信息检索与信息抽取差异性探析

郑彦宁 化柏林 张新民
中国科学技术信息研究所 北京 100038

〔摘要〕通过发表论文、会议组织、出入口、关键技术、发展趋势等方面对信息检索与信息抽取进行比较分析，发现信息抽取与信息检索有着质的不同。信息抽取不是信息检索的发展方向，但信息抽取技术可以很好地应用于信息检索系统。分析两者之间的差异有利于研究的深入，理清它们的关系有利于共同促进。

〔关键词〕信息检索 信息抽取 命名实体识别 模式匹配 规则抽取
〔分类号〕G35 TP391

Study on Differences between Information Retrieval and Information Extraction

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

〔**Abstract**〕 Through comparing information retrieval with information extraction in terms of published paper, conferences, access and output points, key technique and development trend, the paper finds that there is radical difference between information extraction and information retrieval and that information extraction can not stand for the direction of information retrieval, but this type of technigue can be applied into information retrieval system. Analyzing differences between them will foster further research, and exploring their relation will be beneficial to their mutual development.

〔**Keywords**〕 information retrieval information extraction named entity recognizing pattern matching rules extraction

李保利[1]等人从功能、处理技术与适用领域等三个方面介绍了信息检索与信息抽取的不同[1]。李芳等人在阅读大量相关文献的基础上，简要介绍了信息抽取、信息检索与自动文摘的区别[2]，认为信息抽取是“更高级的信息检索”。文献[3]分析了信息检索与信息抽取的特点和不足，认为信息检索技术的研究主要侧重于语料库的方法，信息抽取技术的研究更侧重于自然语言的理解，基于符号的处理方法，并最终提出了一个结合两者优势的信息获取模型。

然而笔者认为，信息抽取不是信息检索的高级阶段，它并不能代表信息检索的发展方向。信息抽取可以应用于信息检索，提高检索质量与精度，反之，信息检索的应用也会对信息抽取提出更新的挑战。

1 信息检索与信息抽取的学术关注度差异

在中国知网上检索相关文献（题名或关键词精确匹配），关于信息抽取的第一篇文章为1997年刊登在《情报学报》上的《基于信息抽取和文本生成的自动文摘系统设计》；关于信息检索的第一篇文章为1980年刊登在《情报科学》上的《全息情报检索QQJ系统简介》；关于文献检索的第一篇文章为1976年刊登在《武汉大学学报（理学版）》的《怎样查找科技文献资料》。从1997至2006十年间关于信息抽取的文章共393篇，年均39篇，关于信息检索的文章达到6 269篇，年均627篇，是信息抽取的16倍。近10年来信息检索与信息抽取的文章发表数量如表1所示：

表1 中国知网全文数据库跨库专业检索结果

年份	信息抽取(篇)	信息检索(篇)	信息抽取增长率(%)	信息检索增长率(%)	年份	信息抽取(篇)	信息检索(篇)	信息抽取增长率(%)	信息检索增长率(%)
1997	1	216	-	-	2003	34	780	48	16
1998	3	256	200	19	2004	73	945	115	21
1999	7	324	133	27	2005	104	934	42	-1
2000	8	450	14	39	2006	120	1 104	15	18
2001	20	589	150	31	合计	393	6 269	-	-
2002	23	671	15	14					

（注：检索条件为题名或关键词精确匹配，检索时间：2007-05-27）

从绝对数量上看，信息检索的文章远远多于信息抽取，甚至高出一个数量级。信息抽取的研究起步比较晚，只有10年的时间，而信息检索的研究比较成熟，已有几十年的时间。关于信息抽取的文章，增长最多的时候出现在2004与2005年，分别增长29篇与31篇。关于信息检索的文章，每年的增量都在100篇左右，只有2005年出现了很小的负增长，而增长率

收稿日期:2007-06-18 修回日期:2007-06-29 本文起止页码:17-20 本文责任编辑:林佳莉

最高的时候出现在2000年前后，从1999年到2001三年间保持着30%左右的增长。为了能在同一幅图里显示信息抽取与信息检索文章的增长趋势，把信息检索的文章数量进行缩小，缩小到与信息抽取的文章处于同一数量级（每年文章数量除以16），如图1所示：

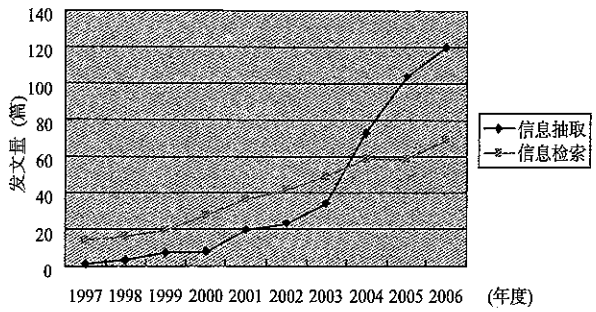

图1　信息抽取与信息检索发文量趋势（1：16）

从图1中可以看出，信息检索得到了持续的关注，从1998年开始迅猛增长，增长的原因主要是搜索引擎的崛起，带动了整个信息检索领域的新发展。而信息抽取从20世纪90年代末开始得到关注，从2003年开始得到迅速发展。目前信息抽取的增长势头非常迅猛，而信息检索相对平稳一些。如果说10年间信息抽取的研究经历了从无到有的过程，那么信息检索的研究就是从弱到强的过程。

2　信息检索与信息抽取的相关会议

关于信息检索的国内会议比较多，其中包括：中国科技情报学会计算机情报检索专业委员会从1980至1986年举办了5届全国机器检索学会交流会，该系列会议后来改名为全国计算机情报检索学术讨论会，后来再次改名为全国计算机信息管理学术讨论会；中国中文信息学会信息检索与内容安全专业委员会举办的全国信息检索与内容安全学术会议，今年将举办第三届会议；随着搜索引擎的迅速发展，关于搜索引擎的会议也显得越来越重要，中国计算机学会互联网专业委员会举办的全国搜索引擎和网上信息挖掘学术研讨会，2007年已举办了第五届；另外，微软亚洲研究院联合清华大学、香港中文大学于2004年共同承办了首届亚洲信息检索研讨会。

关于信息检索的最有影响力的两个国际会议组织当属TREC与INEX。TREC由国际标准和技术委员会及美国国防部共同资助，每届参会的人数很多，提供丰富的评测标准与实验数据，是目前最权威的检索评价会议；INEX由DELOS数字图书馆网络组织和IEEE计算机学会资助，主要针对基于内容的XML检索提供统一评价程序。这两大会议是国际上公认的权威评测机构，而国内关于检索方面的评价还很少，全国搜索引擎和网上信息挖掘学术研讨会近年来主要是针对分类进行评测。

信息检索会议举办得如火如荼，但以信息抽取命名的会议在国内还很少，比较有影响力的是微软亚洲研究院于2005年举办的信息抽取技术暑期研讨班。国际上比较有影响力的当属MUC，它是20世纪80年代末由美国国防部的DARPA发起的，旨在通过一系列国际化的研究系统测评，来推动信息抽取的研究，提高信息抽取的能力，目前已举办了7届会议。

信息检索不仅有大量的学术论文与会议组织，还有成熟的理论模型与经典著作，而信息抽取的理论模型尚不成熟，也尚未出现经典著作。信息检索领域最经典的著作当属Ricardo Baeza-Yates, Berthier Ribeiro-Neto等人著的《现代信息检索》(Modern Information Retrieval)。信息检索的理论模型主要有概率模型、布尔模型、向量模型和逻辑模型[4]。

3　信息抽取与信息检索的出入口

信息检索强调对检索入口进行控制，并不对检索出口进行控制，也就是说，信息检索策略的调整只能决定检索结果的多与少，并不能决定每条检索结果的大与小。通过构造检索表达式与指定检索范围等策略来决定检索结果的记录数，而不能对某条记录的内容进行抽取。例如，要查找中国所有自然语言处理方向的博士生导师，利用搜索引擎进行检索，用户需要遍历每一个网页，然后进行人工汇总。如果将信息抽取技术应用于搜索引擎，在检索之前可以指定内容的范围，也就是说会有两个检索输入框，第一个为检索入口，每二个为检索出口，检索入口输入“自然语言处理 方向 博士生导师”，检索出口输入“姓名、所在单位、专业、年龄、招生人数、考试科目”等信息，利用信息抽取技术就会直接显示出一个二维列表，用户只需阅读一个网页，这种搜索也称之为列表式搜索。

信息抽取不同于信息检索，其粒度要比信息检索的粒度小——信息检索以篇为单位，信息抽取以篇中的信息单元为处理单位。信息检索一般返回整篇文献，而信息抽取返回信息的某个单元；信息抽取存在对与错的问题，如抽取的名词要么是人名，要么不是人名，不存在人名的贴近度问题。而信息检索存在好与坏的问题，是一个程度问题，我们称之为召回率，信息检索所查到的文献，有完全符合需求的，有基本符合需求的，有不怎么符合需求的，所有返回文献的准确率是线性的、连续的。

信息检索的最终用户是人，而信息抽取的用户是计算机。一般来讲，信息检索由人构造检索式，通过系统进行检索，得到检索结果由人来查看，整个过程中体现着人机交互；而信息抽取一般是系统根据模板和预先设定的规则，通过分析文本抽取需要的内容，信息抽取系统一般不单独使用，往往是为其它系统提供技术工具，例如为信息检索、自动分类、自动问答等应用系统解决某些特定的问题，信息抽取过程往往不需要人机交互。

信息抽取按抽取的数据对象结构化程度分为三类：①以PDF文件代表的非结构化文件，利用文件结构、字体、换行符等方面进行分析并抽取[5]，PDF文件只有文件结构信息，没有任何关于内容的信息，而目前全文数据库大都以PDF为存储格式，因此非结构化文件的信息抽取意义重大，难度也很大；②以网页文件为代表的半结构化文件，即以标记语言为格式的文件，按照标记程度分为HTML和XML[6]。基于XML文件对象的信息抽取主要使用DTD[6]以及DOM树附加语义、样本学习生成基于DOM路径的抽取规则，利用遍历DOM树实现信息抽取[7]。标记信息有两种：一种是HTML标签标记，如<title><title>，一种是文本标记，如“相关链接”文本所指示的信息为URL链接信息；③以数据库内容为代表的结构化信息，抽取相对简单，关于这方面的探讨还比较少。

4 信息检索与信息抽取的关键技术

信息检索通常有分析标引与响应检索两大过程，信息抽取的分析过程更复杂、更有针对性。信息检索可以做成通用的，而信息抽取往往是领域相关的或特征相关的。

一般的信息抽取系统包含以下6步过程[8]：

- 用一组信息模式描述感兴趣的信息。
- 对文本进行预处理。采用特征词频率统计和特定模式匹配过滤掉当前文本中与特定领域无关的信息。
- 对文本进行词法分析、浅层句法分析以及简单的语义分析，对文本中包含的特定领域的主要名词短语单元进行识别，同时标注语义信息。
- 使用模式匹配方法实现事件模板的构造，建立实体之间的联系。采用基于知识的句子分析技术，将识别的实体映射到一个结构中，并标注它们的角色。
- 采用语段分析技术实现句子相关性分析，进行上下文关联、共指、引用等分析和推理，对句子层获得的结构实现重载与合并，解决语段的指代和省略问题，构造一个完整的实体事件。
- 格式化分析结果，把抽取的信息输出到预定义好的模板中[8]。

信息抽取的关键是命名实体识别与模板的匹配。命名实体识别有两类特征信息，一类是实体内含信息，如姓名抽取中，根据中国人数较多的姓（如王、张、李、刘等）以及专用于姓的汉字(如姚、闫等)，加之人名所用高频字等信息判断姓名；另一类是前后附着信息，根据实体名的上下文来识别命名实体，如根据机构、职称、职务、职业、称谓等关系确定命名实体（如北京市委书记刘淇、北京大学副教授孔庆东等），一般都是紧密相连。如果针对特定的抽取任务，设计~名高频词、~名低频词、~名停用词等亦可以提高处理的精度。

一个模板就是一条规则，每个模板都是一个约束的序列，这个约束的序列表现为对文本特征的描述，这些特征包括标点符号、词典、大小写、词长、句法分类、句法分块、语义特征等[9]。而事件抽取不一定是整篇文献的内容，有可能只从文献的某一部分内容中进行抽取。例如从新闻中专门抽取事件的经过或事件的影响。事件的描述主要有事情的背景、人物、时间、地点、缘由、开始、过程、结果、影响、评价等，这种事件的抽取涉及场景模板填充任务、命名实体识别、共指关系确定、模板元素填充等。如袁毓林在职务变动事件抽取研究中，根据职务变更动词的有关句法、语义特点，把职务变动的动词分成6个小类，分别描写每一小类动词的论元结构，建立动词的论元角色跟事件模板元素的匹配关系，进行由动词驱动的信息抽取[10]。通过语句的逻辑结构和篇章结构约束信息模板的类型，并约束对当前句中缺失的或以代词等形式表达的信息项目的求解[11]。

5 信息检索与信息抽取的发展趋势

目前信息抽取的模型有很多，包括基于agent的信息抽取[12]、基于隐马尔科夫模型的信息抽取[13-14]、基于决策树的信息抽取[15]以及基于本体的信息抽取[16-17]。基于本体的信息抽取的研究比较多，因为一旦有了本体，信息抽取相对比较容易，所以基于本体的信息抽取不管是期刊论文还是学位论文都特别多①，但如何获取本体才是问题的关键。

现代信息检索的理论模型开始更多地糅合粗糙集、模糊集、潜在语义标引、神经网络等人工智能技术，信息检索的应用也朝着个性化、知识化、智能化的方向发展，垂直搜索引擎也取得了长足的发展，并起着举足轻重的作用。

未来的信息检索与信息抽取，将更多地运用人工智能理论与自然语言处理技术，需要更加丰富的语料库与语言学知识的支撑。只要资源库足够丰富，无论是抽取还是检索都会更加有效。公安系统有全国13亿人口的资料，可以统计出姓和名的用字概率；政府有全国各级行政区划的命名，铁路系统有大小火车站的名录，这些数据库准确度和可信度都非常高，需要增加系统数据的开放性。如果把各行各业的数据统一共享起来，命名实体的识别就会容易得多。再加之各种分类系统、各行业主题词条，概念等级体系也会在很大程度上得以解决。

6 结论

综上，笔者认为，信息检索与信息抽取有着质的不同，信息检索与信息抽取是完全不同的两个概念，无论从处理目标、

①注释：在中国期刊全文数据库通过标题精确检索信息抽取有128篇，“基于”“信息抽取”有64篇，(“本体”或“ontology”)并且“信息抽取”有9篇。

关键技术、理论模型还是应用对象等各方面都有所不同。信息抽取不是“更高级的信息检索”，它不是信息检索的发展方向，也不会取代信息检索，只能是促进信息检索的发展。

信息抽取可以应用于信息检索，但信息检索不是信息抽取的唯一应用。信息抽取除了可用于信息检索外，还可用于自动文摘、自动问答系统、技术跟踪与监测系统、结构化数据获取等很多方面。

在有关信息抽取的学术论文中，硕士生所发表的论文占很大比重，近几年有关这个方面的硕士学位论文也较多，特别是基于本体的信息检索或基于本体的信息抽取尤其明显，这种现象与其它研究领域有很大的不同。因为信息抽取往往是面对特定领域，针对某一具体特征，运用某种方法解决某种特定问题的研究，相对来讲比较容易取得应用性创新，而且有很强的现实意义，不需要太大、太复杂的工程即能实现，无需深奥的理论支撑，理解起来也较为容易，这些特点使得大家纷纷加入信息抽取的研究和探索行列。但是如果要分析更复杂的语言现象、设计更通用的信息抽取模式、抽取更复杂的信息单元，还有很长的路要走。

参考文献：

[1] 李保利,陈玉忠,俞士汶.信息抽取研究综述.计算机工程与应用,2003,39(10):1-5,66.
[2] 李　芳,盛焕烨,姚天昉.信息检索与信息抽取技术的研究.计算机应用研究,2002,22(2):16-18.
[3] 贺　胜.信息抽取技术内涵及应用.南京师范大学文学院学报,2004(2):184-188.
[4] Baeza-Yates R, Ribeiro-Neto B.现代信息检索(英文版).北京:机械工业出版社,2006:24-34.
[5] 李　珍,田学东.PDF文件信息的抽取与分析.计算机应用,2003,23(12):145-147.
[6] 宋艳娟,张文德.基于XML的PDF文档信息抽取系统的研究.现代图书情报技术,2005(9):10-13.
[7] 崔继馨,张　鹏,杨文柱.基于DOM的Web信息抽取.河北农业大学学报,2005,28(3):90-93.
[8] 孙　斌.信息提取技术概述(上).术语标准化与信息技术,2002(3):28-32.
[9] Leonid peshkin and avi pfeffer bayesian information extraction network.[2007-05-27]. http://dli.iiit.ac.in/ijcai/IJCAI-2003/PDF/063.pdf.
[10] 袁毓林.用动词的论元结构跟事件模板相匹配——一种由动词驱动的信息抽取方法.中文信息学报,2005,19(5):37-43.
[11] 袁毓林.用逻辑和篇章知识来约束模板匹配——逻辑结构和篇章结构知识在信息抽取中的运用.中文信息学报,2005,19(4):39-45.
[12] 孟宪福,狄　惠.基于Agent和XML的web页面信息抽取研究与设计.计算机工程与设计,2004,25(8):1411-1414.
[13] 王　胜,朱　明.基于最大熵马尔可夫模型的地址信息抽取.计算机工程与应用,2005,41(21):192-194.
[14] 刘云中,林亚平,陈治平.基于隐马尔可夫模型的文本信息抽取.系统仿真学报,2004,16(3):507-510.
[15] 张树瑜,朱仲英.基于MT决策树的Web信息抽取研究.计算机工程与应用,2004,40(13):69-71.
[16] 陆科进,李新颖.基于Ontology的文本信息抽取.计算机应用研究,2003(7):46-48.
[17] 张成洪,王向安,古晓洪.利用Ontology和规则表达式的Web信息抽取.计算机工程,2004,30(5):58-60.

〔作者简介〕 郑彦宁,男,1965年生,研究员,发表论文4篇;　化柏林,男,1977年生,硕士,助理研究员,发表论文5篇;　张新民,男,1970年生,博士,副研究员,发表论文43篇,译著1部。

(上接第43页)

描述,使潜在用户了解服务的价值并吸引他们利用这项服务,变潜在用户为实际用户。国外图书情报机构有宣传这项服务的优良传统，并且有许多成熟的经验可供借鉴。国内图书馆可通过多种途径宣传数字参考咨询的使用。

此外，关于指标体系的应用[6]，采用一定的标准评价某个数字参考咨询服务系统的质量或是对若干个虚拟参考咨询服务系统进行分析比较，都必须要考虑到数字参考咨询服务的具体环境，包括外部环境（社会、文化、政策、语言、用户等）和内部环境（目标、任务、基础设施、人力、财力等）。具体环境中的数字参考咨询服务评价还要以此指标体系为基础增删一些评价要素，设计具体的操作方法，把数字参考咨询服务指标体系方案做到实处，取长补短，以促进数字参考咨询稳定、可持续发展。

参考文献：

[1] Facets of quality for digital reference services. [2005-04-05]. http//www.Vrd.org/faces-06-03 .shtml.
[2] Charles R M, et al. Statistics,measures and quality standards for assessing digital reference library services guidelines and procedues. [2005-03-02]. http://quartzsyr edu/quality/Quality.pdf.
[3] 张　婉,杨志萍,周宁丽,等.国家科学数字图书馆网络联合参考咨询服务质量控制及评价方案研究.现代图书情报技术,2005(11):30-33.
[4] 初景利.图书馆数字参考咨询服务研究.北京:北京图书馆出版社,2004.
[5] 袁红军.国内外合作式数字参考咨询服务项目比较分析. 情报理论与实践,2007(1):76-79.
[6] 杨慕莲,詹德优.虚拟参考咨询服务评价标准与体系初探.情报理论与实践,2005(4):396-398,443.

〔作者简介〕 袁红军,男,1970年生,副研究馆员,发表论文94篇。