信息检索系统的关联关键词推荐研究

温有奎[1,2]
(1. 中国科学技术信息研究所，北京 100038；2. 北京万方数据股份有限公司，北京 100038)

摘要：目前的信息检索系统对用户是不透明的，用户需要以猜想方式向系统提问并反复浏览检出结果来判断信息价值。大数据加剧了用户因筛选惊人文献量而导致的精神上和时间上的压力，且压力随着跨学科、多关联度信息检索需求的上升而越来越明显。本文提出一种关键词关联推荐的方法，解决过去由用户先猜想输入一个覆盖面大的检索词，再通过浏览缩小检索范围的方法，变为由用户选择系统推荐的内部关键词关联组配的方式来提高检索精确度。实验证明，这种关键词关联推荐方法大大提高了信息检索系统的检索精度，同时减轻了用户的信息检索压力。

关键词：关键词组配；关联推荐；信息检索
中图分类号：G202
DOI：10.3772/j.issn.1673-2286.2016.4.002

随着我国科研投入的加大、创新步伐的加快，中国科研论文数量跃居世界第二。信息数量的急速增长和跨学科创新研究的加剧，给科研工作者快速查明科技文献的精准信息带来了精神上和时间上的极大压力。目前的信息检索系统大多基于用户与系统的关键词匹配的检索原理，这种检索方法简单、快速，但检索系统对用户来讲是不透明的，这种检索方式是一种猜想式的检索，难以解决用户与系统的透明、精确检索要求。因此，需要开发一种根据用户搜索词推荐关联关键词组配的透明检索方法，将以往的用户猜想检索方式转变为用户选择关联关系的检索方式，以提高学术信息检索系统的效率。

1 研究背景

信息检索的基本原理和机制是系统对信息集合与需求集合的匹配与选择。经典的信息检索模型使用一组具有代表性的关键词（索引词）来描述数据库中的每一篇文档。关键词由文档中的一些能反映主题的简单单词构成，通过它们可以与数据库中的文档相联系。经典信息检索模型主要包括布尔检索模型、向量检索模型及概率模型[1]。

目前的信息检索系统大多基于经典的信息检索模型，用户向系统输入搜索词，系统根据用户搜索词查找系统内部的关键词索引，如果关键词索引与搜索词有匹配，系统会给出关键词所代表的检索结果。这种检索方法简单，但检索系统对用户来讲是不透明的，这种检索原理是一种猜想式的检索方法。因此，在用户检索时，用户首先给出一个概念很大的搜索词试探检索，这样系统会给出成千上万条检索结果，需要用户反复浏览检索结果来调整检索词以便达到缩小检索范围的目的。为了提高海量数据检索的精度，大多数学术文献检索系统增加了高级检索功能，即设置了多个检索词来实现逻辑条件（与、或、非）的限定检索方法。高级检索方法的增加起到了一定的限定检索范围的作用，使得检索精度有所提高，检索结果的输出有所专指，但这种方法并没有改变用户与检索系统之间不透明的本质。由于用户与系统之间存在文献信息组织的不透明，若用户利用高级检索功能使用多个假设的关键词进行逻辑（与、或、非）限定，用户使用的自由词与系统的标引词不一致，将会导致检索结果为0的悲剧。其实系统里有与用户搜索词关联的关键词组合，只是用户事先难以知道罢了。用户事先难以了解信息检索系统中他所需要的精确关联信息，因此用户也就难以给出一个理想的关键词从信息检索系统中找到满意的文献信息。尤其是，学术性检索系统比社会网络检索系统要求的检索精度要高，跨学科信

息检索越来越普遍，而目前海量数据常常给用户搜索词的选择上带来了时间上和精神上的极大压力。

大数据、云计算和物联网技术的发展使得科技文献的管理和获取途径大大提升，尤其是目前网络搜索引擎推出的一键式访问模式给用户带来了极大便利，用户只要输入一个简单的关键词，系统就会给出相关联的足够多的信息，用户可以通过浏览选取所需信息。网络信息大多是新闻和消息类的信息，用户浏览信息的多少并不影响用户心情。但对于学术性信息检索系统，用户的需求不仅仅是浏览，而是获取精确的信息。因此，提高检索的精准性是学术信息检索系统的重要目标。而目前的学术信息检索系统模仿了时髦的网络检索系统的一键式检索模型，却对精确检索功能和深度挖掘功能没有加以重视，因而在渐渐地失去用户。美国霍普金斯大学张甲博士指出，“目前的发现系统虽然模仿了Google的一个检索框，却没有抓住读者点击进入后的知识过滤行为的特点和共性”[2]。车天文提出一种用户检索词推荐的方法及系统[3]，这种方法有一定的参考价值，但有其局限性，因为搜索引擎大多依据用户日志文件进行检索词推荐，不具有反映信息检索系统信息关联组合的全面性，因而其推荐会失去全面性功能。岑咏华等研究了用户当前检索关键词的关联推荐方法[4]，但该方法只推荐了单个关联关键词的概率，没有给出关联关键词在数据库中的组合信息，仍需要用户使用展示的关键词进行组合以实现精细检索，这样仍然存在组配信息不明的问题。并且，若用户不使用组配检索，只选用单个推荐词，则会进入另一个相关词的领域，这样就会偏离用户最终目标。本文提出信息检索系统的关联关键词推荐思路，寻求解决用户信息检索过程压力的方法，提高信息检索系统的满意度。

2 关联关键词推荐方法

2.1 关联关键词推荐方法流程

关联推荐方法在电子商务、在线音乐、在线新闻、社交网络、个性化搜索等诸多方面表现出其不可替代的作用。由于其能够实现的可定制功能，使得可以针对不同的用户提供个性化服务，更能够让用户从海量数据中更轻易地定位到自己所需的信息，提高使用体验。目前主要使用的推荐策略包括基于用户的推荐、基于物品的推荐、关联规则推荐、协同过滤推荐，等等[5]。

为解决信息检索系统对用户不透明引起的检索压力过大的问题，本文采用基于用户搜索词的关联关键词推荐方法，提高检索系统的透明度，提高信息检索的精确度。本文的关联关键词推荐方法流程如图1所示。

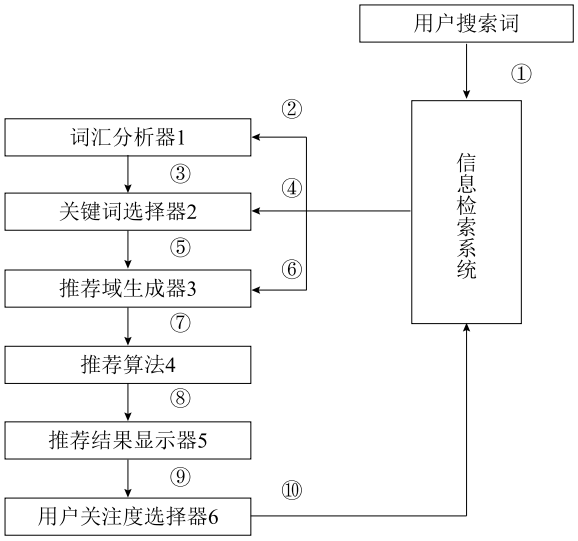

图1 关联关键词推荐算法流程

具体步骤说明：①用户向信息检索系统输入检索词；②从信息检索系统读取用户搜索词；③经词汇分析器分析用户搜索词与信息检索系统关键词匹配情况，若匹配则直接推荐系统关键词，否则，从信息检索系统选择能表达用户搜索词的系统关键词；④从信息检索系统读取与用户匹配的关键词，放入关键词选择器；⑤将关键词推荐给推荐域生成器；⑥从信息检索系统读取推荐域需要的关键词和元数据ID号；⑦将推荐域内的关键词送入关联推荐算法进行运算；⑧将推荐结果送入推荐结果显示器中；⑨用户选择关注度高的推荐结果存放在用户关注度选择器中；⑩将用户选择的推荐结果送入信息检索系统，信息检索系统给出用户满意的检索结果。

2.2 关联关键词推荐算法

①令A为输入搜索词；

②令B为系统匹配关键词；

③若A≠B，则执行用户重新输入搜索词⑩；否则执行④；

④建立以B为推荐域的事务数据库T；

⑤设$I=\{I_1,I_2,I_3,\cdots,I_m\}$是一个有$m$个项的集合，事

务是k个项组成的集合，记为$t\subseteq I$，所有t组成集合事务数据T，其中，

$t=\{(t_1,a_i),(t_2,a_i),(t_3,a_i),\cdots,(t_m,a_i)\}$，$i\subseteq 1,2,3,\cdots,m$；

⑥计算支持度$\text{supp}(X)=|\tau(X)|/|T|$；

⑦计算最小置信度$\text{conf}(X\Rightarrow Y)=|\tau(X\cup Y)|/|\tau(X)|$；

⑧按照最小置信度倒序排列输出推荐结果；用户选择推荐方案，转入系统检索，输出推荐结果；

⑨推荐系统结束；

⑩请用户更换搜索词。

2.3 频繁项集挖掘算法

关联规则推荐最重要的是寻找置信度大于置信度阈值的频繁项集，这个过程包含两个步骤：（1）找出交易数据库中所有支持度满足支持度阈值的频繁项；（2）找出频繁项集中置信度大于置信度阈值的项集，即寻找强关联规则。相比于步骤（1），步骤（2）在执行上所需的内存、I/O和时间都比较小，因此，主要的工作是如何高效地从巨大的数据集中挖掘出频繁项集。本文使用两种方法实现频繁项集挖掘。

（1）采用FP-树频集算法。针对Apriori算法的固有缺陷，采用FP-树频集算法分而治之的策略，在经过第一遍扫描之后，把数据库中的频集压缩进一棵频繁模式树（FP-tree），同时保留其中的关联信息，随后再将FP-tree分化成一些条件库，每个库和一个长度为1的频集相关，然后再对这些条件库分别进行挖掘。

（2）采用Lucene实现频繁项集。Lucene是一个全文检索引擎框架，框架内包括完整的检索及索引模块。利用Lucene的工具包，使用该工具包构建事务（关键词、文摘号）库T的索引功能和索引检索功能实现关联算法的频繁项集挖掘。

3 实验结果分析

3.1 搜索词推荐系统界面比较

首先将本推荐系统与国内著名的学术信息检索系统“CNKI中国知网”和“百度学术”的搜索词推荐功能界面进行比较。实验数据都采用期刊库，检索字段采用关键词，用户检索词采用“知识管理”。“CNKI中国知网”检索推荐界面如图2所示，“百度学术”检索推荐界面如图3所示，本系统的关联关键词推荐界面如图4所示。

图2 “CNKI中国知网”检索词推荐界面

图3 “百度学术”检索词推荐界面

图4 本文的关联关键词推荐界面

3.2 搜索词推荐系统功能比较

3.2.1 推荐词对用户搜索词范围限定功能测试

（1）“CNKI中国知网”的推荐词。用户使用“知识管理”检索，“CNKI中国知网”获取10 938条结果；采用系统推荐词“知识管理系统”检索，获取2 117条结果；采用系统推荐词“知识管理模型”检索，获取269条结果。由此可见，“CNKI中国知网”推荐词有限定检索范围的功能。

（2）“百度学术”的推荐词。用户使用“知识管理”检索，获取相关结果约352 000个；采用系统推荐

词“知识管理系统”检索，获取相关结果约1 100 000个。使用推荐词获取的结果数量是使用原检索词获取结果数量的3.125倍。使用系统推荐词“知识管理模型”检索，获取相关结果约257 000个。由此可见，“百度学术”推荐系统不具有对用户搜索词进行范围限定的功能。

3.2.2 推荐词与检索词组配功能测试

（1）“CNKI中国知网”的推荐词。使用系统推荐的组配检索式“知识管理-显性知识”进行检索，获取结果数为0，即“CNKI中国知网”的推荐词没有与用户检索词进行组配检索的功能。

（2）“百度学术”的推荐词。使用系统推荐的组配检索式“知识管理-显性知识”进行检索，获取相关结果约25 300个。“百度学术”的推荐词具有与用户检索词进行组配检索的功能，但组配检索不止限定在关键词字段，而是限定在标题、关键词、文摘等多个字段。

3.2.3 “CNKI中国知网”“百度学术”与本推荐系统功能比较

（1）“CNKI中国知网”和“百度学术”的系统推荐功能具有以下特点：①推荐了包含用户检索词的左右字符串扩展词汇；②扩大了检索词的各种字面表达形式；③推荐词与用户搜索词检索范围无关；④推荐词不是对用户搜索词进行精确限定；⑤选择推荐词可能会偏离用户最初目标。

（2）本系统的关联关键词推荐的特点：①界定了检索词所覆盖的文献范围，即确定了标引检索词的所有文摘号；②揭示了信息检索系统内部关键词多元关联组配关系和数量；③揭示了内容层面的关联关系；④实现了细粒度的精确检索；⑤实现了用户按关联度选择所需信息。

通过上述分析可以看出，本系统的关联关键词推荐功能可限定搜索词的推荐域，针对用户搜索词的推荐域给出了推荐域内的关键词组合信息。这样解决了系统检索的透明性问题，因而提高了检准率；解决了试探性检索转为选择性检索的方法问题，缩短了检索操作的时间，减轻了用户的检索压力。

4 结语

本文采用基于用户搜索词的关联关键词推荐方法，提高检索系统的透明度，提升信息检索的精确度。与目前学术信息检索系统的推荐功能比较，本文基于学术信息检索系统进行的数据挖掘和关联关键词推荐方法，优点在于针对用户的搜索词，系统以关联关键词组配的方式进行推荐，在解决信息检索系统的透明性问题、提高海量文献信息检索的准确度、减少用户信息检索的压力方面，做了初步的尝试。

参考文献

[1] 孙建军.信息检索技术[M].北京:科学出版社,2004.

[2] 秦鸿.关于发现系统的问题与思考[J].数字图书馆论坛, 2012(7):17-20.

[3] 车天文,雷大伟,石志伟,等.一种用户检索词推荐的方法及系统:中国,201310119667[P]. 2013-06-12.

[4] 岑咏华,邓三鸿,王昊.关联推荐及其在学术资源检索网站中的应用研究[J].图书情报工作,2009,53(6):41-45,99.

[5] 孙明.基于语义的信息检索与关联推荐关键技术研究[D].成都:电子科技大学,2015.

作者简介

温有奎，男，1951年生，博士，教授，研究方向：文本挖掘、知识发现，E-mail：wykui123@126.com。

Information Retrieval Method of Association Keywords Recommendations

WEN YouKui[1,2]

(1. Institute of Scientific and Technical Information of China, Beijing 100038, China; 2. Beijing Wanfang Data Co., Ltd, Beijing 100038, China)

Abstract: The current information retrieval system is opaque to the user, the user needs to guess the way to browse the system and repeatedly questioning the results of detection to determine the value of the information. Big data exacerbated the user filter amazing amount of literature which led to mental pressure and time, with the cross-disciplinary, multi-Correlation gradually increased demand for information retrieval, the pressure will be more and more obvious. This paper proposes a keyword associated with the method recommended in the past by the user to solve the conjecture enter a search term covering a large surface, narrow your search by browsing method, the user selects becomes automatic internal keyword group associated with ways to improve the retrieval Accuracy. Experiments prove that keyword association recommended method greatly improves the retrieval accuracy of information retrieval systems, while reducing the user's information retrieval pressure.

Keywords: Keywords Match; Association Recommendations; Information Retrieval

（收稿日期：2016-04-02）