从检索技术的实现方式看三大全文数据库的发展

化柏林 张新民
中国科学技术信息研究所 北京 100038

〔摘要〕通过三个有针对性的检索实例对清华同方、万方数据、重庆维普三大全文数据库检索技术实现的特点进行分析，从其目前提供的功能和招聘信息分析它们在竞争中所处的位置，指出全文数据库的三大核心竞争力是：历史数据回溯、当前数据质量和未来增值服务。进而对未来的增值服务进行分析与预测，包括计量分析自动生成系统、句子级文献自动审稿系统、参考文献自动标注系统、观点型搜索，并指出这些系统的实现将使文献服务真正走向知识服务。
〔关键词〕全文检索 全文数据库 检索技术 全文数据库商 清华同方 万方数据 重庆维普 核心竞争力
〔分类号〕G35 TP391

Exploring the Future Development of Three Major Full-text Databases through Their Information Retrieval Technology Realization Styles

Hua Bolin Zhang Xinmin
Institute of Scientific and Technical Information of China, Beijing 100038

〔**Abstract**〕The features of application of retrieval technique in three major full-text databases are analyzed. The competitive status of each database is concluded by analyzing its system functions and recruit information. Three core competitivenesses of full-text database are history data recall, quality of current data and future value-added services. Further analyzes and predicts the future value-added services, which include automatic generation system of bibliography, automatic reading and examining manuscript in sentence level, automatic indexing system of reference documents and retrieval by viewpoint. The implementation of these services will lead to the mode transformation of service from document service to knowledge service.
〔**Keywords**〕full-text retrieval full-text database retrieval technique full-text database provider CNKI wanfangdata CQVIP core competitiveness

1 引 言

中国期刊全文数据库（简称清华同方）、数字化期刊全文数据库(简称万方数据)、中文科技期刊数据库（简称重庆维普）是国内公认的三大期刊全文数据库。三大数据库之间既有许多相似之处，又各具特色：从检索形式上看，三大全文数据库都支持初级检索、高级检索、专业检索，都支持复杂的逻辑表达式的提问，都支持跨库检索，支持链接导航。但它们也有许多不同之处，如清华同方支持真正的全文检索，而其他两家目前还不能做到这一点。

目前，讨论全文数据库使用技巧及存在问题的文章非常多，但对全文数据库检索技术的实现以及未来的发展趋势的探讨还很少。本文作者既进行过常规的手工操作，也写过进行自动下载的下载程序。通过对数据库的不断测试，特别是通过某些特殊的、有针对性的系列检索，判断出系统所使用的一些基本的检索技术。进而在对这些技术探讨的基础上，分析并预测这三大全文数据库商的竞争优势及未来的发展方向。

2 通过针对性的检索实例分析检索技术的实现

目前，信息检索大都致力于对主题检索的支持，却忽略了利用特征与结构的检索。向量分词检索在索引空间、检索效率等方面优于字符匹配型，但检索质量并不总是高于字符匹配。例如，想查找某一期刊连载的文章。这种检索需求不是主题相关的，而是从特征入手。通常，连载的文章标题后有（上）、（下）或（一）、（二）或（I）、（II）、（III）等字样。笔者于2006年12月19日分别对清华同方、重庆维普、万方

收稿日期：2007-06-18 修回日期：2007-06-29 本文起止页码：13-16 本文责任编辑：杜杏叶

数据的期刊全文数据库在标题检索里精确匹配“(上)”，检索结果分别为301 137，269 222和61条。查看结果后发现，清华同方和重庆维普都滤掉了括号，不支持括号作为检索条件，而只有万方实现了用户的检索目的。这就说明前两者是基于关键词的索引，而万方数据并没有使用关键词索引，而是使用单字符索引。

在另一个机构检索实例中，对重庆维普，以“机构=北大*核心期刊*年=1989-2006”进行检索，得到23 839条检索结果，结果中主要包括西北大学、东北大学、河北大学、湖北大学等，说明重庆维普的机构检索是字符匹配。而从清华同方的数据库中限定1989－2006年期间的核心期刊，选择模糊匹配得到了33 027条结果，而精确匹配则只有7条。在通过精确匹配所得到的7条检索结果中，作者单位都是直接写的“北大”，事实上这是一种不规范的写法。也就是说，如果这几名作者使用规范的机构名称的话，使用清华同方精确匹配的检索结果将为0条。而在万方期刊论文库以“1989－2006期刊机构=北大”作为检索条件，得到43 073条检索结果，说明它也是使用字符匹配。也就是说没有一个数据库使用同义词，能够让用户输入“北大”，也能把“北京大学”检索出来。万方数据与重庆维普的机构检索都没有采用分词，直接使用的单字符索引或like检索，如果采用分词的话，就不会出现这种情况。清华同方的模糊匹配结果很多，精确匹配结果却很少，说明模糊匹配采用的是全字符索引或like检索，而精确匹配却是分词后的索引。

在第三个检索实例中，在中国期刊全文数据库（清华同方）进行检索，检索范围是1980-2007年的全部数据，检索条件为在篇名中精确检索“图书”，得到21 630条检索结果，而把检索词换成“图书馆”后，检索结果却变成了95 636条。“图书馆”包含“图书”字样，按常规思维，“图书馆”的检索结果应该比“图书”的检索结果要少，可事实上检索词变长了，检索结果却多了。这说明该数据库的篇名检索采取的不是全字符切分，检索时采取的不是单字索引，而是向量切分，可能是正向最大向量切分。如果进一步把检索词拉长，变成“数字图书馆”，检索结果又变成了4 885条，也就是说检索词变长了，检索结果却又变少了。“图书馆”肯定包含“图书”，“数字图书馆”肯定包含“图书馆”，同样的现象，却有不同的结果。为了进一步验证，把检索词换成“数据”，检索结果变成75 792条，而改成“元数据”，检索结果又变成了958条，检索词长了，检索结果却变少了，由此断定该数据库采取的不是正向最大向量切分，而是逆向最小向量切分或逆向最大向量切分。因为如果采取的是正向最小向量的话，“图书馆”的检索结果就不会比“图书”的检索结果多。事实上，根据汉语中心语靠后的特点，逆向切分比正向切分的准确率要高得多。而万方数据与重庆维普都是遵循词条变长、结果一定会少的原则，因此可以断定是全字符索引。实验过程与数据如表1所示。

表1　期刊全文数据库标题检索索引方式实验数据
（检索日期：**2007-05-29**）　（单位：条）

检索词	清华同方	万方数据	重庆维普
图书馆员	3 234	2 039	3 185
图书馆	95 636	45 571	83 846
图书	21 630	50 778	98 927
数据	75 792	56 043	100 714
元数据	958	881	1 084
DC元数据	48	45	49

为了进一步验证索引的方式，采用二次检索来比较结果的变化。用清华同方数据库标题检索“数据”得到75 792条检索结果，从结果中再进行标题检索“元数据”与“元”分别得到958条与1 207条，说明对元数据这个词做了三个索引：元、数据、元数据。用“图书馆员”进行标题检索得到3 234条检索结果，在结果中再检索“图书”，只有9条，这9条记录的标题中既含“图书馆员”，又含“图书”（独立于图书馆员）；在结果中检索“馆员”，只有21条，情况与图书一样；而在3 234条结果中分别检索“图书馆”与“员”，其结果都是3 234条，说明系统把“图书馆员”切分成“图书馆/员/”而不是“图书/馆员/”；而在图书馆的95 634条结果中二次检索“图书”，得到1 425条，说明没有把“图书馆”切分成“图书/馆/”。为了进一步的验证，采用更长的词条进行测试，表2的结果说明“网络信息计量学”被切分成了“网络/信息/计量学/”。如果采用最大向量切分，则不能切分出“网络/信息/”，如果采用正向最小向量应该能切出“计量”。计量学与图书馆一样，却与元数据不一样，把“元数据”切成“元/数据/”，却不把“图书馆”切成“图书/馆/”，说明肯定是逆向向量切分，而且是二次嵌套切分。此外，检查结果还证明系统没有使用MMC（基于上下文的最大向量匹配）进行切分。

3　从文献服务走向知识服务

从文献服务走向知识服务的理念已经得到了广泛认可，

表2　期刊全文数据库（清华同方）标题检索二次检索结果

	检索词	检索方式	检索结果(条)
1	网络信息计量学	初次检索	31
2	网络	在1的结果中检索	31
3	信息	在1的结果中检索	31
4	网络信息	在1的结果中检索	31
5	计量	在1的结果中检索	2
6	计量学	在1的结果中检索	31
7	网络信息计量	在1的结果中检索	1
8	信息计量学	在1的结果中检索	31

但实现什么样的知识服务、如何实现真正的知识服务才是未来图书情报机构获取核心竞争力的关键。中国学术期刊网(www.cnki.net.cn)改名为中国知网，由提供学术文献服务向提供知识服务进军，以引文链接、学术定义、学术趋势等新功能为起点，以回答学术问题、打破以篇为单位的知识组织方式，提供知识点与知识点之间的链接为目标，旨在实现真正的以知识点为处理单元的知识服务，也就是从物理层次的文献单元向认知层次的知识单元转换[1]。除了学术定义外，它还会抽取历史发展、分类、特点、方法、关键技术、国内研究进展、国外研究进展、应用前景、实验数据、实验结果等诸多内容，即按照写作的结构把文章所有内容进行模块化抽取；万方数据于2006年推出了知识链接门户，通过作者、分类号、关键词字段等提供文献之间的链接；重庆维普也推出了知识频道。

数据库商纷纷使用知识概念为知识服务造势，虽然可以从服务观念和目标定位上进行强化，但要实现真正的知识服务还有很长的路要走。实现这一转变的根本是人才，所以通过三大数据库商对人才特别是研发工程师的需求也能看出他们的研发重点和相应进展:①通过招聘信息可以发现他们的研究计划。例如，清华同方的招聘信息中进一步强化需要CNKI文献搜索产品、知识元搜索产品和各种垂直搜索产品的数据采集加工、整合更新和系统开发人才，重点解决文本挖掘、中文信息处理、知识系统等[2];万方数据招聘研发工程师的要求是跟踪信息技术发展，在信息检索、文本挖掘等相关研发领域开展研发工作[3]; 而重庆维普的网站上没有发布招聘研发工程师的信息。②从公司招聘研发工程师的描述中可以看出研发深度的不同。在不涉及商业秘密的情况下，研究专业方向越具体，研究内容越深入，公司的研发力量就越强，推出的产品价值也就越高。因此，从招聘信息可以推断出，目前清华同方的技术研发似乎强于万方数据，而万方数据又强于重庆维普。事实上，通过他们所推出的增值服务，也就是新功能也可以验证这一关系。例如清华同方的知识链接（引文分析）的推出早于万方数据的知识链接。近期清华同方又推出了学术定义、图表搜索、搜索趋势等，这些服务已经开始对文本特别是正文内容进行分析，并充分利用了信息抽取技术。计量分析自动化已经实现了数值的统计计算，尽管还没有实现计量分析报告的自动生成，但以目前的势头来看，相信在不久的将来，也会实现的。③从清华同方招聘研发工程师的任务和要求来看，已不再关注传统的信息检索技术，而是要解决文本挖掘、信息抽取等问题，以实现自动分类与聚类、自动摘要以及问答系统等目标。

4 全文数据库的三大核心竞争力

期刊全文数据库的竞争主要体现在三个方面：历史数据的回溯；当前的数据质量；未来的增值服务，即数据的深加工程度。

对于历史数据的回溯建库问题，目前中国知网走在前列，很多期刊已经回溯至创刊号。历史数据除了扫描全文外，重点是关键词的提取与摘要的自动生成以及自动分类（或归类），因为在20世纪90年代前，很多期刊的文章都没有关键词，更不用说摘要了。

当前的数据质量主要反映在数字化程度的比例。就是从编辑部那儿得到多少篇纯电子版文章，而不是利用纸版进行扫描。因为只有用纯电子版，才有可能提供真正的全文检索，如果不能对正文字段进行检索，那检索只能称之为假全文检索，因为没有比正文字段更能反映文章内容的了。如果没有电子版的数据，增值服务也就无从谈起。

这些增值服务除了提供更好的检索服务外（如中英文摘要语料对齐后的双语检索），还包括（但不仅限于）以下的应用：

- 提供文献计量自动分析的查询，如清华同方已推出的“中国学术期刊文献评价统计分析系统”。
- 提供学术调研报告的自动生成，在文献计量自动分析的基础上，对国内外某领域进展情况进行评述。学术调研评价可以指导论文的选题和前期调研，特别有利于论文选题、项目评审等工作。
- 可以从句子级提供文献自动审稿辅助功能以及参考文献自动标注功能。
- 提供更小粒度的检索，支持句子检索、真正的图片检索（首先是流程图、系统结构图、数据表等的检索，以后会支持图像检索，从颜色、纹理、形状等各个要素进行分析），大量使用信息抽取技术，提供列表式搜索。
- 支持学术问答，支持观点型搜索、流派型搜索，能够提供学者谱系图，利用学位论文的致谢提供导师自动评价系统。
- 提供知识点与知识点之间的链接，实现真正的知识服务。正如由过去买本整刊进行阅读到现在的只看某篇文章，将来可能实现只看某篇文章的某一部分。

5 增值服务是核心竞争力的核心

文献计量自动分析系统，可以统计分析任意一个学科、专业或方向的核心作者，主要研究机构，地域分布，关键词、标题、文摘及分类号的关系，提供研究热点及趋势等统计分析，以Top N、统计图表等形式提供给用户，并用文献计量的定律来进行验证。而现在的计量分析方面文章大都是由人来写的，而且主要分布在图书情报领域。其实自然科学领域也非常需要他们本学科的文献计量统计分析，如果能有这样的一个自动统计分析系统，会为科研人员节省很多时间和精

力，为研究工作提供很大方便（关于计量分析的技术实现请参阅文献[5-6]）。

当前，信息爆炸与信息泛滥的问题日益突出，解决的根本方法是使大量创新性很低的文章没有发表的可能。为此，编辑部会使用“学术抄袭与科学引用自动判定系统”辅助审稿，从而在源头上利用技术手段解决学术抄袭的腐败问题，而这种系统可以由全文数据库商联合提供。这类句子级分析匹配系统既可以对学术抄袭与科学引用进行自动判定，同时也可以帮助作者进行参考文献的自动标注。句子匹配分析系统的难点主要表现在：异构数据的获取；历史数据的回溯建库；跨语言之间的判定。

现在的数据库商以篇为单位提供数据，未来的数据库商不仅能提供句子级的搜索与分析，还能提供以知识点为单位的搜索与分析。将来的系统还将会提供学习型搜索和观点型搜索：①学习型搜索相当于文献自动综述，对于现在的检索而言，如果用户不打开检索结果进行全文阅读就很难判断哪些文章是需要的，哪些文章是不需要的，而看过的文章又有许多重复的内容。如果能够让计算机进行滤重与知识重组来完成这一工作，那将是一件非常有意义的事情。未来的搜索将可以实现知识的重组，把上千篇文献组织成一篇，相当于以百科全书的形式进行组织，用户只需要看“书”中感兴趣的部分就可以了。②观点型搜索是指根据某观点进行搜索，以自然语言形式输入查询，搜索含有某个观点的文章，或者关于某个知识点的所有观点。未来的检索结果将不再是一篇一篇的文章，而是一个列表。列表列出每种观点以及每一种观点的支持人数。当我们想详细了解某一种观点时，就点击相应记录，系统会显示关于这种观点有哪些论述方式，是如何来论述的，也就是真正的知识链。这是解决信息泛滥与知识贫乏的关键途径[6]。

6 结 论

清华同方率先实现了真正的全文搜索，并推出了参考文献与引用文献相关的文献链接，今年又推出了一系列的学术服务，这也是CNKI的价格较高同时市场占有率也很高的原因之一。因此，数字化程度与数据深加工程度将是数据商的核心竞争力的反映。同行数据库商之间的竞争除数据质量外还有：在范围上体现为对历史数据的回溯程度——对历史数据回溯得越多，数据就越有竞争力；在深度上将反映在对数据的深加工程度上，对数据加工的越深越细，将会越有竞争力。

未来针对全文数据库的应用系统可能会很多，新功能的名字也可能有很多不同，但整体上会朝着以下几个方向发展：分析粒度越来越小（句子分析是重点和核心），分析数量越来越大（大规模异构数据综合分析），分析范围越来越广（正文内容分析成为重点），分析程度越来越深（不再以词为重点，会支持结构检索、语用检索等）。

自动问答、信息抽取、列表式搜索、观点型搜索等都是一些新的趋势，但是搜索技术要取得突破性进展，知识获取无疑是关键。三大全文数据库商拥有如此丰富而权威的资源，在这些文献中蕴含着大量的专家知识，如果能把这些知识都抽取出来，就可以进行学术的自动问答了。未来的IT用户所强调的不是拥有技术，而是拥有可以用的知识。发挥计算机的速度优势主要依靠算法，发挥计算机的存储优势主要依靠知识库。建好人用知识库可以解决很多问题，如果这样的知识库（如CYC）同时还能为计算机所用，那么许多问题便会迎刃而解。拥有这样的知识库必将引领未来的IT，必将印证“得资源者得天下”的道理。

参考文献：

[1] 马费成.情报学的进展与深化.情报学报,1996,15(5):338-344.
[2] [2007-05-23]. http://www.cnki.net/cpyc/cpst.htm.
[3] [2007-05-23]. http://soft.wanfangdata.com.cn/Us/job.aspx.
[4] 化柏林.用VBA实现计量分析研究中的数据预处理技术.现代图书情报技术,2007(3):69-72.
[5] 化柏林.用VBA剖析计量分析研究中的统计分析技术.现代图书情报技术,2007(4):70-74.
[6] 化柏林.从IPO分析未来的搜索引擎.情报学报,2006,25(S1):351-355.

〔作者简介〕 化柏林,男,1978年生,助理研究员,硕士,发表论文15篇;
张新民,男,1970年生,副研究员,博士,发表论文43篇,译著1部。

(上接第118页)

上文中讨论的 h_m 指数实际上已经承认在 h 指数相同的情况下，具有更高总被引次数的人更加杰出，所以在具体计算时需要知道待评作者的总被引次数。而若选用其它标准，如认为在 h 指数相同的情况下，发表作品总数更多的人更加杰出，则只需将式（2-1）中的 $N_{c,tot}$ 换成作品总数 N_p 即可，其它情况同理。由于 h_m 指数的定义式中，总被引次数 $N_{c,tot}$ 位于分母，这就要求这个数字不能为0，所以对于那些从未得到过引用的作者（即 h 指数 $=0$ 且 $N_{c,tot}=0$），将无法使用 h_m 指数进行评价，这时可选用其它评价指标，如发表作品总数等来实现评价目标。

参考文献：

[1] Hirsch J E. An index to quantify an individual's scientific research output. Physics.[2007-03-06] .www.pnas.org/cgi/doi/10.1073/pnas.0507655102.

〔作者简介〕 张学梅,女,1978年生,助理馆员,硕士研究生,发表论文5篇。