情报学报 ISSN 1000-0135
第25卷 351-355, 2006年10月

JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION ISSN 1000-0135
Vol 25 351-355 October 2006

从IPO分析未来的搜索引擎

化柏林
(中国科学技术信息研究所,北京 100038)

摘要 本文主要从搜索引擎的爬行范围、对网页内容的分析处理以及用户查询接口三个方面分析了搜索引擎的最新进展,并根据技术发展的规律以及人机交互的需求对搜索引擎的信息采集、信息分析、信息提供三大处理过程和支撑资源的建设等方面的发展作了相应的分析与观测。

关键词 搜索引擎 IPO 发展趋势 信息抽取 知识获取 自然语言处理 多媒体

1 引 言

一代搜索引擎以主题分类为主要特征,由于是人工采编,所以搜索范围较窄,准确度较高。这类搜索引擎一般都索引少于100万个网页,极少重新搜集网页并去刷新索引[1]。而且其检索速度非常慢,一般都要等待10秒,甚至更长的时间[1]。在实现技术上也基本沿用较为成熟的信息检索、网络、数据库等技术,相当于利用一些已有技术实现的一个WWW上的应用[1]。以Yahoo、搜狐等为代表。现在的使用也越来越少。

目前的搜索引擎以网页自动爬行、网页全文标引为技术特征,在自动爬行过程主要利用超链接进行爬行,在标引时主要用到词语的切分技术。从功能上同样分为三大部分:网页爬行、分析标引和用户查询[2]。从这样实现的过程来讲,北大天网等基于FTP的搜索也属于这一代,因为它也同样包括这样三大模块,只不过爬行走的不是HTTP协议,而是FTP协议;爬行的不是Web服务器,而是FTP服务器;不是超链接技术,而是IP寻址技术[3]。其实网页搜索引擎中也有不采用超链接来获取URL列表的,也有采用IP方法的。

以Google、百度等为代表的第二代搜索引擎技术上比较成熟,尽管也在推出一些新的搜索功能,但离人们的所得即所搜还是有一定差距的。下面从信息采集、信息分析处理和信息提供三个处理过程以及分析处理中所需要的资源来论述搜索引擎未来的发展,如图1所示。

从图中可以看出,上维属于信息搜集,是信源(Input);下维属于检索过程,讲的是信宿(Output);左维和右维属于分析标引过程,中间是处理过程(Process)。右维注重分析处理的层面,左维是资源支撑。从内到外数据越来越全,准确度越来越高,人机界面越来越友好。其实,IPO算不上什么理论,可申农的信息论、系统动力学、软件模块设计等理论和应用中又无一不体现着IPO的重要性。

2 信源 Input

二代搜索引擎的URL是直接从html文件中析取出来的,是字符级匹配的过程。搜索引擎只能对html文本中提供的URL进行下一页的爬行,而不能对动态生成的URL进行爬行。二代搜索引擎搜索的主要是静态URL,尽管能对形如*.asp的网页进行爬行,但对真正的动态网页搜索能力很差。特别是对通过URL传递用户输入参数的网页几乎没有能力,目前的Google和百度都不支持以数据记录为内容的网页。

三代搜索引擎能够爬行以数据记录为内容的网页。对于不同页面间的参数传递常用的有两种方法,第一种是直接写在URL中,这种方式显性于URL中,自动分析判别较容易。另外一种方式是通过对话Session设定参数,如JSP里的setParameter()和getParameter(),这种方式独立于URL链接,所以识别起来较困难。不同的网站设计者构造URL的方式不一样,所以搜索引擎无法用一种固定的模式来进行。这样的网页都是根据用户输入查询条件,以数据记录的形式从数据库里取出来,生成网页的,因此它的数据量更大、更专业、更新速度快、价值也高。表单元素[4]的识别较容易,元素值相对困难一些,要让计算机能尝试性输入值。现在有一些专门的下载工具可以对指定的网站进行数据记录的下载。三代搜索引擎能够参照html文件中析取出的URL构造新的URL并下载。这种URL的构造具有尝试性,能够学习。三代搜索引擎应该能够搜索网上公开的、免费的、非注册的动态网页。

四代搜索引擎在爬行过程中还多了一个自动注册机。网上有许多数据库是免费的,但是只有注册用户才能够使用,搜索引擎应该能够根据注册需求自动注册,完成注册过程成为系统用户,然后像三代搜索引擎一样再去下载数据库里的内容。这样爬行范围就更加广泛,获取数据机制与以往有很大改进。本来由人来完成的过程,可由计算机来完成。

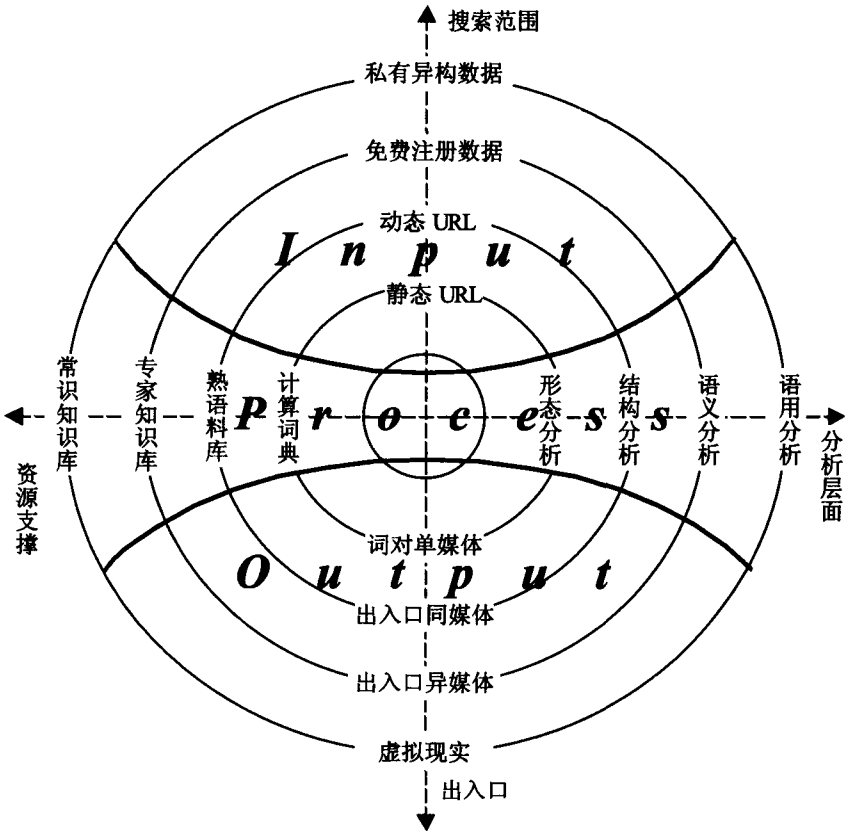

图1 搜索引擎发展趋势图

五代搜索引擎能够对私有数据进行搜索，当然异构数据的查询也早就实现了。这种异构是真正的开放的异构。现在的跨库搜索大都是针对某些特定的已知的数据库和数据库管理系统。不同的数据库可能使用不同的字段名，字段的长度与取值范围也会不同。不同的数据库管理系统的数据类型可能不完全对应，所支持的数据长度也不尽相同，查询语句或存储过程的语法也不尽相同。这一切都是目前的跨库搜索要解决的问题，不仅仅是ODBC等数据接口所能解决的问题。对于未知的新来的数据库，能够自动把数据字典导进去，然后把数据记录移植过来就直接能访问，不需要再改数据接口和重编译源程序。五代搜索引擎要解决的不仅是跨库搜索的技术问题，问题的关键在于数据库商的合作，这更多的是经济与社会问题。如搜索引擎与数字图书馆、各种数据库商的合作会使搜索引擎的搜索范围，数据质量有极大的提升。

从搜索范围来讲，三代搜索引擎与二代搜索引擎的爬行程序会有很大的变化，同一个网址会因为URL参数不一样而爬行多次。四代搜索引擎从技术上并没有多大改进，不能称之为代，但从数据范围上还是有很明显的变化，由公开直接获取到注册用户有限使用，反映了在数据获取机制上的巨大变化。五代搜索引擎主要涉及了数据合作、经济社会因素更多一些，为实现全球信息库而迈出的关键一步。

3 分析处理 Process

从语言的单位来看，文本分析主要有词法分析、句法分析、语段分析、篇章分析等，反映的是不同的信息粒度；而从分析的层面来看，又分为形态分析、语法分析、语义分析、语用分析等，反映的是不同的分析深度。未来的搜索引擎将跨越形式的匹配，深入到内容的匹配，主要涉及分析、理解与生成三大过程。分析的目标是为了理解，在理解的基础上才能生成。不论何种媒体格式，搜索引擎将沿着形态、结构、语义、语用的层面向前发展。

二代搜索引擎主要利用分词技术，词根词干分析技术，词语同现及频率分布。对于屈折语，词法分析主要是词的构成，通常有前缀+词根+后缀+词尾，切分非常容易，一般是空格自然分开，还有少量的标点符号进行分隔标记。对于分析型语言，切分便是最大的问题，拿中文来讲，一般有三类切分方法，自然切分、向量切分和概率切分。自然切分又分为一元、二元或三元等。它的优点是算法简单，不需要任何词典，纯机械切分，缺点是切分结果冗余较多，检索速度随着检索表达式的增长而变慢。这种切分方法适用于不是基于数据库而是基于文件的一些小型系统，网络上许多支持全文检索的小词典多使用这种切分方法。向量切分法按长度分为最大向量与最小向量，按分向又分为正向向量切分、逆向向量切分和双向向量切分，组合起来就有6种向量切分方法。

向量切分法相对二元切分或三元切分来讲，准确率要高，索引量要比自然切分小得多，如使用最大向量切分，就不会出现输入“北大”却查出“东北大学”这种情况。向量切分法应用较为广泛，目前许多系统都使用这种方法。如 CNKI、NSTL 等，这也是为什么有时候输入的检索词长了结果反而多了。如在 CNKI 的期刊论文数据库里通过标题检索“网络信息计量学”比检索“网络信息计量”的结果要多很多。概率切分是在有切分歧义时，通过已知词的概率来确定哪一种切分结果更优。

三代搜索引擎不再停留在词的层面，而是深入到句法层面，对句子的结构、句子成分及词汇短语在句子中的功能进行分析；对于图形图像涉及到颜色、纹理、形状的分析；对于音频涉及基音、音强、音色，对于视频涉及到帧结构、镜头运动方式与切换方式等[5]。这一时期统计型搜索、学习型搜索将会得到长足发展，自动问答系统也会有所进展。信息抽取技术将会在这代搜索引擎中得到广泛应用。查到了相应的文献，在文档中有许多我们不需要的信息，通过信息抽取把想要的信息单元抽取出来，过滤掉其他信息。比如我们想查中国所有自然语言处理方向的博士生导师，我们现在的做法是，搜索引擎返回的是一个个招生单位的链接，然后我们遍历每一个网页，人工地进行汇总。而我们需要博士生导师的姓名、所在单位、研究方向等信息。利用信息抽取技术就会直接出来一个二维列表，也就是只需要阅读一个网页，所以也称列表式搜索。学习型搜索相当于文献自动综述。例如想了解搜索引擎，将不再显示所有含有搜索引擎的成千上万篇文章，而是一篇综合了所有关于搜索引擎的文章，文章会有发展历史，主要分类，使用技巧与方法，关键技术实现，发展趋势等多个主题，相当于百科全书的形式来组织关于搜索引擎的所有知识点。这样由阅读多篇文章变成了阅读一篇文章的不同部分，实现了内容的滤重与重组。

四代搜索引擎将深入语义层面，深入理解句子的意思，理解图像的含义、音频视频的内容，这时对于不同媒体格式的数据可以达到统一。不仅要分析词的义项、分析句子的语义，应该还能够对语篇进行语义分析。这个时期可以进行观点型搜索、流派型搜索，如查持有“数据挖掘不同于知识发现，而是知识发现的一个阶段”观点的文章或与其观点不同的文章。因为越过了符号系统，深入到语义层面，所以跨语言检索也将有长足的发展，“Love lives in cottages as well as in courts.”对于这样一个句子，“爱情不分贫富。”是比较地道的译文，但我们在目前的搜索引擎中输入“贫富”可能很难能查到含有上述英文句子的网页。可它毕竟表达出了贫富的意思，因此真正的跨语言检索是需要同族匹配、提问翻译、文档翻译和中间语言转换等技术[6]。多种媒体格式的数据用统一的语义来表示，语义的表示仍然是个难题，如“竹横麻竖，青黄交错软硬帘；碳去盐归，黑白分明山水货”分别描述的是两幅劳动的场景，除了有颜色、纹理等图像特征外，还有质地、取源等图像难以分析的内容特征，而把场景和上述对联用统一的语义来表示的确有点困难。

五代搜索引擎将穿越语义，在充分理解各种语义的基础上，能够分析文献的写作手法、修辞方式，能够推敲语言的妙用。搜索引擎能够分析出不同媒体格式所带来的不同效果。强调语用是五代搜索引擎的主要特征，如想查询所有与“孔乙己大约的确死了”使用同一写作手法的句子，或者查询一二句描写自然景色三四句抨击社会现象的七言律诗。这样的搜索就穿越了语义而达到了语用的层面，不仅仅是语义搜索，而是语用搜索。

4 信宿 Output

从检索出入口来看，二代搜索引擎输入的是文本，输出的是文本、图像、音频、视频。对于非文本的搜索主要是输入描述性的词语，而这些词是从文件名中或文件说明中抽取出来的词，所以从本质上讲，去数据库里还是用文本来匹配文本的搜索，只不过返回的结果是图片或音频视频而已。二代搜索引擎的入口是词，不同的词之间通过逻辑组配，加之检索范围的限定去数据库进行匹配。现在查数据挖掘与知识发现的区别，往往要输入数据挖掘 AND 知识发现 AND(区别 OR 差别 OR 不同 OR 异同 OR 比较)，从检索入口角度，构造一个表达式很麻烦，一般的检索用户没法穷举表达区别的各种方法；从检索出口角度，即使目标文献符合上述检索条件，也不一定真的是在讲数据挖掘与知识发现的区别，这种形式匹配而不是内容匹配显然不能满足人们的检索需求。

三代搜索引擎可以实现出入口同媒体，通过输入自然语言的句子来进行文本的搜索，而对于图形，可以输入示例图形，也可以草图查询，如输入灭火器的示例图来查询灭火器，或输入日落的草图来查日落的图片。所以输入和输出是同一种媒体。如 IBM 的 QBIC 就是使用草图检索图片的一个系统。想查日落图片，先画一个圆，圆的上半部涂成明亮色，下半部涂成暗灰色，就能查到相关图片，可是这样的方式又如何能确定是日落而不是日出呢？对于视频，从某一系列的视频中要检索某一符合条件的镜头，比如从四十集的《射雕英雄传》中检索郭靖使用降龙十八掌的所有镜头，我们可以找一个或几个典型的关于降龙十八掌的视频帧作为输入，检索出所有郭靖使用降龙十八掌的所有镜头，可以进行聚类分析或其他研究等。三代搜索引擎只能是对已有的媒体格式进行分析，并不能按需求生成其他类型的媒体格式。

四代搜索引擎可以实现输入与输出是不同的媒体，如果没有相应的媒体数据，可以由系统生成。输入一个句子，结果可以是文本，也可以是图像音频视频等。如输入“朱镕基新任上海市长的就职演说”，不但能查到朱镕基那本来 15 分钟实际却长达 110 分钟的演讲稿，还能查到 110 分钟的音频或视频。如果网上事先没有这种音频数据，可以根据文本和朱镕基讲话的特点，生成音频，也就是 TTS(Text To Speech)。新闻媒体库里有当时的朱镕基一些讲话录音，从这些录音中分析提取朱镕基讲话的特点，然后把文本转成音频。当然能够根据所讲内容不同而语气也就不同，就到了语用的层面。

将来的语义模型不仅适用于文本，还应该具有多媒体互通互用，一个句子、一副图像、一句音频，如果是在描述同一个内容，我们应该可以用统一的一种语义来描述不同媒体形式的内容，不同的媒体格式只是交互的方式不一样，对人的输入输出路径不同，接受的信息是相同的，只不过感觉不一样而已。如视频效果可能更直观、文本阅读选择更自由等。只有达到这个层面，媒体格式才能真正的实现互连互通、互转互换、互呈互现。

五代搜索引擎不仅可以生成相应的音频和视频，还能够准确地配以空间属性，以地理属性的可以进行全球定位，如果我们在搜索引擎中输入萨达姆是如何被抓获的，将生成一条文本，描述萨达姆在何时何地如何被发现并被捕，什么样的人参与抓获过程，用了多长时间等一段完整的描述。这也就是正在研究的QA自动问答系统。如果网络上有相应的视频，也会根据视频内容进行分析匹配。如果没有这样的视频可以用搜索到的图片和文本生成动画演示。并把事件的发生通过GPS定位，在GIS上显示出来。把物理属性与地理属性结合起来，并能对历史数据进行回溯，对未来趋势进行预测。基于三维空间坐标的地理信息系统体现虚拟现实的效果会更好。实现真正的数字地球，并能对数字地球任意的搜索，信息可视化程度非常高，信息的自动合成与生成，并能以最接近现实的方式提供信息，更关注人的感觉，是五代搜索引擎显著的特点。

5 资源支撑

从资源的支撑来看，现在的搜索引擎主要是计算词典。对于任何处理，分词是必需的，词性标记用的少一些，还有一些是要考虑词频的。有时词性标记亦可用于一些复杂的切分，特别是有切分歧义的那种可以通过词性标记来消歧。如果仅仅是分词，一般的词典就够用了。可是如果想进行自动摘要，自动分类，聚类分析等就不够了。对于这样的应用，进行句法分析是必要的。所以不仅要有词典，还要有规则。语言知识库越丰富，分析处理会越准确。

目前的智能信息系统都是离不开专家知识的。专家系统主要依据于专家知识库，用一个好的搜索策略找到所需知识，然后用合理的推理机进行推理，做出判断并进行决策。如网上肿瘤诊断系统需要病症和病例库，以及针对不同病情的诊疗方案库。动物识别系统要有各种动物的特征库。如果没有众多著名象棋大师的棋谱，IBM的深蓝又如何能战胜国际象棋大师？所以知识库的建设是相当重要的。网络搜索引擎也在着重某一个方面的深化，想做出一点特色，专家知识库的构建也势在必行。试想新浪的爱问和百度的知道，除了能满足现在一些问题需求外，更重要的是积攒了许多问题及答案。一旦这些问题库与答案库足够大，再加上自然语言处理的成熟，就可以对现有的问题进行分析理解，并把不同问题进行组合与推理，让计算机回答一些新问题也不是没有可能。除了这些问题库，如果能把网页里有的知识和论文里的知识都提取出来，以计算机可理解的形式存到数据库里，形成包罗万象的知识库，那么计算机所拥有的知识岂只是一个专家或一批专家所能比拟的。这也就是真正的知识工程。在知识工程的研究中，知识表示有很多种方法，知识利用也有一些机制，难就难在知识的获取，这也是人工智能向前发展的瓶颈问题。由知识工程师和专家来人工获取毕竟太慢了。绝大多数知识其实都已经表达出来，都蕴含在文献里，如何能从文献里把相应的知识点提取出来，以合理的方式来表示存到知识库里，成了众多研究的基础。有了计算机可识别的知识库，计算机能做的事情就更多，计算机的分析处理能力也会有重大突破。所以在知识获取没有重大突破前，只能有限的获取某一领域的专家知识，让计算机拥有常识知识库还有很长的路要走。

拥有技术并不一定能领航，拥有资源才是真正的“大腕”。像清华同方、万方数据、报社等拥有如此丰富的文本数据，而像中央电视台等拥有如此丰富的视频资料，对这些资料进行深加工，必然会形成最具竞争力的财富。像北大计算语言所等拥有的语言学知识库可以满足这种深加工的需要，这方面的合作迫在眉睫，如自1998年开始加工的《人民日报》的熟语料库，就深受欢迎。另外一个方面就是权威数据库，一些政府机构适当开放部分数据也是非常有必要的。如公安部为社会提供人名数据库，我们可以统计姓氏的数量分布，名字用字的频率分布等，这对我们在信息抽取过程中的人名识别会有相当大的帮助。如果我们拥有国土资源部的地理数据，那我们的地名识别也会有很大的提高，同理如果拥有工商部门的数据，那么公司名的识别也不成问题。因此，为了资源共享，为了提高我们的民族搜索事业，国家行政部门也应该适当开放一些这方面的数据，这不会涉及社会安全稳定及个人隐私等问题。国家统计局和其他行政部门应该开放更多的一些基础数据或统计数据。

6 总 结

除了上述讨论的搜索引擎的发展方向以外，还有其他很多方面都在不断地改进、升级与创新，用户个性化的问题也是一大热点。人机交互是根本，信息分析是关键，只有信息分析技术得到了充足的发展，人机交互才能得以改进。

图灵测试作为测试计算机智能的测试被许多人工智能专家所接受，有人提出不同的人智商不一样，知识背景也不一样，所以图灵测试的可信度也不一样。其实，恰恰是这种不稳定性才能反映出人工智能的特性。正如同一个骗子使用同样的骗术有人会上当而有人就不会。未来的搜索引擎会通过图灵测试，然后超越图灵测试。因为在经过大量的QA以后，你会发现它不是人，而是计算机，因为一个人不会有这么广的知识面，它是一个拥有庞大知识库和推理机的智能机器，你所面对的好像是各行各业的专家们会聚在一起来回答你的问题。

代的划分不是绝对的，在实际发展过程中，不同的维也

可能有不同的发展速度。未来的搜索引擎将是搜索技术与数据库商充分联合，政府数据适当开放，知识库共建共享，全文数据、多媒体数据与空间数据紧密结合，充分利用各种人机交互方式，结果精确、个性突出、充分智能的搜索。

参 考 文 献

1 雷鸣，王建勇，赵江华，单松巍，陈葆珏．第三代搜索引擎与天网二期．北京大学学报(自然科学版)，2001，20(5)

2 化柏林．Google 搜索引擎技术实现探究．现代图书情报技术，2004(年刊)

3 李晓明，闫宏飞，王继民著．搜索引擎－原理、技术与系统．北京：科学出版社，2005

4 严亚兰．面向动态网页爬行的 Crawler 架构．图书情报知识，2003(4)

5 李国辉，汤大权，武德峰编著．信息组织与检索．北京：科学出版社，2003

6 黄国才．汉英双向跨语言元搜索引擎 CELanSE 的设计与实现[学位论文]．中国科学技术信息研究所，2001