探寻式搜索研究述评 *

姜婷婷　高慧琴

摘　要　随着信息搜寻领域的不断细分,探寻式搜索作为一个新兴的研究热点引起了广泛关注。本文重点从探寻式搜索理论和系统两个方面对已有研究成果进行梳理分析。研究发现:探索性既可以表现在搜索的问题背景上,又可以体现在搜索的过程中;探寻式搜索分为学习性搜索和研究性搜索两个层次,要求用户结合启发式和分析式的搜索策略应对不确定性的波动;目前探寻式搜索系统主要是通过搜索结果的分类来支持用户对大规模结果集合的浏览探索,常见的分类方式包括层级分类、多面分类、动态聚类和社会分类,它们在结构、词汇、创建、使用等方面都具有不同特征,适用于不同的搜索系统。然而,过去的探寻式搜索研究普遍忽视了信息搜寻的社会性本质,本文认为:他人的建议或活动能够影响信息决策,探寻式搜索将与直接或间接的社会性导航发生融合,而社会性软件的兴起又为此提供了可行性。图2。表1。参考文献57。

关键词　信息搜寻　探寻式搜索　查询式搜索

分类号　G354

A Review of Research Studies on Exploratory Search

Jiang Tingting & Gao Huiqin

ABSTRACT　As a result of the subdivision of the information seeking domain, exploratory search has become a new research focus which arouses extensive attention. This paper reviews existing related research based on exploratory search theories and systems, and engenders a series of findings: both the search problem context and process can be exploratory; learning and investigation constitute the two levels of exploratory search, and users need to combine the heuristic and analytical strategies to tackle the fluctuation of uncertainty; current exploratory search systems support the exploration of large-scale result collections through classification of search results. The four major approaches to classifying search results are hierarchical classification, faceted classification, dynamic clustering, and social classification, which differ in in terms of structure, vocabulary, creation, and use, thus applicable to different search systems. The social nature of information seeking, however, has been ignored in previous exploratory search studies. As predicted in the paper, exploratory search will be incorporated with direct or indirect social navigation since others' advice or actions may affect people's information decisions, and the rise of social software makes this possible. 2 figs. 1 tab. 57 refs.

KEYWORDS　Information seeking.　Exploratory search.　Query-based search.

1　引言

信息搜寻(Information Seeking)是人类为了改变现有知识状态而搜寻相关信息的活动[1]。在线信息搜寻已经成为我们日常生活和工作中不可或缺的组成部分。明确的信息需求,譬如了解天气、航班、股票等情况,通常可以借助功能强大的网络搜索引

*　本文系国家自然科学基金青年项目“用户探寻式搜索策略分析及系统构建研究”(编号:71203163)及教育部人文社会科学研究青年基金项目“社会性标签系统中用户信息搜寻行为研究”(编号:12YJC870011)的研究成果之一。

通讯作者:姜婷婷, Email: tij@whu.edu.cn

擎(如 Google、Bing、百度等)得到充分满足。这体现了信息检索领域中最常见的查寻模型(Lookup Model):查寻对象一般是"已知条目",查寻机制的核心在于用户提问式与被检索文档集合索引之间的匹配[2]。基于这个模型,具体的提问式能够带来准确的结果,几乎不需要用户对结果进行评价或比较[3]。

遗憾的是,查寻模型对于现实世界中的许多情况都并不适用。比方说,科研人员想深入调查一个新兴的研究热点,自助游客想制定一套合适的旅游方案,年轻人想规划一条通往成功的人生道路等。诸如此类的问题所包含的信息需求很难一次性表述成恰当的提问式,因为用户对搜索目标所涉及的知识领域并不熟悉,或者不清楚如何才能达到目标,又或者目标本身就不太明了[4]。

在以上情况中,用户的首要任务实际上是界定当前的搜索目标,他们需要借助搜索系统逐步向可能与之关联的信息内容靠拢,通过不断吸收新知识来完善自己对问题的认识,从而分辨出已经知道什么以及需要知道什么,两者之间的缺口就是待解决的信息需求。随着信息需求的成形,用户表达提问式和识别相关条目的能力也在增强,此时搜索系统的自动匹配功能才开始发挥最大效用,而最终能否为信息问题找到满意的答案还取决于用户从搜索结果中抽取、理解、整合有价值的信息的技能。这里我们看到的是以用户为导向的非线性搜索,也就是"探寻式搜索"(Exploratory Search)。

探寻式搜索是一种特殊的信息搜寻,它作为一个独立的研究领域开始于 2005 年在美国马里兰大学召开的"探寻式搜索界面"(Exploratory Search Interface)专题研讨会[5]。在此之后的几年时间里,ACM SIGIR、ACM SIGCHI 和 NSF 等学术组织分别举办了"探寻式搜索系统评价"(Evaluating Exploratory Search Systems)、"探寻式搜索与人机交互"(Exploratory Search and CHI)和"信息搜寻支持系统"(Information Seeking Support Systems)等专题研讨会;另外,*Communications of the ACM*、*Information Processing and Management* 和 *Computer* 等多个有影响力的期刊也相继发表了以探寻式搜索为主题的论文特辑。

2　相关理论

近几年来,越来越多来自信息检索、人机交互、信息组织、信息行为等领域的研究人员进入了探寻式搜索领域。当然这并不是偶然的,探寻式搜索研究的确可以在这些领域中寻求理论根源。下面从用户的内在认知和外在行为两个方面对相关理论进行简要回顾。

2.1　交互式信息检索与认知信息检索理论

交互式信息检索(Interactive Information Retrieval)打破了以系统为中心的传统研究视角,更加注重用户对搜索过程的投入与控制,并且在很多时候与认知信息检索(Cognitive Information Retrieval)融合在一起,因为交互的主要目的就是影响用户的认知状态从而使其有效地获取信息[6]。

Ingwersen 认为信息检索中的所有交互活动都可以引起认知过程,他对检索系统的信息空间和用户的认知空间都创建了多元表示(Polyrepresentation):前者由系统环境和信息对象组成;后者包含了用户的任务或兴趣领域、当前认知状态、问题空间和信息需求,这四个要素自下而上存在着因果关系[7]。该多元表示综合模型是最全面的交互认知模型之一,与之类似的有 Saracevic 的分层模型(Stratified Model)[8]。

在"不规则知识状态"(Anomalous States of Knowledge, ASK)假设的指引下,Belkin 为检索交互建立了片段模型(Episode Model),将用户与信息之间的一系列交互视为一个信息搜寻片段,用户的目标、问题、意图、处境等因素决定了特定时间点上交互的类型,而交互本身又依赖于信息的表达、比较、呈现、导航、可视化等支持[9]。

Spink 围绕交互的重复性开展实证研究并生成了交互反馈模型,将搜索过程划分为多个周期,每个周期内又可能出现多次交互反馈循环(Interactive Feedback Loops),描述了用户对内容相关性、用语相关性、信息集合规模的判断以及对策略、用语的考察[10]。

2.2 演进式搜索与信息觅寻理论

在演进式搜索(Evolving Search)理论中,Bates[2]提出了两个重要观点。一方面,在大多数真实的搜索中,用户的提问式是在不断变化的,而且可能不止停留在用语的修改上,因为搜索时遇到的新信息会给他们带来新的想法,信息需求也会发生动态的演进。另一方面,用户的信息需求并不是由一组最优的结果满足的,他们在搜索的每个阶段采集到一些有用的信息,将这些碎片串联起来才能实现总体目标。形象地说,演进式搜索遵循了"一次一点点"的采莓(Berrypicking)模式。

信息觅寻(Information Foraging)理论则更关注搜索活动本身的演进。人类觅寻信息类似于自然界中的动物觅食,最好的觅寻者能够在每单位成本上得到最多的有用信息。Pirolli & Card[11]发现,信息环境通常具有"分块"结构,即信息包含在信息块内,觅寻者凭借信息嗅觉判断信息块的价值,这是他们根据一些邻近的中间信息所形成的不完整认识。为了提高信息觅寻的效率,觅寻者可以尝试降低在信息块间移动的平均成本或增大在当前信息块中获取信息的收益。

3 探寻式搜索理论阐释

以上相关理论对于探寻式搜索都具有一定的解释作用:演进式搜索与信息觅寻强调的是外界信息环境对用户搜索方向的影响,交互式信息检索与认知信息检索则强调了搜索目标确定后用户的主观心理与信息客体之间的相互影响。然而它们都未涉及探寻式搜索中的"探索性"这个根本特点。根据White & Roth[12]对探寻式搜索内涵的概括,这个概念既可以表示"无唯一答案、持续发生、包含多个层面的信息搜寻问题背景",也可以表示"方向不确定、多次循环、依赖多种方法的信息搜寻过程"。这两个方面其实是紧密关联的,因为解决复杂或模糊的信息问题必然有赖于非线性的探索过程。

3.1 探索性的问题背景

人类之所以会搜索是因为意识到了信息问题的存在。为了保证生活和工作的正常运转,我们每天都不得不执行各种各样的任务(Work Tasks),这为搜索活动提供了必要的问题背景[13]。Byström & Hansen[14]、Kim & Soergel[15]、Li[16]等基于不同的维度框架对任务类型进行了详尽的分类。在众多的维度中,有三个是最根本也是最通用的任务属性,即预期答案的确切性(Specificity)、容量(Volume)和时间性(Timeliness)。确切性高的答案更倾向于呈现单个事实,用户有足够的把握确定其是否有效;确切性低的答案则更着重作出全面的诠释或评论,但用户对于目标是否达成缺乏清晰的判断。与确切性对应的是容量,事实性答案所容纳的信息不多,也许只有一个名称、一串数字或一张图表等;而诠释性或评论性的答案可能贯穿一个或多个文档,从不同角度反映相关现象的本质。时间性指的是得到答案所需的时间,这可以是一瞬间、几分钟,也可以以小时、天、月来计算,甚至更长[1]。

在图1中,我们利用连续变化区间来表示以上三个属性。毫无疑问,触发查寻式搜索的信息问题在这三个连续区中都向左端趋近。也就是说,用户对明确的、有限的搜索目标有着即时的预期。而探寻式搜索并不单纯是为了获取信息,它更像是与之交织在一起的学习(Learning)与研究(Investigation),通常都会牵涉到结构不完整的信息问题[3]。结构越不完整的信息问题对人类认知资源的消耗越多,在确切性、容量、时间性三个连续区中就越趋近右端,同时"无唯一答案"、"持续发生"、"包含多个层面"这些特征也表现得越显著。

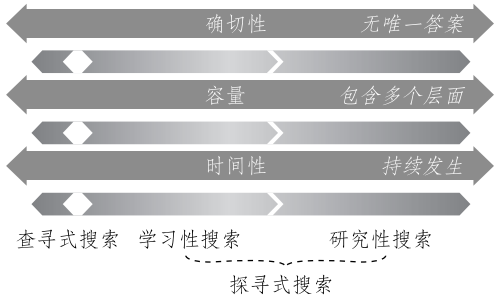

图1 信息问题的三个连续变化区间

为了表现探索性强度的不同,我们对学习和研究这两个探寻层次作了进一步的区分(见图1)。学

习性搜索往往针对某一主题或领域积累相关知识，用户期待能够消除未知的诠释性答案。他们寻获的大量信息对象，包括文本、图片、音频、视频等，有的相互印证、补充，有的又相互对立、反驳，因而需要花费额外的时间对其进行查看、比较、判别，正是这样的内部认知处理活动加强了人类的知识根基。研究性搜索则更进一步创造新的知识。基于对从信息对象中提炼的内容进行深入分析、综合、评价，用户得以实现智慧的决策、规划以及预测。这是一种更高级的认知处理活动，可能持续更长的时间，并且在很大程度上依赖于他们已有的知识水平，最终将得到融入其自身观点的评论性答案。

3.2　探索性的搜索过程

搜索过程是在一定的问题背景下发生的，Wilson[17]将其划分为四个阶段：问题的识别、问题的定义、问题的解决以及答案的陈述，其中两个相邻阶段的过渡总是伴随着不确定性（Uncertainty）的大幅下降。不确定性是造成焦虑和不自信等情感表现的认知状态，是信息搜寻中普遍存在的负面因素[18]。根据 Shannon 的交流理论，人们接收到的信息越多，其不确定性就越低；信息科学的观点认为，新信息的出现有时会导致不确定性的反弹，尤其是在搜索初期[19]。

由探索性问题背景引起的不确定性可能呈现出更明显的波动，这种波动一般会随着时间的推移而趋于缓和，不确定性也不断下降。但有一些特殊情况，譬如信息范围的扩大和（或）复杂程度的增加，会使得不确定性继续波动甚至上升[12]。后一种现象可能发生在以上任一阶段，用户将不得不回到上一阶段重新降低不确定性[17]。因此，探寻式搜索过程是由这四个阶段的逐步推进和反馈循环共同组成的（见图 2 上半部分）。

与不确定性不同，用户行为是搜索过程中可触知、可观测的变量。Wilson[20]、Choo[21] 和 Bates[22] 等都曾全面探讨过各种信息搜寻行为模式，大家普遍认同提问（Querying）与浏览（Browsing）是两种基本的主动模式，即用户是有意识地投入时间和精力去获取信息的。而它们的区别又在于，提问者需要从记忆中唤起适合的用语来表达信息需求，浏览者则需要从周围环境中辨认出有用的信息[1]。探索性的搜索过程是以提问与浏览的相互融合、交替为特点的[3]，图 2 的下半部分展现了这一特点。

从信息问题的初步识别到充分定义，用户在很大程度上依赖以浏览为主导的启发式策略：向具有潜在价值的信息集合导航，通过快速扫视在文档中定位与问题有关的概念，利用横向思考建立起概念之间的联系，从而为进一步明确核心的信息需求提供了前提。在为信息问题寻求答案时，用户则更多地采取以提问为主导的分析式策略：将信息需求分解为操作性更强的子需求并转化为提问式，串行或并行地从搜索系统获得结果，凭借纵向思考深化对有关概念的理解，逐渐提高提问式的准确度直至最后获得满意的答案。

浏览前的尝试性提问是启发式策略的一部分，提问后的针对性浏览是分析式策略的一部分。由于浏览是由外界信息驱动的，因而用户有机会对偶遇的概念产生兴趣，这为新需求的产生提供了可能，使得搜索朝新的方向发展。虽然内驱的提问一般不会带来搜索方向的明显改变，但如果一个提问式对应的结果为空，用户很可能由此发现一个新的问题。在整个探索性的搜索过程中，用户会沿着不可预知的路线曲折前进（见图 2）。

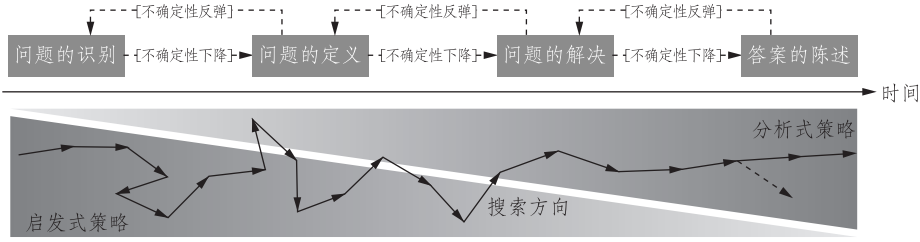

图 2　探寻式搜索过程模型

4 探寻式搜索系统调查

随着探寻式搜索的理论基础日益巩固,各种探寻式搜索系统(Exploratory Search Systems)不断出现,试图通过新的功能和界面增强用户与系统的交互水平以提高提问与浏览的效率[23]。基于探寻式搜索系统对用户构造与改进提问式、浏览搜索结果分别给予相应的支持,已经成为探寻式搜索领域的普遍研究思路。其中,针对提问的支持在传统搜索引擎中也比较常见,例如提问式建议(Suggestion)和扩展(Expansion)工具[24]。而专属于探寻式搜索系统的浏览支持则引起了更为广泛的关注,主要是通过信息分类来加强用户个体对大规模信息集合的理解和操控能力[12]。

众所周知,主流的网络搜索引擎十分注重查准率,尤其是第一页搜索结果的高度相关性。不同的是,探寻式搜索系统却更加偏重于查全率,即使是排序靠后的结果页面上也可能出现有用的内容[3]。这无疑会使用户在浏览结果时遭遇导航困难。为了减轻用户的浏览负担,探寻式搜索系统引入了多种分类思想,将大量的搜索结果划分到少数类别中去,从而降低信息空间的密度,让用户更容易识别关键的信息[25]。而分类本身是基于类别特征分析对其进行系统排列的过程;对一个类别起到定义作用的那一组特征,每个都是必不可少的,它们联合起来又足以将该类别与其他类别区分开来[26]。

我们从信息分类的角度对已有的探寻式搜索系统展开了调查。为了保证调查的全面性,我们不仅考察了目前用户可以正常使用的实际系统,而且还从相关文献中选取了人们经常讨论的一些原型系统。调查结果显示,探寻式搜索系统对搜索结果进行分类的方式主要有层级分类(Hierarchical Classification)、多面分类(Faceted Classification)、动态聚类(Dynamic Clustering)和社会分类(Social Classification)。

4.1 基于层级分类的探寻式搜索系统

层级分类按照逻辑顺序对信息空间进行层层划分,建立上位类和下位类之间的父子关系,形成一套列举式细分体系,一般以树形结构表示[27]。以往已有研究对利用层级分类组织搜索结果进行了探讨。Chen & Dumais 采取机器学习的方法将搜索引擎返回的网页实时划分到网络目录 LookSmart 的既有分类中[28],他们通过比较研究表明这样能使用户花费的搜索时间缩短一半。在 CITIDEL 数字图书馆的 CitiViz 搜索界面上,结果文档是以计算机领域的主题层级体系 ACM Computing Classification System 为基础分类的,其有效性在各种探索性任务(即除给定标题搜索以外的任务)中都大大超越了列表形式[29]。此外,层级分类也可以改善网站内部搜索,曾应用于加州大学伯克利分校网站的 Cha-Cha 系统将站内搜索结果呈现于站点本身的层级结构之中[30],类似的还有马里兰大学人机交互实验室的 WebTOC 系统[31]。

以上研究是构建探寻式搜索系统的有益尝试,为用户提供已知的、稳定的层级分类有助于他们迅速建立关于整个搜索结果空间的心理模型,并了解自己在这个空间中所处的位置。虽然这些研究仅仅揭示了用户搜索效率的上升,但我们可以进一步推断,层级分类的已知性还能够降低分类本身的认知难度,其稳定性又能够减少用户搜索时的焦虑情绪。然而,基于层级分类的搜索系统一般很难跳出原型阶段,这可能是因为根据主题对结果进行自动层级分类实现过程比较复杂,不仅要考虑层级宽度和深度的平衡,还要应对多元层级(Polyhierarchy)[32]的问题,即一个结果条目同时适于两个不同的类别。目前人们常用的网络资源层级分类目录 Yahoo! Directory① 和 Open Directory Project② 就分别是由专家和用户编辑而成的。

4.2 基于多面分类的探寻式搜索系统

多面分类区别于层级分类的首要特征在于它

① http://dir.yahoo.com/

② http://www.dmoz.org/

包含了多个分面,分别代表信息集合的多个属性,然后每个分面又包含了各种属性值的若干个分类[33]。Flamenco[34]、mSpace[35]和 Relation Browser[36]等都是将多面分类应用于搜索的开拓性研究。虽然它们针对的是不同的信息集合,在界面设计细节上也存在区别,但都以一组小型的分面层级取代单个大型的主题层级,允许用户通过依次浏览各层级选择相关类别来逐步确定搜索范围。可用性研究表明,多面分类易于理解,能够帮助搜索系统消除结果为空的情况,支持探索与发现,受到用户的青睐[37]。

近几年来,多面分类在各种搜索环境中得到越来越广泛的应用。尤其是在电子商务平台上,如 C2C 的 eBay①、淘宝②和 B2C 的 Overstock③、Bestbuy④,多面搜索能够充分利用商品已有的结构化元数据提高其可查找性,带来巨大的商业价值[38]。除此之外,多面搜索也逐渐成为下一代图书馆目录的基本特征之一,许多大学图书馆,例如杜克大学⑤、哈佛大学⑥和匹兹堡大学⑦的图书馆,都在依赖 Endeca、AquaBrowser 和 Summon 等发现服务商的技术支持为其读者探索馆藏资源提供多面浏览体验[39]。

不难看出,多面分类通过使用多个概念维度满足了不同用户从不同的角度理解信息的需求,可以有效应对复合性概念给信息组织带来的挑战。而分面搜索本身就是一种探寻式搜索,它建立了搜索结果到多面分类体系的映射,用户可以灵活地按照任意顺序选取自己关心的任意分面来一步步探查结果,而且被他们识别并选中的所有类名连起来也相当于一条复杂的布尔提问式。我们可以预测,多面分类结构的逻辑性将使得多面搜索系统成为数字环境中信息搜索的主流工具。

4.3　基于动态聚类的探寻式搜索系统

聚类的基本思想是利用特定的算法对信息条目进行分组,同一组内的条目应该表现出相似性或关联性,而不同的组又应该具有明显的区别[40]。聚类搜索的兴起始于 Vivísimo 企业搜索,其特点在于检索后聚类(Post-Retrieval Clustering),包括根据结果内容形成分类结构,将各结果条目分插入合适的分类中,选择准备向用户展示的主要类别[41]。目前影响力最大的网络聚类搜索引擎 Yippy⑧(原为 Clusty)就是源于 Vivísimo 的技术支撑,其竞争对手主要有 iBoogie⑨、PolyMeta⑩、Carrot2⑪ 等,而早期著名的 Grokker、KartOO、WebClust、Mooter 却由于各种原因已经无法正常使用。这些系统一般只对最优的几十到几百个结果进行聚类,产生的分类结构可能是单层或多层的[25]。

由于类名的质量决定了分类结构的可用性,聚类搜索系统通常采用以描述为中心的聚类算法,主张类别的描述必须简洁明了、易于理解并且清楚地表现类别中的结果,而无法描述的类别对于用户来说没有价值,应该去掉[35]。聚类搜索系统通常具有元搜索(Metasearch)特征,例如通过 API 方式从 Google、Bing 等处获取并聚合搜索结果,本身则专注于聚类过程。元搜索解决了单个搜索引擎在索引范围上的局限性,使得用户能够在统一界面上更

① http://www.ebay.com/
② http://www.taobao.com/
③ http://www.overstock.com/
④ http://www.bestbuy.com/
⑤ http://library.duke.edu/
⑥ http://lib.harvard.edu/
⑦ http://www.library.pitt.edu/
⑧ http://www.yippy.com/
⑨ http://iboogie.com/
⑩ http://polymeta.com/
⑪ http://search.carrot2.org/stable/search

全面地查看搜索结果[42]。

聚类技术起源于信息检索领域[43],对于探寻式搜索具有十分重要的意义。聚类搜索为每个提问式的结果集合动态创建一个分类结构,这首先为用户查看特定主题下的结果条目提供了快捷方式,并且在很大程度上解决了搜索中一词多义的问题,将表现不同词义的结果区分开来,以便用户有选择地浏览;同时还将原本散落在不同页面上而又存在关联的结果聚集到一起,让所有结果中的重要主题全面呈现出来,有利于用户更为系统地考察整个结果空间。

4.4 基于社会分类的探寻式搜索系统

社会分类又名大众分类(Folksonomy),由大众贡献的自由标签组成,呈现出扁平而松散的形态[44-45]。社会分类最早出现在社会性标签系统(Social Tagging Systems)中,是系统依赖用户通过加标签对资源进行自组织的产物[46]。但由于加标签权限的不同,社会分类又存在狭义与广义之分[47]。前者的代表性系统有 Flickr①、Vimeo②、Reddit③、LiveJournal④ 等,而 BibSonomy⑤、Folkd⑥、LibraryThing⑦、豆瓣⑧等都属于后者。当社会性标签系统的用户作为信息搜寻者时,他们倾向于浏览发现已经被其他用户加过标签的那些资源[48]。习惯基于标签发现资源的用户往往是标签的积极贡献者;标签代表着明确的主题,以此为中间媒介有助于增强浏览过程的方向性[49]。

目前,Amazon⑨ 这个多元化的电子商务平台也允许顾客对商品添加标签,当他们进行标签搜索时,即提问式为标签,搜索结果为所有加了该标签的商品,而对相关标签的选择可以帮助他们进一步精炼结果。在 Amazon 中,社会分类实际上是独立于已有的层级式商品门类而存在的。与此类似,宾夕法尼亚大学和密歇根大学的图书馆也独立于原有图书层级分类引入了社会分类,分别开发了 PennTags⑩ 系统和 Mtagger⑪ 系统[50]。

作为 Web 2.0 平台上的基本分类方法以及 Web 1.0 平台上的辅助方法,社会分类凭借其低廉的创建成本和对外界变化的快速响应在探寻式搜索中表现出应用前景。人们根据自己的理解独立为资源条目添加标签以方便日后检索,这原本是一种个人行为。然而标签聚合技术却赋予其社会性:在微观层面上,添加到一个条目上的所有标签聚合成了它的标引记录;在宏观层面上,来自所有用户的所有标签聚合成了一套分类体系。因此,人们在贡献标签的时候也为别人提供了一条探索资源的途径。

4.5 探寻式搜索系统结果分类方式比较

以上展示了一系列具有代表性的探寻式搜索系统,它们采用不同的分类方法来组织搜索结果。为了给以后的探寻式搜索研究及系统实践提供一个全面综览,我们从结构特征、词汇特征、创建者、创建过程、使用者等方面对各分类方法在探寻式搜索中的应用进行比较,并总结出其所适用的搜索系统(见表 1)。

① http://www.flickr.com/
② http://vimeo.com/
③ http://www.reddit.com/
④ http://www.livejournal.com/
⑤ http://www.bibsonomy.org/
⑥ http://www.folkd.com/
⑦ http://www.librarything.com/
⑧ http://www.douban.com/
⑨ http://www.amazon.com/
⑩ http://tags.library.upenn.edu/
⑪ http://www.lib.umich.edu/mtagger

表1　各分类方法在探寻式搜索中的应用

	层级分类	多面分类	动态聚类	社会分类
结构特征	单个大型的主题层级	一组小型的分面层级	主题聚类	扁平松散的命名空间
词汇特征	受控词表	受控词表	基于内容	任何语言
创建者	专业人士	专业人士	算法自动	普通用户
创建过程	自上而下预先创建	自上而下预先创建	自上而下实时创建	自下而上分散创建
使用者	有足够经验	有足够经验	无需经验	无需经验
适用搜索系统	正式领域专业、有序、集中的信息	特定领域结构化、同质的信息	任何领域大规模、非结构化、异质的信息	非正式领域大规模、混杂、分散的信息

从分类结构的完整性和严谨程度来看，层级分类最高，而大众分类最低，多面分类和动态聚类居于两者之间。层级分类必须准确地反映搜索结果空间中的既有顺序关系，由总到分形成多个层级，各类别之间互无交叠。大众分类却恰恰相反，它不仅缺乏层级，表现出扁平的形态，而且其类别也缺乏排他性。如果说层级分类关心的是如何将结果条目放置到合适的位置，那么多面分类和动态聚类则分别更关心如何去描述和区分这些条目，因而对分类结构本身的要求依次降低。

从分类类名的词汇来源来看，层级分类和多面分类均采用受控词表，即自然语言中经过精确定义的那部分子集，为一般人所理解与熟悉。动态聚类基于搜索结果内容生成类名，这有可能因为算法上的缺陷造成一些无法预料的情况，比如有的类名同时出现在不同的层次上，或者有的类名令人无法理解。社会分类中的标签完全来自用户，他们在不受控制的前提下可能使用任何语言，这很容易引起词汇问题（Vocabulary Problem）和基本层次问题（Basic Level Problem），甚至是垃圾标签的问题。

从分类创建的人力成本来看，这四种方法也表现出递减趋势。层级分类和多面分类都需要既具备信息组织技能又拥有领域知识的专业人士事先根据搜索系统的需要来手工创建，而且他们还得负责创建后可能发生的类别增删与调整工作。由于层级分类本身的复杂特性，其创建成本会更高些。虽然动态聚类克服了这点不足，但是我们也不能忽视前期聚类算法设计与分析的人力投入。社会分类则是在大量用户的共同努力下不停发展的，他们分摊着创建和维护分类的成本。

从分类方法对使用者的要求来看，如果没有足够的经验，用户在层级分类或多面分类中浏览搜索结果可能无法达到预期效果，他们最好对信息分类规律和搜索所涉及的背景知识有一定了解。相比较而言，动态聚类和社会分类对用户的使用经验没有特别要求，任何人都可以在很短的时间内上手。

最后，我们对四种分类方法所适用的探寻式搜索系统进行了归纳。层级分类最适合于正式领域，如科技文献、百科知识的搜索，用于组织专业、有序、集中的信息。多面分类则对搜索领域的内部特征提出了具体要求，即领域中包含结构化、同质的信息，这是构建分面的基础，如图书的作者、语言和年代，或商品的价格、品牌和产地等。动态聚类突破了领域的限制，能够应对大规模、非结构化、异质的信息，是一般网络信息搜索的最佳选择。社会分类因其结构和词汇上的劣势一直以来只在非正式领域中得到推广，例如以用户参与和合作为特点的社会性软件中的信息搜索，或作为其他分类方法的补充手段。

5　探寻式搜索领域展望

信息分类技术的进步固然可以推动探寻式搜索系统的发展，不过已经有少数研究人员将目光投向更为广阔的社会空间。Evans & Chi[51] 通过对150位参与者的调查发现，人际信息沟通贯穿着整

个搜索过程,可以作用于人们搜索前的问题陈述、搜索中的信息搜集与选择以及搜索后的结果分享。Kammerer[52]利用来自社会性书签(Social Bookmarking)站点的大众标签数据对搜索结果进行标引,并根据用户反馈进一步提高结果列表的相关性,实验表明这样能够有效地支持用户在结构不完善的领域内探索新知识。

的确,无论是显性的还是隐性的社会交互都是探寻式搜索领域未来的发展方向。Chi[53]曾指出,用户在信息搜寻的过程中并不是互相隔离的,他们会出于各种各样的理由从其他人那里获取信息,而且这种倾向是非常强烈的。Morville & Rosenfeld[32]也认为,向别人寻求帮助是与提问、浏览同样重要的基本搜寻模式。然而从以上对探寻式搜索系统的调查可以发现,尽管用户的探索活动在系统提供的信息线索(Informational Clues)的帮助下确实可以变得更加高效,但他们仍然只是信息检索传统观点中独立活动的个体。

在社会性软件蓬勃发展的 Web 2.0 时代,信息的交流与共享更加广泛、频繁,人类的信息搜寻已经不可避免地与其社会交往融合起来,这为探寻式搜索带来新的可能性。一方面,人与人的交谈对话可以跨越词汇的障碍,让用户以更自然的方式提问,有助于降低他们的认知负荷;另一方面,群体的"聚合智慧"(Collective Intelligence)可以创造出社会线索(Social Clues),即后来的用户利用以往的用户留下来的动作痕迹,选择大多数人都认为合理的浏览路径。Svensson[54]将这两种社会交互形式分别称为直接和间接的社会性导航。

导航是没有明确目标的搜索,而社会性导航是以人为指引的导航[55]。直接社会性导航是指依靠双向交流产生个性化的导航建议,协助主体不仅能够根据具体情况回答"在哪里"之类的基本导航问题,更重要的是有机会帮助导航主体明确其目的地并选择通往那里的正确路线。而间接社会性导航则表现为协助主体对导航主体的单向交流,但他们通常又没有意识到自己为其他人提供了导引,交流是以"累积信息"(Cumulative Information)的形式发生的,具有突出的动态特征,人的进入和占据打破了信息空间的预先规划并促使其不断发展,就像森林里的路是人们走出来的一样[54]。

早期的社会性导航支持系统都是以间接导航为中心的历史富集型信息环境(History-Enriched Environments)。自 2005 年左右社会性软件兴起以来,以此为基础开展的社会性导航研究也随之增多。Millen & Feinberg[48]在研究社会性书签服务 dogear 时发现查看别人的书签收藏和点击标签查看相关的书签是最常见的社会性导航形式。Vosinakis & Papadakis[56]在虚拟世界(Virtual Worlds)的 3D 环境中整合了空间、语义、社会这三种导航形式,他们提出来的原型框架包含了主题讨论、用户足迹与标签、语义过滤、内外数据互用等几个主要特征。Shami[57]设计的社会性文件分享(Social File Sharing)系统 Cattail,通过近期活动流和共享下载历史支持社会性导航,系统评价结果表明 Cattail 能够帮助用户发现更多相关的人和内容。

探寻式搜索在国际上仍是处于上升阶段的新兴领域,而在国内该领域尚未起步。总的来说,探寻式搜索研究目前还停留在对用户独立搜索活动的关注,忽视了社会化支持对于信息探寻的必要性和必然性。为了不让搜索社会化研究的匮乏成为探寻式搜索领域发展的一块短板,社会性导航领域的融入是一个自然而然的趋势。我们可以从直接或间接社会性导航的研究中获得启发,同时社会性软件的兴起又进一步增加了在探寻式搜索过程中实现社会交互的可行性。社会交互以他人的建议或活动影响着人们的信息决策,为探寻式搜索研究朝更多元的方向推进提供了指引,对于信息搜寻领域乃至整个图书馆学、情报学的发展都将产生重要的推动作用。

参考文献

[1] Marchionini G. Information seeking in electronic environments[M]. Cambridge, UK: Cambridge University Press,

1995.

[2] Bates M J. The design of browsing and berry picking techniques for the online search interface[J]. Online Review, 1989, 13(5): 407 -427.

[3] Marchionini G. Exploratory search: From finding to understanding[J]. Communications of the ACM, 2006, 49(4): 41 -46.

[4] Nolan M. Exploring exploratory search[J]. Bulletin of the American Society for Information Science and Technology, 2008, 34(4): 38 -41.

[5] White R W, Kules B, Bederson B. Exploratory search interfaces: Categorization, clustering, and beyond[EB/OL]. [2012 -12 -26]. http://research. microsoft. com/en-us/um/people/ryenw/papers/WhiteSIGIRForum2005b. pdf.

[6] Saracevic T. Modeling interaction in information retrieval (IR): A review and proposal[C]// Proceedings of the 59th Annual Meeting of the American Society for Information Science, 1996: 3 -9.

[7] Ingwersen P. Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory[J]. Journal of Documentation, 1996, 52(1): 3 -50.

[8] Saracevic T. The stratified model of information retrieval interaction: Extension and application[C]// Proceedings of the 60th Annual Meeting of the American Society for Information Science, 1997: 313 -327.

[9] Belkin N J. Intelligent information retrieval: Whose intelligence?[C]// Proceedings of the 5th International Symposium for Information Science, 1996: 25 -31.

[10] Spink A. Study of interactive feedback during mediated information retrieval[J]. Journal of the American Society for Information Science, 1997, 48(5): 382 -394.

[11] Pirolli P, Card S K. Information foraging[J]. Psychological Review, 1999, 106: 643 -675.

[12] White R W, Roth R A. Exploratory search: Beyond the query-response paradigm[J]. Sythesis Lectures on Information Concepts, Retrieval and Services, 2009, 1(1): 1 -98.

[13] Ingwersen P, Järvelin K. The turn: Integration of information seeking and retrieval in context[M]. Dortrecht, NL: Springer, 2005.

[14] Byström K, Hansen P. Conceptual framework for tasks in information studies[J]. Journal of American Society for Information Science and Technology, 2005, 56(10): 1050 -1061.

[15] Kim S, Soergel D. Selecting and measuring task characteristics as independent variables[C]// Proceedings of the 68th Annual Meeting of the American Society for Information Science, 2005.

[16] Li Y. Exploring the relationships between work task and search task in information search[J]. Journal of the American Society for Information Science and Technology, 2009, 60(2): 275 -291.

[17] Wilson T D. Models in information behavior research[J]. Journal of Documentation, 1999, 55(3): 249 -270.

[18] Kuhlthau C C. The role of experience in the information search process of an early career information worker: Perceptions of uncertainty, complexity, construction, and sources[J]. Journal of the American Society for Information Science, 1999, 50(5): 399 -412.

[19] Kalbach J. On uncertainty in information architecture[J]. Journal of Information Architecture, 2009, 1(1): 48 -56.

[20] Wilson T D. Information behavior: An interdisciplinary perspective[J]. Information Processing and Management, 1997, 33(4): 551 -572.

[21] Choo C W, Detlor B, Turnbull, D. Information seeking on the Web: An integrated model of browsing and searching [C]// Proceedings of the 62nd Annual Meeting of the American Society for Information Science, 1999.

[22] Bates M J. Toward an integrated model of information seeking and searching[J]. The New Review of Information Behav-

ior Research, 2002, 3: 1 - 15.

[23] White R W, Marchionini G, Muresan G. Evaluating exploratory search systems: Introduction to special topic issue of information processing and management[J]. Information Processing & Management, 2008, 44(2): 433 - 436.

[24] Croft W B, Metzler D, Strohman T. Search engines: Information retrieval in practice[M]. Addison Wesley, 2009.

[25] Jiang T, Koshman S. Exploratory search in different information architectures[J]. Bulletin of the American Society for Information Science and Technology, 2008, 34(6): 11 - 13.

[26] Jacob E K. Classification and categorization: A difference that makes a difference[J]. Library Trends, 2004, 52(3): 515 - 540.

[27] Taylor A G. Wynar's introduction to cataloging and classification[M]. Libraries Unlimited Inc., 2004.

[28] Chen H, Dumais S. Bringing order to the Web: Automatically categorizing search results[C]// Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2000: 145 - 152.

[29] Kampanya N, Shen R, Kim S, et al. Citiviz: A visual user interface to the CITIDEL system[C]// Proceedings of the 8th European Conference on Digital Libraries, 2004: 122 - 133.

[30] Chen M, Hearst M, Hong J, et al. Cha-Cha: A system for organizing Intranet search results[C]// Proceedings of the 2nd USENIX Symposium on Internet Technologies and Systems, 1999: 11 - 14.

[31] Nation D A, Plaisant C, Marchionini G, et al. Visualizing websites using a hierarchical table of contents browser: WebTOC[C] // Proceedings of the 3rd Conference on Human Factors and the Web, 1997.

[32] Morville P, Rosenfeld L. Information architecture for the World Wide Web: Designing large-scale web sites[M]. O'Reilly Media, Incorporated, 2006.

[33] Tunkelang D. Faceted search[J]. Synthesis Lecture on Information Concepts, Retreival, and Services, 2009, 1(1): 1 - 80.

[34] Hearst M A. Clustering versus faceted categories for information exploration[J]. Communications of the ACM, 2006, 49(4): 59 - 61.

[35] Schraefel M C, Smith D A, Russel A, et al. The mSpace classical music explorer: Improving access to classical music for real people[C] // MusicNetwork Open Workshop, Integration of Music in Multimedia Applications, 2005.

[36] Capra R G, Marchionini G. The relation browser tool for faceted exploratory search[C] // Proceedings of the 8th ACM/IEEE-CSJoin Conference on Digital Libraries, 2008: 420 - 420.

[37] Yee K P, Swearingen K, Li K, et al. Faceted metadata for image search and browsing[C] // Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2003: 401 - 408.

[38] Dash D, Rao J, Megiddo N, et al. Dynamic faceted search for discovery-driven analysis[C] // Proceedings of the 17th ACM Conference on Information and Knowledge Management, 2008: 3 - 12.

[39] Yang S Q, Wagner K. Evaluating and comparing discovery tools: How close are we towards next generation catalog?[J]. Library Hi Tech, 2010, 28(4): 690 - 709.

[40] Manning C D, Raghavan P, Schütze H. Introduction to information retrieval[M]. Cambridge: Cambridge University Press, 2008.

[41] Koshman S, Spink A, Jansen B J. Web searching on the vivisimo search engine[J]. Journal of the American Society for Information Science and Technology, 2006, 57(14): 1875 - 1887.

[42] Morville P, Callender J. Search patterns[M]. O'Reilly Media, Inc., 2010.

[43] Van Rijsbergen K. The geometry of information retrieval[M]. Cambridge: Cambridge University Press, 2004.

[44] Shirky C. Ontology is overrated: Categories, links, and tags[EB/OL]. [2012 - 12 - 26]. http://www.shirky.com/

writings/ontology_overrated.html.

[45] Kroski E. The hive mind:Folksonomies and user-based tagging[J]. Library, 2007, 2: 91 - 103.

[46] Smith G. Tagging:People-powered metadata for the social web[M]. New Riders, 2008.

[47] Golder S A, Huberman B A. Usage patterns of collaborative tagging systems[J]. Journal of Information Science, 2006, 32(2): 198 - 208.

[48] Millen D R, Feinberg J. Using social tagging to improve social navigation[C] // Workshop on the Social Navigation and Community based Adaptation Technologies, 2006.

[49] Jiang T. An exploratory study on social library system users' information seeking modes[J]. Journal of Documentation, 2013, 69(1): 6 - 26.

[50] Pirmann C. Tags in the catalogue: Insights from a usability study of library thing for libraries[J]. Library Trends, 2012, 61(1): 234 - 247.

[51] Evans B, Chi E H. Towards a model of understanding social search[C] // Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work, 2008: 485 - 494.

[52] Kammerer Y, Nairn R, Pirolli P,et al. Signpost from the masses: Learning effects in an exploratory social tag search browser[C] // Proceedings of the 27th International Conference on Human Factors in Computing Systems, 2009: 625 - 634.

[53] Chi E H. Information seeking can be social[J]. Computer, 2009, 42(3): 42 - 46.

[54] Svensson M. Social navigation[M] // Dahlback N. Exploring Navigation:Towards a Framework for Design and Evaluation of Navigation in Electronic Spaces. Swedish Institute of Computer Science, 1998: 73 - 88.

[55] Svensson M. Defining, designing and evaluating social navigation[D]. Stockholm, Sweden: Department of Computer and Systems Sciences, Stockholm University, 2002.

[56] Vosinakis S, Papadakis I. Virtual worlds as information spaces:Supporting semantic and social navigation in a shared 3D environment[C] // Proceedings of 3rd International Conference on Games and Virtual Worlds for Serious Applicants, 2011: 220 - 227.

[57] Shami N S, Muller M, Millen D. Browse and discover:Social file sharing in the enterprise[C] // Proceeolings of the ACM 2011 Conference on Computer Supported Cooperative Work, 2011: 295 - 304.

姜婷婷　武汉大学信息管理学院副教授,硕士生导师。通讯地址:武汉市武昌区珞珈山。邮编 430072。

高慧琴　武汉大学信息管理学院情报学硕士研究生。通讯地址同上。

(收稿日期:2012 - 11 - 05;修回日期:2013 - 01 - 24)