用户短期检索需求的反映。

2)文档上下文：文档是指信息检索的目标对象。文档的内容特征、使用范围、产生的时间和地点以及其它元数据(meta data)信息等都属于文档上下文。文档内容特征包括文档的书写语言、术语化程度(专业性程度)、布局特点等因素。除了文档内容外，对Web网页而言，超链分析(link analysis)、布局分析(block analysis)可以较准确地挖掘出其特征；对Pdf、Word等文档而言，元数据分析、布局分析能更有效地发现其特点。另外，整个文档集的组织结构、文档之间的关系等独立于单个文档之外的信息也属于文档上下文的范畴[22]。

3)系统上下文：反映了信息检索系统实现过程中的相关特征，如采用的索引机制、检索模型、检索界面等。

4 CIR研究现状

4.1 用户上下文

4.1.1 用户静态上下文

用户静态上下文包括用户的专业背景、工作内容、爱好、经验、生活习惯、理解水平等各种和用户个体相关的许多因素，用户建模(user modeling)就是对用户上下文中的因素进行模型表示。当前，研究者比较关注对用户上下文中用户认知特点(cognitive characteristics)的建模，如兴趣、技能、偏好等[23,24]。随着近年来语义Web(Semantic Web)和本体(Ontology)技术的发展，许多研究纷纷以本体为工具来分析和描述用户上下文。文献[25]在具有层次关系的轻量级本体ODP(dmoz Open Directory Project)上对用户查询兴趣进行扩展：把用户兴趣归纳到ODP上的不同类别上，把用户对某个类别下的实例兴趣度的50%加到其父类别上，达到由下层到上层对用户兴趣进行扩展的目的。文献[26]提出基于Lycos的目录层次结构构建一个表示用户兴趣的个性化层次树，以帮助实现Web的个性化浏览。文献[27]把用户的研究兴趣建立在ODP之上，并通过计算搜索结果网页所属类别和用户兴趣所属类别之间的语义距离来实现个性化的检索。这些研究使用的本体大都集中在一些大型的通用本体，所利用的语义关系大都是父子关系，还缺乏对本体信息更充分的利用和进一步挖掘，如本体上概念之间的关联关系、建立在更细粒度上(如领域本体)的分析等。

信息检索系统中常通过用户描述文件(user profile)为每个用户刻画其用户特征。用户描述文件可以表示成加权向量模型、层次结构模型、加权语义网模型、书签和目录结构等，存储时可以采用纯文本文件、XML文件、关系数据库、XML数据库等各种形式[3,28]。

获取用户上下文最为直接简单的方法就是由用户自我提供确认。系统可以在用户注册该系统时获取相关的用户上下文信息，如年龄、专业、兴趣等。NEC研究所(NEC Research Institute)著名的Inquirus-2项目[6,29]就是通过用户手工选择查询类别来获取相应的用户上下文信息；Google Personal也是由用户选择兴趣所属类别来创建用户描述文件的。然而，通过许多研究调查表明人工方式获得的用户上下文并不很准确，原因在于大部分用户不愿意花费精力认真准确地填写自

己的相关信息。

针对许多系统并不能获得用户准确上下文信息的问题，人们提出许多自动获取方法来获得用户静态上下文，如相关反馈(Relevant Feedback，RF)[13]、机器学习(Machine Learning，ML)、数据挖掘(Data Mining，DM)等。这些方法通过对用户操作历史、用户访问过的资源特征、用户访问日志等信息的统计分析来获取某方面的用户上下文，帮助创建用户描述文件。例如，文献[30]介绍了使用关联规则挖掘用户的 Web 日志以构建用户的个性化描述文件；WY. Men 等人[31]提出根据用户的点击历史自动把用户兴趣定位到 Yahoo 的某个类别层次上，从而确定用户的个性化信息；文献[32]通过增量式文本挖掘方式(incremental text mining)发现用户兴趣。为了获取更准确的用户上下文信息，这些自动方法或者需要长期用户的检索任务上下文信息，或者需要与用户上下文人工获取方式结合起来。

4.1.2 检索任务上下文

检索任务上下文提供围绕用户检索目的的上下文信息，包括检索任务的内容特点、发生时间、发生地点(IP 地址)、作用范围、发生背景(客户端背景)、用户的反馈信息、响应操作等许多因素。根据这些因素的变化性和复杂性，我们把检索任务上下文分为简单因素和复杂因素两类，两类因素比较如表 1 所列。简单因素包括用户提交检索时可以获得的一次性数据，这类数据在用户的一次查询过程中基本没有变化，如检索内容、发生时间、发生地点、发生背景等因素。根据这些简单因素，结合文档集的特点，检索系统可以在第一次返回检索结果时向用户提供更适合其需求的文档。例如：根据发出查询用户的 IP 地址，搜索引擎可以判断用户的使用语言偏好以向用户提供适合用户语言阅读的检索结果；根据用户发出查询的时间，例如是在普通工作时间内还是在休假时间内，搜索引擎可以把用户更满意的结果排在前面。Yahoo 公司的基于上下文的信息检索工具 Y! Q[1] 首先做到的就是从用户在 Web 页面上选取的上下文中识别出用户需求并把相关页面返回给用户[33]。复杂因素是指和用户进行该检索时对检索过程和检索结果的响应相关的因素，和简单因素相比，这类数据动态不定，如用户的反馈信息、响应操作、查询持续时间等。根据复杂因素，系统可以将反馈结果应用在用户查看下一页的结果排列上。Steve Fox 把复杂因素又划分为结果级别(Result-Level)和会话级别(Session-Level)[34]，并详细列出了每一级别上更为细致的因素。

表 1 检索任务上下文中的简单因素和复杂因素的比较

	简单因素	复杂因素
共同点	都是围绕用户检索任务的上下文信息	
区别	在用户的一次查询过程中基本没有变化	和用户在检索过程中的响应行为有关，动态不定
具体因素包括	检索内容、发生时间、发生地点、发生背景等	用户的反馈信息、响应操作、查询持续时间等

复杂因素与用户在检索过程中的响应行为有关，而相关反馈技术是获得这些因素最为重要的方法之一。相关反馈技术不仅在创建用户描述文件时可用于获取用户的兴趣和偏好，对即时提高检索性能、满足用户短期查询也有很好的效果[13]。相关反馈分为明确反馈(explicit feedback)、伪反馈(pseudo feedback)、潜在反馈(implicit feedback)3 种形式。

明确反馈是指由用户明确给出是否满意检索结果的评价。由于大部分用户在检索过程中不愿主动参与，因此在 Web 检索系统中单独应用较少；即使有用户主动参与了明确反馈，效果也不是很好[35]。

伪反馈是一种没有用户参与的方法，它假设第一次检索结果中 Top-N 篇文档正是用户所需，并把这种假设的反馈信息通过查询扩展(query expansion)技术调整新查询的结果排序[36]。伪反馈中可提取出许多信息，如段落和概念等，用于优化检索性能，提取出真正有助于增强了解用户个性化的上下文信息将更具有意义。伪反馈是应用较多的一种方法，但它基于的前提假设“Top-N 篇文档与用户所需相关”值得进一步关注。文献[37]曾通过实验发现伪反馈的效果受 N 值的影响较大，因此提出了两阶段混合模型的解决方法。

潜在反馈是指在用户检索和浏览检索结果的过程中由检索系统自动收集有关用户响应行为的反馈信息，并把反馈信息及时应用到当次检索结果的优化调整上。潜在反馈由于具有不需用户主动配合、能即时修正检索结果的优点，因此成为当前研究领域获取检索任务上下文最主要的方法。也有人对潜在反馈的效果存有疑虑，但研究[28]表明通过潜在结构化的个性化信息进行的个性化 Web 检索性能要比明确反馈信息的效果好，文献[38]也得出了类似的结论，并且通过进一步研究表明在越复杂的检索任务中，潜在反馈的效果越明显。随着人们对检索任务上下文内容更细致的挖掘应用，针对检索任务上下文的潜在反馈模型也成为最近研究的重点，如文献[39]针对用户的点击流(clickthrough)信息提出一种基于决策理论的潜在反馈模型；文献[40]针对用户与 Top-N 文档交互的上下文信息提出一种基于启发式的二元投票模型(Binary Voting Model)。

4.2 资源上下文

超链分析技术主要针对 Web 文档中的超级链接(hypertext)信息，早期曾在 Lawrence Page 和 Sergey Brin 等提出的 PageRank 算法中实现[41]。考虑到重要的文档会有更多的链接指向它，PageRank 算法从文档页面上的进链(backward link)和出链(forward link)数量出发计算每个页面的权重。近年来，人们又提出了面向主题(Topic-sensitive)的 PageRank算法[42,43]和基于 PPV(Personalized PageRank Vector)的个性化 PageRank 算法[44]，这些算法都是在原有 PageRank 的基础上增加了主题特征、用户偏好等其它上下文因素来计算页面的权重。除了 PageRank 算法，Kleinberg[45] 提出的 HITS(Hypertext Induced Topic Search)算法也是超链分析技术中的一个重要算法，与 PageRank 的全局平均思想不同，HITS 算法针对一个查询请求分析权威页面(Authority)和枢轴(Hub)页面来计算页面的重要程度。然而 HITS 算法还是单纯从文档中的超级链接出发，忽略了文档中的其它因素。结合语义[46]等其它上下文因素，又有许多改进算法如 SALSA 算法、PHITS 算法等，见文献[47]等。

除了文档中的文本、超级链接等信息，文档的布局也是反映文档特征的一个重要方面。有许多算法研究文档如何分块，见文献[48,49]等，这些文献大都从视觉位置、内容模式方

[1] http://yq.search.yahoo.com/

面着手；也有一些研究专门从文档分块角度研究特征，如分块的重要性、分块的吸引性、分块的语义性、分块的指向性等；这些研究在分析分块特征时常常依据的是块中词语的熵信息[48]，或者根据链接的统计信息[50]，或者根据分块中的语义信息。布局分析的一个重要意义在于充分挖掘文档特征，以用于提高信息检索性能。

4.3 系统上下文

信息检索系统所采用的检索模型是系统上下文中关键的一种。信息检索领域中经典的3种检索模型分别是布尔模型、向量模型和概率模型，它们分别基于集合论、代数论和Bayesian概率论。布尔模型基于简单的关键词匹配但检索效果很差；向量模型虽然提供了更好的改进但缺乏一个规范的框架；Bayesian概率论最大的优势在于提供了一个完整的框架以便人们把检索中的各种因素组合在一起考虑[51]。各种模型及其相应的模型扩展在文献[51]中介绍得比较详细，本文不再一一列举。

检索系统中检索界面决定了人机交互(human-computer interaction)的内容，和检索系统中的其它上下文信息配合使用，对实现智能个性化检索非常关键。检索界面主要包括接受用户的查询输入和显示结果两个部分。

对查询输入界面，一方面可以在布局设计上考虑满足不同用户的使用偏好和习惯，另一方面可以在功能上考虑向用户及时推送其感兴趣的检索信息。如当前的Google，一方面针对不同地区用户自动推出不同的语言版本以适应用户的语言习惯，另一方面结合世界新闻事件不断更换其标志图案(logo)以向用户推送最新消息。

根据我们使用Web搜索引擎的经验，结果显示界面往往是把从海量信息中筛选出的大量信息显示给用户，因此除了检索性能，结果显示界面的设计常常影响用户对该检索系统是否偏好。好的结果显示界面一方面在布局上要简洁清晰、便于浏览查看，另一方面在功能上还能帮助用户理解个性化的检索结果、提高用户的检索效率。例如，在检索结果列表中加入准确的文档摘要信息，高亮度显示影响文档排序的关键词，按类别显示文档列表等都是比较有效的方法。

除了接受查询部分和显示结果部分，Jaime Teevan认为增加个性化参数控制(control over key personalized parameters)部分也非常重要[28]。虽然这部分功能用户可能较少使用，但提供给用户简易快捷的调整功能还是可以帮助用户获得更加满意的检索效果的。

另外，我们把独立于检索系统之外的社会环境也看作系统上下文的一部分。这些上下文是指隐藏在社会生活、国际背景和文化趋势中的一些外界常规或突发信息。拥有及时社会环境的信息检索系统可以向用户提供更准确更及时的信息。这类上下文有两种方式可以获得，一种是人工收集，另一种是系统自动收集。人工收集是由工作人员根据现实生活，人工收集这类上下文；系统自动收集是指对所有用户检索日志进行统计分析及对比比较，发现这类上下文。两种方法相比而言，人工收集方式具有响应速度快、准确率高等特点，而系统自动收集往往能发现潜在的社会环境信息，从而更易于满足大部分人群潜在的检索需求。如谷歌搜索引擎[2]能自动向用户提供和用户输入字面最接近的、最常出现的前10个查询，而网易[3]总是把系统统计得出的热门搜索显示在其主页面上，以向用户传递人们当前最为关心的信息。

5 CIR研究面临的挑战

尽管自从上个世纪五六十年代开始的信息检索技术研究历史已达半个多世纪，并且已经发展到当前依赖上下文信息满足“查询背后的需求”的第三代检索技术；尽管各地研究者在理解用户个性化兴趣、解析文档特征、发展不断具有适应性的检索模型等方面进行了多角度研究，并不断从机器学习、人工智能、自然语言处理、数据库系统、数据挖掘等领域借鉴方法和思路；但是面对不断出现的超大规模在线数据，面对快速发展的语义网资源，面对用户对检索效果越来越高的检索要求，基于上下文的信息检索技术还面临着多重挑战。

1)深入理解用户需求并建模

无论是当前第三代依赖上下文信息满足“查询背后的需求”的检索技术，还是未来第四代实现“上下文驱动的信息推送”(context driven information supply)技术[52]，只有深入理解用户个性化的需求才能达到真正使用户满意。虽然有许多研究在用户静态上下文和检索任务上下文领域进行了有效的尝试，但如何深入理解用户需求有待进一步研究，尤其在用户需求背景比较复杂、需求周期变换不定、需求形式多样化的情况下。文献[53]曾把用户上网搜索的需求形式分成3类：对信息的需求(例如：找新闻、找评论、找帖子等)、对导航的需求(找某个特定网站)和对交易的需求(例如：下载软件、在线购物、订机票等)。然而，面对不断增长的海量数据，我们还需要在理解用户的个体需求背景下加深用户需求动机的分析，例如了解用户已有的和查询相关的知识背景等。当然，在深入理解用户需求的同时也可能会带来暴露隐私的危险性，这也是个性化信息检索中一直期待研究的重要问题。

2)加强语义理解

语义网描述了信息资源的语义数据模型，提供了计算机理解内容的基础。随着语义Web和本体技术的发展，大家普遍认为按照本体标注和组织资源可以方便计算机之间基于语义的交换和处理[54]。当前的检索系统虽然大量利用了文档上下文信息，特别是文档内容中的信息如超链接、标签、文档视觉形式以及其它各种形式的元数据类别等信息。但是，本质上这些方法仍然还是靠句法结构，基本上是用单词来匹配文本，缺乏对文档含义的真正理解。如何适应语义网上的处理方式以实现和语义网资源和服务的无缝连接以及如何深入理解文档含义和用户需求的含义都需要深入的语义分析。已有的研究在利用大规模通用本体如WordNet、ODP等上面取得了提高，但面向领域、粒度细致的语义分析和改进仍需深入研究。

3)提供融合上下文的检索模型

检索模型是检索系统的核心算法，信息检索领域中已经成功发展了向量空间模型、概率模型和统计语言模型等3种经典模型，并且还出现了新型的检索模型如基于引力的检索

[2] www.google.com

[3] www.163.com

模型(Gravitation-based model)[55]等。对经典的向量空间模型而言,虽然已经有潜在语义分析(Latent Semantic Analysis)、向量空间基(vector space bases)等方法把文档的上下文融入到向量空间模型中,但如何把各种上下文信息合理地融入到检索模型中的研究还不多见。近十年来统计语言模型是被强烈看好的一种支持融入上下文信息的检索模型。对统计语言模型而言,线性插值法是研究过程中常用的方法之一。然而线性插值项的系数并不是一件容易确定的事情,特别是在涉及到多种不同类型的上下文信息时。因此,对如何把各类不同上下文信息合理地融入到检索模型而言,还有许多值得研究的问题[56]。

4)CIR 标准测试数据集和基准测试查询

众所周知,TREC 会议上的测试数据集已成为信息检索领域公认的标准数据集。其中的 HARD 评测上也专门提供了考察像用户位置、文档风格、文档语言等上下文信息的标准评测数据[57,58]。但信息检索过程中包含用户、文档、系统等不同种类的上下文,在 TREC 标准数据集中加入更多标准化的不同类型的上下文信息,特别是便于语义理解的上下文信息如提供标准的领域本体,对考察 CIR 查询效果的影响具有重要意义。

5)由被动要求转为主动推送

事实上,"Contextual Information Retrieval"具有一语双关的含义,可以具有两种解释形式,分别是基于上下文的信息检索(retrieval determined by context)和上下文中的信息检索(retrieval determined in context)。这两种解释分别代表着当前第三代信息检索技术"满足查询背后的需求"和未来第四代信息检索技术"上下文驱动的信息推送"的研究方向。虽然现在已有像 RSS(Really Simple Syndication)这种在线订阅推送服务,但这并不是一种根据用户上下文信息主动变化而适时推送的服务,距离真正的主动推送服务还有很大差距。深入了解并理解用户所处的上下文环境,并充分利用系统、资源等上下文信息是未来实现向用户主动推送信息的前提条件。第三代信息检索技术的发展将为第四代技术奠定坚实基础。

结束语 本文全面介绍了上下文信息检索的研究现状,概括了国内外研究者对上下文信息检索过程中涉及的上下文因素及其分类,并从用户上下文、文档上下文和系统上下文 3 个角度对国内外有关上下文信息检索技术的研究作了概述。分析探讨了 CIR 研究领域面临的挑战,指出对用户检索背后需求的探究、基于语义的理解、创建融合上下文的检索模型等将是上下文信息检索领域面临的一些问题。

参 考 文 献

[1] Berners-Lee T, Hendler J. Publishing On The Semantic Web - the Coming Internet Revolution Will Profoundly Affect Scientific Information[J]. Nature, 2001, 410(6832): 1023-1024

[2] Berners-Lee T, Hendler J, Lassila O. The Semantic Web-A New form of Web Content That is Meaningful to Computers Will Unleash a Revolution of New Possibilities[J]. Scientific American, 2001, 284(5): 34-43

[3] Dou Z, Song R, Wen J-R, et al. Evaluating the Effectiveness of Personalized Web Search[J]. IEEE Trans Knowl Data Eng, 2009: 1178-1190

[4] Sheng Q, Shi Z. A knowledge-based data model and query algebra for the next-generation web[J]. Advanced Web Technologies And Applications, 2004, 3007: 489-499

[5] WordNet Homepage[EB/OL]. http://wordnet.princeton.edu/

[6] Lawrence S. Context in Web Search[J]. IEEE Data Engineering Bulletin, 2000, 23(3): 25-32

[7] Pitkow J, Schutze H, Cass T, et al. Personalized search[J]. Communications of the Acm, 2002, 45(9): 50-55

[8] Allan J, Aslam J, Belkin N, et al. Challenges in information retrieval and language modeling: report of a workshop held at the center for intelligent information retrieval[J]. SIGIR Forum, 2003, 37(1): 31-47

[9] Hoi S C H, Lyu M R, Jin R. A unified log-based relevance feedback scheme for image retrieval[J]. IEEE Transactions on Knowedge and Data Engineering, 2006, 18(4): 509-524

[10] White R, Ruthven I, Jose J, et al. Evaluating implicit feedback models using searcher simulations[J]. ACM Transaction on Information Systems, 2005, 23(3): 325-361

[11] Shen X, Tan B, Zhai C. Context-sensitive information retrieval using implicit feedback[C]// the Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Salvador, Brazil, ACM Press, 2005

[12] Ruthven I A N, Lalmas M. A survey on the use of relevance feedback for information access systems[J]. The Knowledge Engineering Review, 2003, 18(2): 95-145

[13] Kelly D, Teevan J. Implicit Feedback for Inferring User Preference: A Bibliography [J]. SIGIR Forum, 2003, 37(2): 18-28

[14] Soumen C. Dynamic personalized pagerank in entity-relation graphs[C]// the Proceedings of the 16th international conference on World Wide Web. Banff, Alberta, Canada, ACM, 2007

[15] Haveliwala T H. Topic-sensitive PageRank: A context-sensitive ranking algorithm for Web search[J]. IEEE Transactions on Knowledge and Data Engineering, 2003, 15(4): 784-796

[16] Yang Q, Zhang HH. Web-log mining for predictive Web caching [J]. IEEE Transactions on Knowledge and Data Engineering, 2003, 15(4): 1050-1053

[17] 崔航,文继荣,李敏强. 基于用户日志的查询扩展统计模型[J]. 软件学报, 2003, 14(9): 1593-1599

[18] Broder A Z. The Next Generation Web Search and the Demise of the Classic IR Model[C]// the Proceedings of the ECIR. 2007

[19] Ingwersen P, Belkin N. Information retrieval in context [J]. ACM SIGIR Forum, 2004, 2

[20] Aleksander H R N. Contextual insight in search: enabling technologies and applications[C]// the Proceedings of the 31st international conference on very large data bases. Trondheim, Norway. VLDB Endowment, 2005

[21] Olstad B, Seres S. Contextual Search[J]. Supplement to KM-World, 2005(November/December): 10-11

[22] Thirunarayan K. On embedding machine-processable semantics into documents[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(7): 1014-1018

[23] Ioannidis Y, Koutrika G. Personalized Systems: Models and Methods from an IR and DB perspective[C]// the Proceedings of 31st VLDB Conference. Trondheim Norway, 2005

[24] Wen J, Dou Z, Song R. Personalized Web Search[J]. Encyclopedia of Database Systems, 2009: 2099-2103

[25] Middleton S, Shadbolt N, De Roure D. Ontological user profiling in recommender systems[J]. ACM Transactions on Information

Systems,2004,22(1):54-88

[26] Chaffee J,Gauch S. Personal Ontologies for Web Navigation[C]// Proceedings of the 14th International Conference on Information and Knowledge Management(CIKM2005). McLean,Va. ,USA, 2000

[27] Chirita PA,Nejdl W,Paiu R,et al. Using ODP Metadata to Personalize Search[C]// the Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval. Salvador,Brazil,ACM Press,2005

[28] Teevan J,Dumais S T,Horvitz E. Personalizing search via automated analysis of interests and activities[C]// the Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval. Salvador,Brazil,2005

[29] Glover E,Lawrence S,Gordon M,et al. Web search-Your way [J]. Communications of the Acm,2001,44(12):97-102

[30] Mobasher B,Cooley R,Srivastava J. Automatic personalization based on Web usage mining- Web usage mining can help improve the scalability,accuracy,and flexibility of recommender systems[J]. Communications of the Acm,2000,43(8):142-151

[31] Liu F,Yu C,Meng W. Personalized web search for improving retrieval effectiveness[J]. IEEE Transactions on Knowledge and Data Engineering,2004,16(1):28-40

[32] Liu R L,Lin W J. Incremental mining of information interest for personalized web scanning[J]. Information Systems,2005,30 (8):630-648

[33] Kraft R,Maghoul F,Chang C C. Y! Q:Contextual Search at the Point of Inspiration[C]// the Proceedings of the 14 ACM CIKM International Conference on Information and Knowledge Management(CIKM2005). Bremen,Germany,ACM Press,2005

[34] Fox S,Karnawat K,Mydland M,et al. Evaluating implicit measures to improve web search[J]. ACM Transactions on Information Systems,2005,23(2):147-168

[35] Anick P. Using terminological feedback for Web search refinement:a log-based study[C]// the Proceedings of the 26th annual international ACM SIGIR conference on research and development in informaion retrieval. Toronto,Canada,ACM Press,2003

[36] Singhal A. Modern Information Retrieval:A Brief Overview[J]. IEEE Data Engineering Bulletin,2001,24(4):35-43

[37] Tao T,Zhai C. A two-stage mixture model for pseudo feedback [C]// the Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. Sheffield,United Kingdom,ACM Press,2004

[38] White R W,Ruthven I,Jose J M. A study of factors affecting the utility of implicit relevance feedback[C]// the Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Salvador,Brazil, ACM Press,2005

[39] Shen X,Tan B,Zhai C. Implicit User Modeling for Personalized Search[C]// the Proceedings of the CIKM'05. Bremen,Germany,2005

[40] White R W,Jose J M,Ruthven I. An Implicit Feedback Approach for Interactive Information Retrieval[J]. Information Processing and Management(IP&M),2004,42(1):166-190

[41] Brin S,Page L. The Anatomy of a Large-Scale Hypertextual Web Search Engine[C]// Proceedings of the World Wide Web Conference. Brisbane,Australia,1998

[42] Haveliwala T H. Topic-sensitive pagerank[C]// the Proceedings of the 11th International World Wide Web Conference. Honolulu,Hawaii,USA,2002

[43] Haveliwala T. Topic-sensitive pagerank[C]// the Proceedings of the 11th International World Wide Web Conference. Budapest, Hungary,2002

[44] Jeh G,Widom J. Scaling Personalized Web Search[C]// the Proceedings of the 12th International World Wide Web Conference. Budapest,Hungary,2003

[45] Kleinberg J M. Hubs, authorities, and communities [J]. Acm Computing Surveys,1999,31(4):1-3

[46] Halkidi M,Nguyen B,Varlamis I,et al. THESUS:Organizing Web document collections based on link semantics[J]. Vldb Journal,2003,12(4):320-332

[47] Borodin A,Roberts G O,Rosenthal J S,et al. Finding Authorities and Hubs From Link Structures on the World Wide Web [C]// the Proceedings of the Tenth International World Wide Web Conference(WWW 10). Hong Kong,China,ACM Press, 2010

[48] Chen J,Zhou B,Shi J,et al. Function-based object model towards website adaptation[C]// the Proceedings of the Tenth International World Wide Web Conference (WWW 10). Hong Kong,China,ACM Press,2001

[49] Yu S,Cai D,Wen J-R,et al. Improving pseudo-relevance feedback in web information retrieval using Web page segmentation [C]// the Proceedings of the Twelfth International World Wide Web Conference (WWW2003). Budapest, Hungary, ACM Press,2003

[50] Song R,Liu H,Wen J-R,et al. Learning Block Importance Models for Web Pages[C]// Proceedings of the 13th World Wide Web Conference. New York,NY USA,2004

[51] Baeza-Yates R,Ribeiro-Neto B. Modern Information Retrieval [M]. 北京:机械工业出版社,2004

[52] Broder A. The Next Generation Web Search and the Demise of the Classic IR model [C]// Rome,Italy:29th European Conference on Information Retrieval 2007

[53] Broder A. A taxonomy of Web search[C]// Proceedings of the ACM SIGIR Forum,2002,36(2)

[54] Stanley L,Leandro Krug W,Jos,et al. Concept-based knowledge discovery in texts extracted from the Web [M]. ACM Press, 2000:29-39

[55] Shuming S,Ji-Rong W,Qing Y,et al. Gravitation-based model for information retrieval[C]// the Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval. Salvador,Brazil,ACM Press, 2005

[56] Bai J,Nie J-Y,Cao G,et al. Using query contexts in information retrieval[C]// the Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval(SIGIR 2007). Amsterdam,The Netherlands, ACM Press,2007

[57] Allan J. HARD Track Overview in TREC 2005 High Accuracy Retrieval from Documents[C]// Proceedings of the The Fourteenth Text REtrieval Conference(TREC 2005). NIST,2005

[58] Allan J. HARD Track Overview in TREC 2004-High Accuracy Retrieval from Documents[C]// Proceedings of the The Thirteenth Text Retrieval Conference(TREC 2004). NIST,2004