●马张华（北京大学 信息管理系，北京 100871）

后组式检索系统的组织体系研究

［关键词］后组式系统；组织体系；关键词搜索引擎；文本检索系统

［摘　要］针对后组式检索工具研究中存在的局限，提出后组式检索系统的组织体系必须根据组配特点，结合其与组织相关的基本构成成分及要素加以考察，并以关键词搜索引擎为例，分析了构成成分及其特点，考察了贯穿在其中的基本要素以及研究内容，认为传统信息组织领域的专业人员应结合长期探索的积累，在后组式系统的研究中发挥应有作用。

［中图分类号］G354
［文献标志码］A
［文章编号］1005—8214(2008)04—0029—04

随着网络搜索引擎的大量使用以及文献数据库的日益发展，文本检索这类后组式系统已经逐步发展成为信息检索的主流形式，并在组织成分、构成要素、应用形式等方面经历了一系列的发展。但迄今各国对信息组织的研究，主要仍集中在先组式系统上，对于后组式系统，一般停留在控制词表的构建和应用特点等的讨论，而缺乏对其组织系统的完整研究；许多与组织相关的内容，被作为检索问题讨论，使得信息组织的研究被限制在一隅，不能有效与计算机环境下的发展结合。这种情况严重限制了对后组式系统的研究，制约了对主题法，尤其是文本检索系统的组织规律和方法的探索，使信息组织失去了研讨当前发展最为迅速、最有生气的关键领域的机会，这对完整进行信息组织的研究是致命的。本文试图以关键词搜索引擎等为例，在分析其组成的成分、特点、构成要素等的基础上，对后组式检索系统的组织体系的特点进行讨论，希望能引起学界的重视，逐步从信息组织的角度加强对后组式检索系统的研究。

1　后组式检索系统的组织特点和构成

众所周知，传统先组式检索系统的特点是先组、定组、显性。以等级列举式分类法为例，其类目体系是用户检索以前就预先组配好、句法关系确定、并可以加以完整显示的。而后组式检索系统，包括关键词搜索引擎、文献数据库等建立的文本检索系统则恰恰相反，表现为：

•后组。不像先组式检索工具那样，预先建立完备的系统，其组织体系是在检索阶段结合用户的检索提问形成的。

•自由组配。不像先组式检索工具例如分类法那样，按照预先设定的结构展开体系，而是根据检索的需要建立起基本的组配模式，可以根据用户的检索需要灵活组配，存在无数种组配和构造的可能性。

•隐含。不像先组式系统那样，可以对形成的组织系统加以完整显示，而是隐含的、只有在检索后才能显示相应的部分。尽管就其能力而言，它存在着无数种检索和揭示的可能性，但在实际使用中，这一系统只显示与检索相对应的那部分内容，不可能也没有必要对其组织体系进行完整的显示。

在过去的某些文献中，人们往往将这类组织系统的基础成分，例如文献索引库的构成作为研究的基本内容，将它对应于先组式系统的体系，这种认识至少是不完整的。实际上，不同组配方式的系统，其组织体系的构成特点和要素是不同的，以分类法为例，对传统分类法的组织系统可以通过详细列举的类表加以了解；但在分面分类表中，这些内容则表现为分面单元概念表和引用次序。而要了解文本检索的组织系统，必须根据后组式系统的特点，结合其与组织相关的基本构成成分及要素加以考察。以关键词搜索引擎为例，其直接构成部分包括采集模块、存储模块和检索模块。其中，采集模块是组织的前提，决定资源的组织的处理对象；存储模块中的索引及相关工具是组织的基础结构，是检索提供的基础条件；检索模块中，检索界面及其采用的检索提供方式，是根据用户需要确定的实施组织的条件和形式；此外，种类多样的检索优化形式则为系统提供了优化重组机制。上述几个方面的结合，构成了从基础结构到检索形式和提供方式等的完整内容。而贯穿在整个组织和检索操作过程中的、对检索结果起影响的组织要素则是其词法、句法，以及与资源组织相关的其他因素，包括链接因素、用户点击因素等。关键词搜索引擎的组织特点、构成成分及其相关因素可以表1的方式简单表示如下。[1]

应考虑系统的知识化服务能力的建设，从资源组织到系统设计，全面提升整个系统的知识化服务能力，由一个数据库检索平台逐渐向专业知识服务平台迈进。

［参考文献］

[1] THOMSON PHARMA［EB/OL］.［2006－06－12］. http://www.thomsonscientific.com/media/Pdfs/pharma-infopack.pdf.
[2] 万方医药信息系统镜像网站［EB/OL］.［2006－06－15］. http://wanfang.llas.ac.cn:90/.
[3] Rachel Buckley. The Life Sciences Information Challenge-Focussing Multidisciplinary Delivery in a Information Environment［EB/OL］.［2006－06－15］. http://www.infonortics.com/ch04/chemical/slides/buckley.pdf.
[4] 万方数据网站［EB/OL］.［2006－06－15］. http://www.wanfangdata.com.cn/.

［作者简介］马建玲（1969－），女，毕业于武汉大学信息管理学院，中国科学院资源环境科学信息中心信息技术部副研究馆员，发表论文多篇。

［收稿日期］2007－10－22　　［责任编辑］王　岗

表1　关键词检索系统的组织特点和影响因素

组织特点	—后组。根据检索需要实时组配显示，没有预先建立的完备显示的系统。 —自由组配。存在多种组织系统。 —隐含。组织系统是隐含的，检索后才能显示。
构成部分	—采集机制。采集器及接受递交窗口。 —基础机构。索引及相关工具。 —检索机制。检索界面和提供形式。 —检索优化机制。提供优化重组的形式。
组织要素	—词法 —句法 —其他相关要素，如链接、用户因素、技术应用等。 —算法

可以看出，这类检索系统的组织体系不像先组式系统那样，是一个预先建立的显性系统，而是上述构成成分和组织要素的综合。上述的特点也决定了，在这样的系统中，组织系统是与检索密切结合的，尤其是其中的检索优化机制，往往是在检索的基础上动态形成的。因此，要探索这类系统组织体系的规律和特点，不能如对先组式系统那样，通过对显性系统的分析加以了解，而必须根据它们的特点，对其构成成分、组织要素及其运行规律等进行考察。

2　后组式检索系统的构成及其特点

常见的后组式检索系统包括主题词检索系统、以文献数据库为对象的文本检索系统、关键词搜索引擎等多种类型，尽管不同的系统在具体构成成分上存在着差异，但在整体结构方面并无本质的不同，其中，以关键词搜索引擎的发展最为充分，其组成部分及其特点最具有典型性。关键词搜索引擎的基本构成部分涉及到：[2]

•采集模块。包括建立供人工操作的网络编目平台和开发自动采集软件，如 crawlers、robots 等。后者采用与人相似的方式，访问和下载软网页资源，通常从一组范畴名开始，访问主页，同时下载主页中的链接数据，扩展对网页的访问。采集模块一般须根据系统的特点确定搜索范围，制订搜索策略等，用以规定系统组织的资源对象。

•索引及相关工具。通常建立多种类型的索引，包括顺排索引、倒排索引、链接索引、各种实用索引等，同时，还发展检索日志及相应的检索词典等多种相关工具。[3]关键词搜索引擎中索引的常见构成成分可以简单归纳如表2。可以看到，关键词搜索引擎的索引系统不仅容量大，构成成分也远比传统文献数据库充分，为系统的组织和检索提供了适用的基础。

•检索界面和检索提供形式。基本上继承了传统数据库检索界面的形式，其发展是：其一，定型化，确定了简单检索、高级检索、专类检索等基本检索方式；其二，重视易用性，如在简单检索中设置默认的组配检索方式和自然语言语句转换机制、在高级检索界面采用易于操作的组配检索形式等；其三，提供多因素结合的可能性，如高级界面中提供多种限定检索设置，便利用户进行复杂检索；其四，在检索结果返回时，采用检索匹配加权的形式加以排序显示，在保障检全率的情况下，提高检准率。这些努力较好解决了搜索引擎的组配句法，以及显示形式等问题。

表2　关键词搜索引擎常见构成成分

索引类型	索引对象	补充说明
顺排索引	顺序记录资源文本。通常是全文。	
倒排索引	词后记录网页 ID 和位置信息，利用标识语言，记录附加信息，如，粗体字（以标签＜B＞标注的字），标题（以标签 <H1> 或 <H2>标注的字），包括锚定文本、URL中的词等。	传统系统不包括锚定文本、URL 中文字。
结构索引	即链接索引。描绘为点和边的图，记录网页之间的链接。可作为判断网页间联系的依据。	传统系统中接近的形式为引文索引。
实用索引	结合查询界面提供的实用检索功能建立的索引类型。如，系统提供域名检索、网站检索等即意味着需要建立相应域名、网站等的索引。	
检索日志	记录用户查询操作，可用来优化检索提供。	
词典	收入所有的索引中的词，包括各种控制词表等。	

•检索优化机制。指以交互方式对用户查询提供新的选择方案或将用户的检索结果加以重组，以改进检索效果。这类方式虽然在传统检索系统中一直存在，但未得到充分开发。网络资源的特点和终端检索的需要，使其迅速发展成为一个受到广泛关注的领域，如，以“similar to”提供相似文献；利用用户检索查询，提供查询优化；在返回结果的基础上聚类，作为二次检索依据；此外结合用户信息进行个性化提供等形式正在逐步发展之中。

可以看到，关键词搜索引擎的组织成分具有多元的特点，并且是随着资源和使用的需要发展的。上述构成成分是根据网络资源组织的特点决定的，同时也具有一定的普遍性。如各种文本检索系统往往也都采用类似索引及词表等工具，有的并纳入引用关系等改进相关揭示。与传统先组式检索系统相比，关键词搜索引擎的特点表现在：

首先，其组织匹配的中心是由以文本词汇为对象的倒排索引，正是它提供了以语词为中心的匹配基础，从而不同于基于先组式体系的分类或标题组建的文档。

其次，它是通过词汇的附加信息，如结合标识语言记录的结构信息，语词的位置信息，以及链接索引等多种索引工具的结合应用，对检索词价值和网页有效性进行遴选，改进组织和检索效果的。多因素的结合是文本检索和网络资源的特点决定的，虽然先组式系统也可以结合多种因素进行检索提供，但其迫切性远不如文本系统。

其三，其组成的整体构建起了完备而又适用的组织机制。其中，索引及相关工具提供组织匹配的对象；组配和检索优化机制提供匹配句法；采集模块则用于解决组织对象的选择，三者结合构成了一个不同于先组方式的组织框架，从而可以动态地、更加灵活地对资源进行组织和提供。

显然，后组式系统的构成特点不同于先组式系统，如何根据检索系统的资源和应用环境确定各个部分的构成，探索相应成分的规律及技术方法，并按照对各部分相互关系的了解加以优化，是这类检索系统应解决的基本内容。因此要有效进行这类系统组织体系的研究和构建，至少应涉及以下几个层面的内容，如：

•根据资源的特点和检索需要，对各部分构成成分及特点、作用、构成规律等的探索。如在存储模块中，对索引及其相关工具的类型、构成及其规律等进行研究；

•各个构成成分的技术方法研究，如，存储模块中文本索引构建技术的研究，自动采集、检索优化领域的各种技术方法的研究等；

•各个组成部分之间关系及其相互影响与作用等的研究。如，检索界面的设置和检索优化机制是与索引类型和词典的特点相适应的，同时也会反过来影响索引和词集的编制；再如，采集决定了索引和检索的对象，同时它本身又是随着用户需求和索引技术的发展不断调整的，检索日志等的数据除作为排序因素加以应用以外，也会同时影响采集方针的调整等。关键词搜索引擎各个组成部分之间的影响可简单列举如表3。

表3　组成成分之间的相互影响

基本模块	成分	相互影响
采集模块	采集器、网络编目平台	决定索引的对象，同时也受用户检索需求和索引技术水平的影响。
存储模块	索引与词表等相关工具	资源组织和提供的基础，索引技术的水平影响采集对象的选择。同时其处理对象、特点受采集对象特点、检索使用需要、检索优化形式等的影响。
检索模块	检索界面和提供形式	在索引的基础上进行，同时也反过来影响索引的编制。
	检索优化机制	受索引、词典等的影响，反过来也影响对它们的编制。

上述研究内容虽然不同于先组式系统，但许多是与传统信息组织的理论方法密切联系的。例如，上述各部分中，索引文档以及词表的构建，即与词汇控制相关，而检索界面的设置，则主要是解决组配句法问题，都是传统信息组织关注的基本内容，只是在后组环境下研究探讨而已；有些内容，如自动采集、检索优化中采用的一些新技术，突破了传统应用形式，也应是信息组织在新环境下的应有之义，是应当拓展和了解的。而其中，影响系统建立和运转的关键内容之一，则是贯穿在各个组成部分之间的检索要素。

3　检索要素研究

检索要素是贯穿在检索过程中，改进检索效果的重要因素。随着计算机处理能力的增强，多种检索要素的结合正在日益成为文本检索系统组织和检索的基本特点。某种程度上，对要素及其应用规律的研究，直接影响到文本检索系统的应用水准，因此是这类系统研究中应予关注的重要内容之一。

表4列举了网络关键词检索系统中涉及的常见检索匹配因素。表中前7项均为文本词汇因素，包括查询词匹配数量、匹配位置、匹配单元、分解因素和反文献频率等，同时还大量结合了文本词汇以外的因素，如链接因素、用户因素等。显然，网络资源缺乏质量控制等因素增加了相关因素应用的迫切性。S.Brin，L.Page [2] 说明引入PageRank的原因时提到，在单纯采用词汇匹配作为资源排序依据的搜索引擎中，多数排列在前的资源虽然有较高相关度，但其本身并没有使用价值。将链接因素作为网页重要性的判断指标，则可以在一定程度上解决资源有效性问题。目前，多种因素的纳入已成为一种共识，包括结合用户使用情况，作为个性化服务的依据等。给出新网页的新鲜度数据，则是为平衡新资源链接、用户点击数等的差距所提供的一个调整值，此外，还应包括对于各种商业因素的排除。

表4　网络检索中检索匹配涉及的要素

—查询词匹配数量
—多个查询词匹配的完备程度
—匹配单元和分解问题
—匹配词的接近程度
—网页中术语的位置
—术语所属的成分，例：<title>、link tent、body text
—网页词频与总词频之比
—指向本页的链接分析
—点击分析
……
—对新网页，结合考虑新鲜度
排除商业因素。例：规定凡与人为增加检索要素的网站建立联系的资源，一律不予排序。

显然，各种基本要素的纳入是与相应要素及其应用规律的研讨密切联系的，对词汇控制、链接规律、用户日志等的研究，是将要素有效纳入的重要条件。自20世纪90年代网络搜索引擎出现以来，对基本因素的研究取得了许多进展，但仍然存在进一步改进的空间。仅以词汇控制研究为例，至少涉及到：

•组织和检索的语词单元问题。如在中文系统中，如何在以单元词汇为基础的同时，适当纳入词组，包括文本环境下词组的发现和以有效结构收录、应用等。

•词间关系控制问题。如文本环境下同义词相关词识别方法的研究，以及词汇控制应用方法的探讨等。

•检索句法问题。如基本组配模式的设置和优化，组配成分的权值设定，以及对自然检索语句的有效识别和切分等。以自然检索语句的处理为例，其分解就涉及分解单元、分解层次、分解策略等方面的因素，目前的关键词搜索引擎，如百度、中文google、中文yahoo!等的处理方案都仍存在进一步改进的需要。

•词汇控制与多种相关因素的结合应用问题，例如，词汇控制与链接因素、用户因素的有效结合和权值的合理设定问题。

•词汇控制在多种环境下的应用规律问题，包括各种形式自动标引中的应用、自动文摘编制、相似文献的提供，一直到结合用户需求情况下的知识发现等。

有人根据Google等不采用截词检索，认为网络检索中检准率是主要矛盾，同义控制的意义不大。但实际上，尽管多数搜索引擎不直接使用同义词检索，但一般并不排斥结合词间关系控制来改进检索效果。常见的如：

•将同义控制、相关控制等作为检索扩展的选项。如在使用“百度”检索“北京大学”时，其相关搜索栏同时提供北大、pku等在内的待选词，供必要时选用。

•作为容错检索手段。只要收入检索频率高的常见错误检索词，并将其与对应的检索词加以联结就可以了，这实际上也是等同关系词汇的特殊应用形式。

•作为检索优化处理的依据。如在动态聚类中对同义词、等级关系词等加以控制，以便在概念层次上实施聚类操作，改进聚类效果。

•作为进行自动标引的手段。例如，一些检索系统往往对网络资源进行自动分类或主题词标注，结合词汇控制有助于改进自动标引的效果。

此外，各种知识组织系统的应用也是改进文本系统检索效果的重要方式。

显然，在词法、句法等的研究和应用方面，目前仍然有大量的工作要做，文献领域工作者长期积累的对词汇控制规律的了解，是推进词汇控制应用的重要力量。同样，对链接因素、用户因素和其他相关因素的研究和纳入也需要类似的努力，如计算机界已经对链接因素、用户因素等结合进行了许多探索。[4,5] 近年来相关因素纳入的另一个例子，是在检索返回资源的显示中，将文献类型，如百科全书词条、个人主页、机构官网等作为改进排序的因素之一，取得了较好效果。如何在检索系统中发现并有效纳入这类因素，仍是这类系统关注的一个内容。

关键词搜索引擎的处理技术为文本检索系统的改进，提供了十分有价值的经验。从表5可以看出，传统文献数据库的数据特点与网络资源既有不同又有相似之处。事实上，目前在一些文献数据库中，已开始逐步将网络关键词搜索引擎的技术方法引入系统，用以改进和优化检索效果。

表5　网络资源数据与文献数据处理中部分因素比较

网络资源因素	文献资源因素	说明
HTML标记	文献结构数据	
URL	发表来源、地址	文献库人工标引
链接	引用数据（部分数据库）	
锚定文本		新相关因素
相关数据，如主题指南中类目数据	元数据中的分类、主题标识	文献库人工标引
检索日志	检索记录	

4　结语

后组式检索系统的后组、自由组配、隐含等特点，决定了对其组织体系的研究不能照搬先组式系统的方法，而应当结合其资源对象和应用环境，对其基本构成成分及要素加以考察。显然，将上述内容单纯作为检索问题进行探讨是不合适的，必然会严重限制对后组式系统组织规律和方法的探索，应当改变。后组式检索系统，尤其是关键词搜索引擎、文本检索系统等多元结合的特点决定了它们的发展需要多个领域专业人员的共同努力去加以推进。传统文献组织领域的专业人员不能局限于传统的检索语言领域，而应当结合长期研究和实践的积累对研究范围进行拓展。对后组式系统的研究可以适应时代发展，扩展视野，将我们的知识贡献于这一新的领域的开掘，同时也可以在对后组式系统充分了解的基础上，在更加广阔视野的基础上从事信息组织领域的研究和探索。

[参考文献]

[1] 马张华，黄智生．网络信息资源组织［M］．北京：北京大学出版社，2007：92－102．

[2] Brin S，Page L．The Anatomy of a Large-Scale Hypertextual Web Search Engine［J］．Computer Networks and ISDN Systems，1998，30（1－7）：107－117．

[3] Arasu A．Searching the Web［EB/OL］．［2007－08－13］．http://oak.cs.ucla.edu/~cho/papers/cho-toit01.pdf．

[4] Page，etal．The PageRank citation ranking：Bringing order to the Web［EB/OL］．［2007－08－13］．http://www-db.stanford.edu/~backrub/pageranksub.ps．

[5] 王建勇，等．海量web搜索引擎系统中用户行为的分布特征及其启示［J］．中国科学E辑，2001，31（4）：372－384．

[作者简介] 马张华（1948－），男，教授，出版有《网络信息资源组织》《信息组织》等著作。

[收稿日期] 2007－11－30　　[责任编辑] 王　岗

（上接第22页）多家图书馆加入了该链接，国家图书馆则是国内首家。读者通过Google Scholar进行信息检索后，可方便地检索国家图书馆数字资源门户，进一步获取数字资源全文，获取国家图书馆馆藏信息，获取文献传递、馆际互借服务。[1] 北京师范大学图书馆也成功地将Google Scholar纳入了图书馆学术资源服务体系，与万方数据资源系统、维普资讯等国内外著名数据库商的合作也日趋紧密。合作发展的最终目标是共赢，将有越来越多的图书馆和数据库供应商加入Google Scholar发展计划。[2] Google Scholar不仅推动了中文学术资源走向世界，而且方便了中国用户获取与利用国外科研学术成果，同时也为自身发展赢得了机会。

4．2　中国引文数据库将加快数据库建设步伐

中国引文数据库是中国知网的增值服务产品，有较为固定的信息检索用户（图书馆用户和一些网络散户），数据库检索技术成熟可靠，可利用这些优势，扩充信息源，如图书和报纸等类型资源，收录一些尚未收录的学术期刊，加快全文回溯建库速度等。同时，为了长远发展目标，加入了《中国学术期刊文献评价统计分析系统》的合作计划，成为该系统的基础数据源之一，为各学科期刊之间的比较与评价提供了准确、客观、公正的数据参考，促进了我国期刊发展和科学文献计量研究，并满足了国内学术研究人员的多层次需求，将在竞争中立于不败之地。

4．3　学术研究工具助推学术资源向开放存取方向发展

尽管目前Google Scholar与中国引文数据库都不能向用户提供免费学术资源，但都能提供免费检索，就说明它们已经掌握了数字化的资料，至于什么时候把它变为“开放存取”，那仅仅是时间，或者说是商业角度和版权制度的问题了。[3]

[参考文献]

[1] 夏旭．基于Google学术搜索的引文检索研究［J］．情报理论与实践，2006（6）：697－701．

[2] 张文彦．Google给图书馆带来的十大机遇与挑战［J］．图书馆杂志，2005（10）：62－63．

[3] 奇迹文库．也说Google的学术搜索［EB/OL］．［2007－05－02］．http://www.qiji.cn/drupal/node/7538．

[作者简介] 冯向春（1977－），女，广东茂名学院图书馆馆员，发表论文10篇。

[收稿日期] 2007－10－27　　[责任编辑] 王　岗