维普资讯 http://www.cqvip.com

信息抽取与信息检索技术比较研究

史树敏[1,2] 刘东升[1]

(内蒙古师范大学计算机与信息工程学院 呼和浩特 010022)[1](南京理工大学计算机科学技术学院 南京 210049)[2]

摘 要 面对日益激增的信息量,人们迫切希望能够拥有快速、便捷获取有用信息的技术或方法。信息检索及稍晚发展起来的信息抽取技术应运而生。本文旨在介绍并分析比较信息抽取与信息检索技术各自的发展历程、相关研究方法等重要问题,为笔者及相关研究人员今后研究提供一项基础性调研报告。

关键词 信息抽取,信息检索,比较研究

1 基本概念

本文从信息抽取与信息检索技术的基本概念入手,分别就各自的发展历程、适用领域、主要研究内容和未来值得关注的若干问题等,通过分析比较,展开论述。

信息抽取(Information Extraction, IE)是把原始文本里包含的事实信息(Factual Information)进行结构化处理,变成表格一样固定格式的信息点的组织形式。信息检索(Information Retrieval, IR)是指从非结构化的数据记录,特别是包含自由格式的自然语言文本的数据记录中获取与用户的信息需求相关的数据记录。

1950年,美国学者 Calvin N. Mooers 首创了"信息检索"这一术语。随后 Luhn 推出了统计 IR 的基本理论和方法;Marson 和 Kuhns 提出了 IR 的概率模型;Gerard Salton 等人提出了 IR 空间向量模型。1966年,在 Cranfield 项目中首次提出 IR 系统的评价方法。1972年,Lockheed 公司推出了世界首例商用在线信息查询服务系统:DIALOG 系统。20世纪80年代,IR 技术发展进入相对沉寂的时期,但也产生了一些典型的理论成果。90年代,IR 技术进入高发展期,出现了潜在语义索引技术、Bayes 网络和神经网络技术等理论成果;Google、Lycos 等大型互联网搜索引擎[1]。信息抽取技术则是近十几年来发展起来的,其产生、发展乃至学科的建立都与一系列重要的会议分不开,下节将详细介绍。

2 重要项目与评测会议

2.1 TIPSTER 项目与 MUC 系列会议

由美国国防部高级研究计划署(DARPA)和中央情报局(CIA)联合资助,并由国家标准与技术局(NIST)协作的 TIPSTER 文本项目(TIPSTER Text Program)是 IE 研究的主要发起和推动力量。该项目1991年实施,1998年中止,目的是通过政府、产业界和学术界的研究、开发者的合作来提高文本处理水平,并将其研究成果应用于包括 CIA 在内的13家情报部门。主要关注三项技术内容:文档侦测、信息抽取和自动文摘。在 TIPSTER 项目下,开展了 TREC、MUC 和 SUMMAC 等以互相交流、促进发展为目的的评测会议[2]。

消息理解系列会议(Message Understanding Conference, MUC)对 IE 领域建立及发展的推动作用不容忽视。美国政府支持下 MUC 于1987～1998年先后共举办了7届,对 IE 技术进行评测。MUC 实际上首开大规模自然语言处理系统评测的先河,在此之前,评价基本没有章法可循,测试也仅限于在训练集上进行;由 MUC 所定义的各种 IE 任务规范及评价体系已经成为其研究事实上的标准。随着应用需求推动和技术普及与整合,英语和日语姓名识别的成功率达到了人类专家的水平。一些系统已具备相当程度的处理大规模真实文本的能力。MUC 英文信息抽取的各项指标(最好情况)大体是:NE 任务约 97%,CO 任务约 60%～70%,TE、TR 任务约 60%～70%,ST 任务约 60%～70%[3]。关于 MUC 会议的详细内容也有很多文献进行专门介绍[4,5],这里不再赘述。

2.2 TIDES 项目与 ACE 系列会议

由 DARPA 资助正在进行的跨语言信息侦测、抽取和摘要(Translingual Information Detection, Extraction and Summarization, TIDES)项目是 TIPSTER 的后继,二者都被称作"评测驱动的计划"。项目研究:发现并翻译对美国国家安全至关重要的、不同语种的文本信息的自动技术及严格、客观的评测方法。阿拉伯语和汉语是项目关注的重点。所涉及的技术包括 IR、MT、IE、问题回答、文档理解和自动文摘等。项目外部组被邀请参加 NIST 每年举办的 IR 评测,主题侦测和跟踪(Topic Detection and Tracking, TDT)评测,ACE(Automatic Con-

维普资讯 http://www.cqvip.com

tent Extraction)评测和 MT 评测等。ACE 是当前正在推动 IE 研究进一步发展的主要动力之一，ACE 与 TIDES 存在着协作关系：ACE 提供基础的人类语言理解技术，以支持更高层的自动语言处理应用，起初只面向英语；TIDES 项目则把基本技术集成化和工程化，面向多语。与 MUC 相比，目前的 ACE 评测不针对某个具体的领域或场景，采用基于漏报和误报为基础的一套评价体系，还对系统跨文档处理(Cross-document processing)能力进行评测，将 IE 技术研究引向新的高度[6]。

ACE 评测 2000 年 12 月正式启动，分阶段进行。刚结束的是 2005 年举办的 ACE evaluation 2005(ACE05)。ACE05 支持三种语言：英语、汉语和阿拉伯语，有八项任务(Primary recognition task)，五个主要任务：EDR (Entity Detection and Recognition)、VAL(Value Detection and Recognition)、TERN (Time Detection and Recognition)、RDR(Relation Detection and Recognition)、VDR (Event Detection and Recognition)；三个提及层次的任务(mention-level tasks)：EMD(Entity Mention Detection)、RMD(Relation Mention Detection)、VMD(Event Mention Detection)。另外还有三个诊断性任务(diagnostic task)：EDR co-reference(Given correct mentions)、RDR Given correct Entities, values and timex2 和 VDR Given correct Entities, values and timex2。此次评测共有 15 家单位参加，国内北京大学、哈尔滨工业大学、中科院、厦门大学和香港理工大学等 5 家单位参加了 ACE05 的测评[7]。

2.3 TREC 会议

TREC(Text REtrieval Conference)会议也是在 TIPSTER 和 TIDES 项目下最权威的国际评测会议，由 NIST、DARPA 和 ARDA 共同主持，始于 1992 年，已举办了 14 届，刚结束的是 TREC2005。TREC 追求的主要目标是：以大规模测试集为基础，推动 IR 研究；经由开放式论坛，使与会者交流研究成果与心得，增进学术界、产业界与政府交流互通；经由对真实检索环境的模拟与重要改进，加速实验室研究技术转化为商业产品；发展适当且具应用性，供各界遵循采用的评估技术。TREC 提供一个4.5 GB 的用 SGML 语言标记过的文档集，作为训练和测试集，包括财经报道、新闻报道、领域文章和公文等。会议针对当前热点设置若干主题，早期主要有两个任务：Routing 任务和 Ad hoc 任务。近年来，评测主题转移到更新的研究项目上，如从文本库到文本流的检索、自动问答、跨语言检索、多媒体对象检索等。评测采用 Judgment Pools(判定池法)，将参评者提交的检索结果运行排序，按照一定原则，选择一些合并到判定池，生成测试集。虽然该方法遭到些质疑，但面对超大规模测试语料，采取手工评测显然不可能的情况，它还是一种可行的、相对公平的方法。参加 TREC 的单位包括许多国际著名的大学及公司，国内先后参加该项评测的单位也很多，如中科院计算所，哈尔滨工业大学，大连理工、北京大学、清华大学，复旦大学等[8]。

2.4 国内 863 评测

为了解国内外中文信息处理与智能人机接口技术领域现状，检查 863 计划信息领域相关课题进展情况，推动技术进步与成果的应用及产业化，国家 863 计划计算机软硬件主题专家组从 1991 年起，举办了 8 次该领域的技术评测，2003 年以来，组织方式更规范合理并逐步向国际惯例靠拢，评测项目和参与单位都大大增加，产生了广泛影响。刚结束的 2005 年度 863 评测，采用通过 Internet 发布数据和提交结果的方式，共设三大类评测，分别为机器翻译、信息检索、语音识别。IR 评测只设置一个项目：相关网页检索。目的是检测互联网环境下大规模数据的中文 IR 技术的研究现状和系统有效性。测试集是 CWT100g，主题(topic)模拟用户需求，由 4 个字段组成：编号(num)、标题(title)、描述(desc)和叙述(narr)。本次评测参加单位是中科院自动化所、哈工大信息检索研究室、清华大学系统与技术国家重点实验室、北京邮电大学模式识别实验室和北大计算机研究所等五家[9]。

另外，北大网络实验室和计算语言学研究所于 2004 年建立了以大规模中文 Web 信息为测试集的 IR 研究论坛：中文 Web IR 论坛(CWIRF)，目标是推动中文 IR 技术。已举办两届，由 CWIRF 提供统一的评分程序，通过开放论坛，供参加者交流。即将举办的第三届 SEWM06 评测有两个任务：中文 Web 检索和中文网页分类。重点是有效的主题提取和导航搜索方法及大规模网页分类的核心技术，其中包括网页文本分类以及链接关系辅助等。

2.5 其他国家、地区相关评测

欧洲一体化进程推动下，对于众多欧洲语种，有效实现跨语言检索意义重大，2000 年开始，在欧盟 IST 资助下，由意大利、德国、瑞士、西班牙等国及美国 NIST 共同发起了跨语言评测论坛(Cross Language Evaluation Forum, CLEF)，旨在通过提供对欧洲单语和跨语检索系统进行评测的基础设施，建立可重用的基准测试语料库和查询集，鼓励单语检索向多语检索转换，促进欧美和亚洲研究人员学术交流。会议就一些独立检索任务，如单语检索、双语检索、多语检索，科学文献检索等展开评测，近几年，开始对多语问答、跨语语音文档检索及图像库跨语检索等任务进行评测。2005 年 9 月刚结束的 CLEF 2005，共 74 个机构参加，欧洲 43 个，北美 19 个，亚洲 10 个，南美和澳大利亚各一个[10]。

维普资讯 http://www.cqvip.com

日本国立情报学研究所(National Institute of Informatics, NII)等部门针对日语和其他亚洲语言的文本检索,于1998年举办了搜索引擎评价型国际会议(NII Collection for Information Retrieval Systems, NTCIR),旨在通过为实验提供大规模可重用语料和允许跨系统比较的通用评测基础设施,促进IR、IE及相关研究;为研究团体提供交换意见的非正式平台;研究文本处理技术的评测方法和构建大规模可重用语料的方法。刚结束的NTCIR Workshop5既关注传统实验室型的检索系统评测,又关注从文本检索到IR的转变、自动文摘以及评测方法本身的技术革新。第二届开始,国内的中科院等单位开始参加该评测[11]。

尽管针对IE和IR的评测会议从内容、主题、任务等差别很大,但二者有明显的共性:大规模的数据资源建设与共享和公开的技术评测已成为推动信息处理学科发展的两个重要驱动力,评测后的技术交流与学术讨论也是成为学术和产业界未来更大发展的契机。

3 应用目的与适用领域

IE与IR不仅仅在发展轨迹,评测方法上存在不同,在其他诸多问题上也各具特点。

首先,二者的应用目的不同。IR技术从文档库中检索相关文档,用户必须从找到的文档子集合中翻阅自己所要的信息。核心在于从文档集中为用户检出最相关的子文档集,或者按检出文档的相关程度进行排序,作为对检索用户所提出查询的回应。IE的任务是将源文本所包含的信息内容析构出来,并按模板的结构组织存储,形成结构化的信息库。通过对原文本信息内容的分析,抽取出有意义的事实,生成满足用户要求的简洁的信息。

其次,二者的适用领域也不同。IR系统通常是领域无关的,主要是从大量的文档集合中找到与用户需求相关的文档列表;而IE系统则通常是领域相关的,旨在从文本中直接获得用户感兴趣的事实信息,只能抽取系统预先设定好的有限种类的事实信息。

另外,多数IE研究是从以规则为基础的计算语言学和自然语言处理技术发源的。而IR则更多地受到信息理论、概率理论和统计学的影响[12]。

4 主要研究方法

4.1 信息检索常用方法

广义上讲,IR方法分为基于统计的和基于语义的方法两类。传统方法有全文扫描、倒排文件、向量模型及聚类等。从处理的技术来讲,当前包括自然语言处理、人工智能、模式识别、机器学习、神经网络、数理统计、运筹学等在内的技术纷纷被应用于现代IR系统。近年来一些引入语义的检索方法,如自然语言处理、隐性语义标引、神经网络等,通过计算语义相似度(semantic similarity)来对用户的查询词汇进行语义相似词汇扩展,或者改进用户查询与文档相似度的计算方法。词汇间语义相似度的计算需要依靠语义辞典[13]。

从其发展看,基于统计的方法长期占据主流地位;随后针对统计方法的不足,将语言处理技术引入到了IR领域。基于统计方法的IR模型一直是该领域的重要研究课题,较典型的是布尔模型、向量空间模型、概率检索模型等。在此类模型中文档被表示为关键词(也称索引词)的集合。布尔模型最简单,文档索引词权重只有0、1,表示是否含有索引词,用户查询用布尔表达式进行,但其分类能力有限,早期的商用IR系统普遍采用该模型。向量空间模型中文档和用户查询都分别表示成由索引词的权重$W_{i,j}$构成的N维向量D和Q,该模型的关键问题是索引词权重$W_{i,j}$的计算和用户查询与文档相似度Similarity(q,d)的计算。与布尔模型相比,向量空间模型在实际系统中的表现更佳,且计算效率较高,因而获得了广泛的应用,如SMART系统。在概率模型中文档和用户查询也被表示为索引词集合的形式,通常采用索引词在文档中的统计分布等参数,计算任意文档d与给定用户查询q相关概率P(q|d)。典型模型如Bayes推理网络模型,它将计算用户查询与文档的相关度问题转化为由Bayes网络计算用户查询与文档的联合概率问题[14]。

与基于统计的IR方法相比,自然语言处理方法以对文本的语言结构分析和语义分析为特色,将信息处理的层次深入到文本内容,而非仅依据文本索引词的统计信息,如专名的自动识别与分类。因为IR对实时性有较高要求,所以通常采用的自然语言处理方法一般只进行浅层次处理(shallow parsing)。另外由于统计方法也是自然语言处理的主要方法,因此自然语言中处理中基于平均互信息的方法,基于相关熵的方法等都有应用于IR技术中[13]。

4.2 信息抽取常用方法

IE研究方法基本上也划分为两类:基于知识工程方法(规则方法)和基于统计学习方法(统计方法或学习方法),有时也称作自动训练方法(Automatic Training Approach)。

知识工程方法主要靠手工编制规则使系统能处理特定知识领域的IE问题,要求编制规则人员对该领域知识有深入了解,开发过程耗时耗力;统计方法不一定需要专业知识人员,系统主要通过学习标记好的语料库获取规则。任何对该知识领域比较熟悉的人都可根据事先约定的规范,标记语料库,经训练后的系统能处理未见过的新文本。两类方法各有利

维普资讯 http://www.cqvip.com

弊，都有对应的已实现的 IE 系统，目前较多成功的 IE 应用都是针对统计方法。采用统计学习方法的基本思想是建立一个统计学习模型，如概率模型，然后从大规模真实语料中选取自然语言输入：$x_1 \sim x_n$ 为样本，并以此对概率模型进行统计学习，再对给定的输入 x_{n+1}，用学习好的概率模型进行预测。该方法比知识工程方法投入资源少，开发周期短，但通常需要足够数量的标注语料作为训练数据，才能保证处理质量。统计方法中有很多语言模型，如 N-元文法模型等。在 IE 中涉及最多的统计语言模型是隐马尔可夫模型和最大熵模型。

研究人员已经意识到两种方法的互补性，理论研究和实践经验都表明，统计学习方法在解决自然语言理解的浅层问题方面比较有效，知识工程方法解决自然语言理解深层问题不可或缺。二者的有机互补，应该是解决问题的合理途径。

4.3 评价方法

IE 评测指标有三个：查准率、查全率和 F 评价(F-Measure)。查准率 P (Precision)表示在系统得到的全部答案中，正确答案所占的比值，描述系统抽取的信息中，有用信息是多少；查准率/ 查全率 R (Recall)指在所有答案中(包括系统得到的和不应该忽略的)，正确答案所占比值，表示在应该得到的信息中，已抽取出了多少；F 值评价(F-Measure)是查准率和查全率的加权几何平均值，用于综合评价系统性能，计算公式为：$F=\frac{(\beta^2+1)PR}{\beta^2P+R}$ 其中 β 是一个预设值，决定抽取系统最终输出结果，是对准确率 P 侧重还是对查全率 R 侧重，通常设定为 1。用一个 F 值可以分析比较，得出系统的综合性能评定。

早期衡量 IR 系统的指标也是上述三种，其含义类似。随着测试集规模的扩大及对评测结果理解的深入，更准确反映系统性能的新评价指标逐渐出现，包括平均准确率(Mean Average Precision, MAP)：单个主题的 MAP 是每篇相关文档检索出后的准确率的平均值；主题集合的 MAP 是每个主题的 MAP 的平均值，MAP 是反映系统在全部相关文档上性能的单值指标。R-Precision：单个主题的 R-Precision 是检索出 R 篇文档时的准确率，其中 R 是测试集中与主题相关的文档的数目；主题集合的 R-Precision 是每个主题的 R-Precision 的平均值。P@10 是系统对于该主题返回的前 10 个结果的准确率，P@10 常常能比较有效地反映系统在真实应用环境下所表现的性能。

5 当前值得关注的问题

5.1 信息检索方面

目前的热点主要集中在两个方面：一是网络化，是 IR 对于网络环境的适应，突出表现为 Web 搜索，在某种程度上已经成为 IR 技术的代称，提高 Web 查询质量是目前智能化 IR 研究中最关键的一个问题，目前主要途径是检索算法的改进；二是自然语言处理，这是关键词检索技术发展到一定程度的必然需求，是 IR 在深度上的发展，改进和优化搜索引擎的途径应该是增加语意理解(如主题词和关键词的提取)。在广度方面，多媒体信息检索中基于内容的检索(CBR)是近几年的研究热点。目前该方法主要是针对底层语义特征，高层语义检索能力不足，今后会在结合人工或半自动领域知识标注，利用高级语义进行多媒体检索方面做更多的尝试。多领域学科交叉的趋势将更加明显，以机器学习、数据挖掘为代表的统计方法和计算语言学相关的计算模型与知识库已经逐步与 IR 相融合。一些原本用于自然语言处理领域的方法，如最大熵、HMM 模型、马尔可夫随机场模型将进一步应用于 IR。

同时检索系统评估对于系统的研发和应用影响显著。从 TREC 来看，目前任务设置向高精度、细粒度和大规模三个方向倾斜，如：高精度文档检索任务(HARD)、新信息检测任务(Novelty)、问答任务(QA)、TB 级检索(Terabyte)等。总体上从结果、性能及稳定性等方面来看，国外主流的 Web 检索技术已比较成熟，并已在日常信息获取中发挥作用；更高精度和更细粒度的检索技术仍处于实验室阶段。另外，国外由 TREC 设立 Web Track 测试项目，提供英文 Web 测试集；NII 提供日文 Web 测试集。国内缺乏大规模的中文 Web 测试集也是制约中文 IR 技术前进的障碍。目前北大以中文为主的 Web 测试集 CWT(Chinese Web Test collection)就是力图解决这一问题，以推动中文检索技术的发展。

5.2 信息抽取方面

在 IE 研究中，命名实体识别(NER)是目前最有实用价值的一项技术，NER 的完成水平直接关系到 IE 的质量。关于名字的识别国内外都已经有了大量的研究工作，尤其在人名、地名识别方面，而机构名识别相对较少，涉及中文机构名识别的更少。今后这方面的研究工作深度还将加大。指代消解是一项重要且非常困难的研究，在 IE 中是一个关键问题，也是自然语言处理的重要内容。到目前为止，还没有较好的自动的指代消解技术和方法。

随着互联网的应用，用户对于 Web IE 的有效性越来越期待，很多商业应用的需求也越来越明显和迫切，所以这一领域的发展，必将继续成为 IE 技术的一个热点。为了能在多语种上实现 IE，探讨自动抽取体系结构也是有必要的。而多语信息抽取技术(Multilingual IE)[15]更是一项新的发展。英国 Sheffield 大学的 ANNIE 信息抽取系统起初也是针对英语单语种的，但其在后来开发的 GATE 平台上，很多模块可以直接重用[16]。

维普资讯 http://www.cqvip.com

语义网络给 IE，特别是基于本体的信息抽取(Ontology-Based Information Extraction，OBIE)带来了全新的挑战。传统 IE 和 OBIE 最重要的不同就是后者使用了形式的本体而不是平面词典或专有名词字典结构，并且 OBIE 不仅是找到特定类型的待抽取实体，同时能通过在本体中把它链接到其语义描述上来识别它，使得实体可以在文档内被跟踪，并在 IE 过程中丰富其描述。尽管存在本体统一标准的问题，但在类似于互联网的"求同存异"的思想下，OBIE 技术必将得到更大的拓展空间，其发展也将对 IE 的语义资源建设及篇章分析与推理等深层次的问题的解决产生促动。另外，研究将多文档文摘和信息抽取相结合，从相似的文本中自动获取模板，将模板合并生成多文档文摘，再将 IE 与多文档文摘的优点结合起来应用于开放域信息检索，也具有很大的应用前景。早些时文[17]也做过类似的研究即把多文档文摘技术看作 IR 的后处理，IE 的技术应用[17]。

结束语 本文旨在分析比较并介绍信息抽取与信息检索技术各自发展历史及相关重要问题。在如此短小的篇幅下去穷尽这两个领域当前的所有问题，几乎是不可能的；仅希望通过这一基础性调研工作对今后的研究有所启发帮助。文中论述存在偏颇之处，恳请批评指正。

致谢 衷心感谢中科院计算机语言信息工程研究中心冯冲博士提供的宝贵意见与资料！

参 考 文 献

1 王晓龙，关毅. 自然语言处理[M]. 北京：清华大学出版社，2005. 136～137
2 TIPSTER Text Program. [EB]http://www. itl. nist. gov. iaui. 894. 02/related_projects/tipster
3 Cunningham H，Bontcheva K. Named Entity Recognition (presentation). [Z] RANLP (European NLP Conf. on recent Advances in Natural Language Processing). Borovets，Bulgaria，Sep. 2003
4 孙斌. 信息提取技术概述. 语言信息处理，2002
5 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. 10
6 The ACE02 Evaluation Plan. [EB] ftp://jaguar. ncsl. nist. gov/ace/doc/ACE-EvalPlan- 2002. pdf
7 NIST 网站[EB] www. nist. gov/speech/tests/ace
8 TREC 网站[EB] http://trec. nist. gov/
9 863 评测网站[EB] http://www. 863data. org. cn
10 CLEF 网站[EB] http://www. clef-campaign. org/
11 NII 网站[EB] http://research. nii. ac. jp/ntcir/index-en. html
12 Gaizauskas R，Wilks Y. Information Extraction：Beyond Document Retrieval. [J] Computa- tional Linguistics and Chinese Language Processing，1998，3 (2)：17～60
13 Lee L. Similarity-Based Approaches to Natural Language Processing：[Ph. D. thesis]. Harvard University Technical Report TR-11-97
14 Yates R B，Neto B R. Modern Information Retrieval. [M]. Addison-Wesley，New York，1999
15 Maynard D. Multi-Source and Multilingual Information Extraction. ppt. [J]. September 2003
16 Bontcheva K，Maynard D，Tablan V，et al. [J/OL] GATE：A Unicode-based Infrastructure Supporting Multilingual Information Extraction，2003
17 White M，Korelsky T，Cardie C，et al. Multidocument Summarization via Information Extraction [A]. In：Proceedings of the First International Conference on Human Language Technology Research [C]. 1998. 36～44

(上接第 132 页)

时，各个 agent 之间通过相互协作来完成挖掘的全过程。其中，多维文本分析引擎以文本预处理为基础，以文本挖掘为支撑。文本超立方体中的维来自于文本预处理所得到的文本特征属性，例如时间、作者等。而文档主题类别的生成以及文档之间关系的聚类分析又依赖于文档挖掘技术。反过来，多维文本分析引擎又为文本挖掘提供了有效的可视化手段和特征修剪工具。文档集合的特征修剪结果可以展现给用户，也可以作为挖掘对象输入到文本分类 agent 和文本聚类 agent，如图 2 所示。

结束语 在 Web 信息充斥的情况下，Web 挖掘是一个具有极大潜力的研究方向。一些国际会议，例如 KDD97、IJCA I'99 等，已经或即将举行有关 Web 挖掘的专题讨论，对其理论、体系结构、算法等展开研究，Web 挖掘的一个重要应用方向将是电子商务系统。而与电子商务关系最为密切的是用法挖掘(Usage Mining)，也就是说在这个领域将会持续得到更多的重视。另外，在搜索引擎的研究方面，结构挖掘的研究已经相对成熟，基于文本的内容挖掘也已经有许多研究，下一步将会有更多的研究者把多媒体挖掘最为研究方向。基于 Agent 的研究也大量的出现，总之，Web 挖掘已经成为数挖掘和 Internet 结合的一个重要方向，吸引着越来越多的研究者投入到这种新的技术中去。

参 考 文 献

1 Kosla R，Blockeel H. Web mining research a survey. SIGKDD Explorations，2000，2：1～15
2 Pal S K，Talwar V，Mitra P. Web Mining in SoftComputing Framework：Relevance，State of the Art and Future Directions. IEEE Transactions on Neural Networks，2002，13(5)：1163～1177
3 Madria S K，Bhowmick S S. Research issues in web data mining. In：The 1st International Conference of Data Warehousing and Knowledge Discovery，1999. 303～312
4 Srivastava J，Cooley R. Web usage mining：discovery and applications of usage patterns from Web data. SIGKDD Explorations，2000，1(2)：12～23
5 陈莉，焦李成. Internet/Web 数据挖掘研究现状及最新进展. 西安电子科技大学学报，2001，28(1)
6 王继成，潘金贵，张福炎. Web 文本挖掘技术研究. 计算机研究与发展，2000，37(15)