基于智能 Agent 的中文元搜索引擎模型研究

王浩鸣　张曰贤　吴志军　史西兵
（西安财经学院计算机科学与技术系，西安 710061）
E-mail：hmwang@mail.xaufe.edu.cn

摘　要　论文讨论了现有搜索引擎技术的缺点，比较了中文与英文分词方法的差别，描述了中文文档的基于无词典信息抽取方法。通过分析用户搜索信息的历史，构建用户的个性化搜索模型，并将这些文档进行分档，在本地服务器上进行整理与保存。文中对系统涉及的关键技术：文档类关键词提取方法、用户特征的建立方法、页面价值评比算法等进行了描述。最后，对进一步研究指明了方向。

关键词　智能代理　个性化模型　信息检索　类关键词

文章编号 1002-8331-(2005)31-0154-03　**文献标识码** A　**中图分类号** TP18；TP311.1

Research of the Chinese Meta-Search Engine Model Based on Intelligent Agent

Wang Haoming　Zhang Yuexian　Wu Zhijun　Shi Xibing
（Dept. of Computer Science，Xi´an University of Finance and Economics，Xi´an 710061）

Abstract：This paper introduces the method extracting the words from documents without dictionary，builds the personalized model according to the web pages the user has visited the Internet，and discusses the key technologies，such as classifying methods and calculating the page value.The relative methods，the storing and the querying of semi-structure information are put forward.In the last paragraph，the future work is mentioned.

Keywords：intelligent agent，personalized model，information retrieval，class keywords

1　问题的提出

由于 Internet 是一个开放、分布的信息空间，它本身所固有的特点已经明显地阻碍了人们充分地使用 Internet 内的信息资源。用户在 Internet 内进行信息检索时可能会出现“信息过载”或“资源迷向”，即用户不知道如何有效地利用资源，以致达不到所期望的高查全率与高查准率[1,2]。

本文提出一种以智能信息 Agent 为工具的 Internet 中文信息智能化获取方法，从用户过去浏览的网页中自动学习用户的浏览习惯与基本需求模型，从而为用户提供具有个性化 Web 信息导航服务。

1.1　目前已有的技术

智能 Agent 是智能化程序的集合，它们能够学习用户的需求，并利用搜索引擎等系统提供的现有服务来帮助用户检索所需的信息，这类系统的组成基本类似：由代理服务器模块及学习模块构成。代理服务器模块用于实现用户与 Web 之间的交互，而学习模块则向服务器提供用户模型信息，从而使用户与 Web 的交互更具个性化。代理服务器储存已访问过的文件地址或访问内容，学习模块则使用这些信息提取并建立用户兴趣模型。涉及到的关键技术有：

(1)关键词提取：分为基于词频的提取技术和基于语义分析的提取技术。但由于对自然语言理解的研究尚未达到一定深度，目前基于词频的关键词提取技术仍然占据统治地位。

(2)查询结果的聚类分析：搜索引擎的查询结果中经常出现重复的内容，即相同的内容在很多网页中出现，从而导致返回给用户的结果很多但有用信息很少，因此通常采用统计学中的聚类分析对查询结果进行分类，剔除相同的结果，并从每一类结果中挑选出一个最具有代表性的结果提交给用户。

1.2　现有系统的缺陷

(1)非个性化检索方式适应用户兴趣变化的能力较差。现有大部分信息检索系统采用关键词输入方式进行检索，对任何用户都采用同一种模式，很容易让用户感到迷茫，有时用户也无法准确地表述自己的兴趣。尽管有些系统为此进行了改进，确实改善了检索效率。但由于没有不同个性化模式之间的相互学习和信息共享机制，并不能很好地适应用户兴趣变化。

(2)用户与检索系统的交互方式比较单调。针对不同需求的用户，提供不同的输入方式是目前现有系统所缺少的。

(3)缺少分布式智能信息检索和适应信息源信息变化的能力。现有系统主要通过学习用户的历史关联信息，在线引导用户检索感兴趣的信息。这种为用户导航的方式无法避免用户浏览以前已经浏览过而现在不需再看的文档或链接。此外，由于没有有效地适应信息源信息变化的机制，不能及时为用户提供新的信息，因而无法为用户快速定位感兴趣的主题。

应该指出，上述几项开发成果，基本都是基于英语信息的获取。由于英语与汉语的差异，对广大使用汉语的 Internet 用户而言，困惑依旧，而其浏览、获取汉语信息占所需信息的大多数，因此，如何有效改善中文信息获取的质量，已成为影响 In-

基金项目：陕西省自然科学基金项目(编号：2004F06)资助
作者简介：王浩鸣(1968-)，男，汉族，江苏省靖江人，副教授，西安交通大学电信学院博士生，主要研究方向：信息检索与安全，数据库应用

ternet 中文信息资源优势发挥程度的重要因素。

本文的目的是利用智能 Agent 技术构建元搜索引擎，通过在本地服务器数据库中记录用户的行为特点，并构造用户访问模型，从而为用户提供具有个性化的服务。

本文的内容按如下结构组织：第 2 节描述系统的结构并介绍了系统各部分的功能；第 3 节讨论系统构建中的关键技术；第 4 节与第 5 节对本文进行总结并对下一步的研究指明方向。

2 系统结构描述

系统的最终使用者是广大的 Internet 使用者，其目的就在于帮助广大用户能更好地利用网络信息资源，因此，本系统应该能够自主处理网络的信息资源、收集用户感兴趣的信息并将它们加以过滤，将“有用的”信息在本地服务器加以保存，它应具有软件的易使用性、界面的友好性、推理的适用性、系统的可移植性等特点。

系统的研究成果显然不是想取代 Yahoo、Google 等通用搜索引擎，而是想为用户提供一个具有个性化的搜索引擎入口，系统的结构可以概括为：信息的收集、分类与发布等方面。结构如图 1 所示：

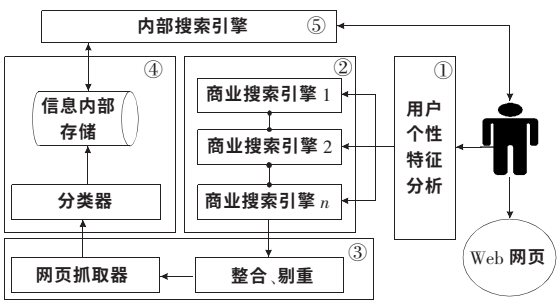

图 1 系统结构示意图

用户在使用时可以直接访问 Internet 也可以通过本系统提供的具有个性化分析功能的门户网站访问 Internet。系统由 5 部分构成，其中：

部分①为用户个性特征分析模块，本模块通过自主学习的方式完成指定功能，用户在刚开始使用时系统并没有用户的兴趣目标，系统通过学习用户检索的目标网页逐步建立起用户的个性特征，并将该内容写入到本地数据库中，在用户下次使用时，上次建立的个性特征将起指导作用。

部分②为若干个商业搜索引擎的集合，系统首先分析商业搜索引擎的搜索语法表达，再分析用户的个性特征，然后将用户的搜索意愿以合适的方式提供给这些商业搜索引擎，并将返回的结果交由部分③进行处理。

部分③为信息获取模块，其功能是将从多个商业搜索引擎中获取的结果进行整合并剔重，得到相关网页信息列表，然后通过“网页抓取器”将这些内容抓取到本地，其中网页抓取器为本模块的关键部分，它需要根据用户的个性特征及网页价值判断算法，对欲读取的页面进行判断以决定是否需要下载到本地，显然如果将网页内容全部下载到本地再对页面进行价值判断是比较简单的做法，但它可能引起较大的网络流量；如果能够在网页下载前先对其价值进行判断可以解决网络流量的问题，但会占用对方主机的资源，两种方法各有所长，从已得到的实验数据来看，Internet 网上有些网站的内容其实是其它一些站点内容的重复，这些信息对用户来说可能并不需要，因此，如果将这些内容全部下载到本地，再进行剔除，未必是个好的方法。

部分④为信息分类与存贮模块，其功能是将部分③得到的内容根据规定的分类原则进行分类，并保存到本地的存储介质上，保存的信息根据网页内容的不同分为两类：一类是保存信息本身，另一类是保存网页链接地址。“分类器”是本模块的关键部分，对于已经保存的信息需要定时更新，总是将最“有用”的信息存放在最容易找到或查询代价最小的地方。

本模块涉及到非结构化信息的存储技术，在本系统中第一步采用简单的变通方法，该技术本身将作为进一步研究内容。

部分⑤是面向用户的服务窗口，表现为内部搜索引擎，它与其它商品化搜索引擎的不同之处在于它只对系统内部存储的数据进行检索，相比而言它得到的结果应该比商用网络搜索引擎更具权威性。

总的来说，信息获取不外乎是“查全”与“查准”两方面，要找到一个对所有用户都能满足这两方面要求的系统并不容易，因此，本系统设想首先在特定的几个学科提供个性化的服务，在时机成熟后再推广到其它学科。

3 系统关键技术探讨

从总体上看，本系统需要进行的研究主要包括：

(1)知识的获取和融合。主要研究：非规范知识的获取，矛盾知识的融合，时变知识的融合，不确定知识的融合和多表示知识的融合等。

知识的获取首先要解决检索需求的表达问题，人们在进行信息查找时，往往难以准确表达自己的信息需求，这就需要系统利用自主学习的功能对用户的信息需求进行分析，主动搜集用户平时感兴趣的信息，分析用户需求的个性化特点，从而建立用户个性化需求模式，引导和帮助用户正确表达其思想。

(2)知识的转换和传播。重点解决：①不同类(非规范)知识表示的相互转换；②内涵和外延、定量和定性知识表示的转换；③不同抽象层次之间(面向知识内涵发掘)的知识表示转换。

以因特网上知识为代表的非规范知识大部分是非结构化或半结构化的，与通常存放在数据库里的结构化信息不同，非结构化的自由文本通常使用自然语言处理技巧，其抽取规则主要建立在词或词类间句法关系的基础上，需要经过句法分析、语义标注、专有对象的识别和抽取规则的制订等步骤进行处理。半结构化数据介于自由数据和结构化数据之间，其特点是没有事先给定的数据模式，或者数据模式对数据的约束不强，模式的规模比较大（有时甚至可以大过数据），或是经常变动的，数据未赋予严格的类型。非规范知识进行转化后，才能得到有效的共享和利用。

用户个性获取中的关键问题是不确定性问题。根据处理不确定性问题的方法可以将这些技术分为以下几类：

(1)基于贝叶斯网络的用户兴趣获取技术；

(2)基于合作过滤(Collaborative filtering)的用户兴趣获取技术；

(3)基于 DST(Dempster-Shafer Theory of evidence)的用户兴趣获取技术；

(4)基于模糊逻辑的用户兴趣获取技术；

(5)基于机器学习的用户兴趣获取技术等。

本系统采用的具体方法为：

(1)建立各个具体学科领域的学科类关键词分布情况，以

特征向量的形式表示，如 $S_1,\cdots,S_m$；

(2)根据用户访问的历史页面，从中抽取出用户特征类关键词，以特征向量的形式表示，如 U_k；

(3)计算 U_k 与 $S_1,\cdots,S_m$ 之间的余弦距离，其中距离最小的就是最接近的兴趣领域。

可以看到，相似度计算方法都以余弦角公式为计算基础，在向量模型及其扩展模型中广泛使用。但正如 Cornell 大学的 Salton 所言：利用测试余弦角获得向量相似度的方法并没有严格的理论根据[11]。

3.1 确定学科领域类关键词分布

因为表示学科领域特征所采用的词条数量有限，所以有必要对文本信息进行词干抽取处理。传统的方法是通过切分词的方法，但使用这种方法的前提是已经有比较精确的词典存在，但任何常用词典和专业词典都不可能涵盖所有的词语，据吴立德教授统计：在含有 15 000 个词条的语料库中，即使使用具有 70 000 个词条的词典，仍然有 30%以上的词条没有被收录[6]。切分歧义和词典生词限制了机械分词的分词准确度，而且，词典对分词精度造成的影响远大于分词方法自身产生的歧义切分错误[11]。

为了达到领域无关性和时间无关性，本系统采用文献[11]提出的不需要词典的词干抽取方法，该文献指出：在中高频词条的处理上，精确率达到 94%，完全能够符合词条频度敏感的中文信息处理工作对分词准确度的要求。系统第一次通过自主学习系统提供的、已经经过人工标引过的材料，从而建立特定学科领域的类关键词分布模型。

设有已经标注好类别的文档 D_{ij}（其中：$i\in(1,\cdots,m)$ 表示 m 个分类类别；$j\in(1,\cdots,n)$ 表示某个类别已标注好的 n 篇文档），分别对每篇文档进行抽词处理，设得到类关键词 $k_{ij1},\cdots,k_{ijp}$，在此基础上，统计每个分类类别 i 的类关键词分布情况，按其出现概率的大小顺序进行排序 $k_{i1},\cdots,k_{ip}$，使用这 p 个类关键词构成该分类类别的特征向量 S_i，使用同样的方法可以得到其它各个分类类别的特征向量 $S_1,\cdots,S_m$。

3.2 用户个性特征的获取

本系统采用基于关键词提取的用户兴趣获取技术[5]。设用户浏览过并保存在本地的文档为 $D_i(i=1,\cdots,m)$，分析每篇文档，提取出每篇文档的类关键词 $k_{i1},\cdots,k_{ip}(i=1,\cdots,m)$，统计得到的所有类关键词并采用 3.1 节所示方法构建用户特征向量 U_k，以建立用户个性化模型。通过计算 U_k 与 $S_1,\cdots,S_m$ 之间的余弦距离，确定 U_k 的最大归属度。同时将与此相关的数据，如用户的登录名、专业领域等内容写入本地数据库。当用户再次访问时，这些数据将用作为用户提供个性化服务的依据，同时用户的最新查询结果可用来修正用户的个性化模型。

显然基于出现概率的统计方法最为简单，但存在着一些缺点，更为有效的统计方法需作进一步的研究。

需要指出的是，在本系统中用户的个性化服务并不是必须的，如果用户不选择个性化服务，可以直接进行信息的搜索。

与本步骤同时进行的还有信息分类工作：提取文档 $D_i(i=1,\cdots,m)$ 的类关键词 $k_{i1},\cdots,k_{ip}(i=1,\cdots,m)$ 构成文档特征向量 $D_i(i=1,\cdots\mathrm{m})$，通过计算 D_i 与 $S_1,\cdots,S_m$ 之间的余弦距离，确定 D_i 的最大归属度，考虑到某篇具体的文档可能涉及到多个领域，因此，设置一个类别阈值，如果余弦距离超过该阈值，即认为该文档 D_i 属于该学科领域。

3.3 半结构化信息的转换与保存

Web 上的数据与传统的数据库中的数据不同。传统的数据库都有一定的数据模型，可以根据模型来具体描述特定的数据。而 Web 上的数据非常复杂，没有显式的模式描述，每一站点的数据都各自独立设计，并且数据本身具有自述性和动态可变性。因而，Web 上的数据具有一定的结构性，但又是一种非完全结构化的数据，即半结构化数据。半结构化是 Web 上数据的最大特点，常使用 OEM(the Object Exchange Model)模型进行描述。

OEM 是自描述对象模型，专为表达半结构化数据而设计。它最初的目的是为异构数据源间的数据交换提供高度灵活的转换工具。不同应用中的 OEM 模型大多在原模型的基础上作了一些小的改动，在 OEM 模型中，数据的组织可以看作是一张图，它由节点和带标签的边组成。所有的实体都是对象位于节点处，边表示对象之间的联系。对象以惟一的对象标识符来表示，可分为原子对象和复合对象。原子对象是仅含有一个原子型值的对象，如：整型、实型、字符串型、html 型、Java 型等。复合对象是对象参量的集合，以一系列(对象，边)的数据对来表示。针对 Web 的半结构化数据的表示与存储，国内外已有许多相关研究，并有多种半结构化数据模型及查询语言被提出。

其查询语言通常采用两种途径来研究：一种是以 SQL 或 OQL 语言为基础，增加必要的机制，使其能够表达一组查询；另一种是以某种语法进行适当的变形，成为一种便于使用的查询语言。根据这两种途径所设计出的查询语言非常相似。

在本文中，这部分内容将不作为研究的重点。

4 下一步研究内容

从系统已经实现的功能来看，对于特定的学科取得比较满意的效果，但还不能称作是通用的元搜索引擎，原因是涉及到的以下几项技术需作进一步的研究：

(1)网页价值判断算法。传统的网页评价技术是根据网页的内容，利用单词匹配、词频统计来评价网页。1998 年出现了在内容关联的基础上进一步利用 Web 的超文本链接结构来评价网页的算法，最近几年来，国外许多文献作了关于利用 Web 的超文本链接评价网页的研究[8~10]，其中以 PageRank 和 HITS 算法为代表。本系统将在这些技术的基础上研究合适的网页价值算法。

(2)页面信息抽取技术。在本系统中采用支持向量机(SVM)的方法判断目标页面的学科类别，这对于能够确定所属学科的页面来说比较准确，但 Internet 网上存在大量的、以商务为主的页面内容，这些信息的组织和表现形式与学术类页面迥异，因此需要采取另外的方法对这些页面中的内容加以分析、保存与再发布。

5 小结

本文首先指出通用搜索引擎难以为用户提供个性化的服务，设想在通用搜索引擎的外围附加一层个性化服务接口，以元搜索引擎的形式出现，系统将元搜索引擎返回的结果进行类关键词提取、文档分类以及与用户个性特征匹配等处理，从而给用户提供具有个性化的服务。

在元搜索引擎的设计中，讨论了用户个性化特征获取、关

(下转 204 页)

图 4　管线测量记录

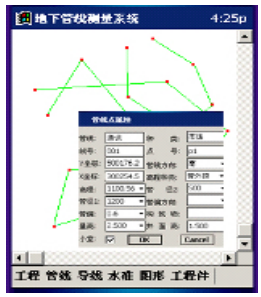

图 5　管线草图与属性查询

4　结语

基于 Windows CE 嵌入式操作系统 PDA 开发管线普查与测量系统软件，实现测量成图一体化的前端数据采集，具有现实性和可行性，与传统的电子手薄和电子平板等作业方式比较，作业速度与精度大幅度提高，显示出较大优越性。该系统仍需在以下两方面进一步加强：基于管线点多向性和管线点属性的数据库结构的简化；数据质量检查模块的开发。

（收稿日期：2005 年 4 月）

参考文献

1.方允治，赵斌臣，张晓峰.城市地下管线信息管理系统建设中的测量方法[J].山东交通学院学报，2004；12(2)：48~50

2.张金波，周蕙，王鹏等.PDA 用户参考手册[M].北京：中国水利水电出版社，2002

3.Lisa Guadagno，Carla VandeWeerd，Dirk Stevens et al.Using PDAs for data collection[J].Applied Nursing Research，2004；17(4)：283~291

4.崔铁军，李玉，饶欣平.嵌入式 GIS 的发展及开发实践[J].测绘学院学报，2004；21(2)：128~130

5.张时煌，方裕.微型嵌入式 GIS 软件平台的重要意义及发展动态[J].中国图形图象学报，2001；6(9)：900~906

6.北京博彦科技发展有限公司译.[美]Microsoft Windows CE 程序设计[M].北京：北京大学出版社，1999

7.李现勇.Visual C++串口通信技术与工程实践[M].北京：人民邮电出版社，2002

8.Chenchen Qiu，Mark E Orazem.A weighted nonlinear regression-based inverse model for interpretation of pipeline survey data[J].Electrochimica Acta，2004；49(22)：3965~3975

（上接 156 页）

键词提取、信息分类等关键技术，对非结构化信息的存取、网页页面价值判断、网页信息提取等内容作为下一步研究的重点。

（收稿日期：2005 年 4 月）

参考文献

1.汪晓岩，胡庆生等.面向 Internet 的个性化智能信息检索[J].计算机研究与发展，1999；36(9)：1039~1046

2.王继成，萧嵘等.Web 信息检索研究进展[J].计算机研究与发展，2001；38(2)：187~193

3.刘振宇.基于 Agent 技术的 WWW 信息查询系统设计[J].计算机应用研究，2001；(9)：74~76

4.林锦贤，钟春芳.基于 Agent 的网页自适应检索模型[J].福州大学学报（自然科学版），2000；28(3)：72~76

5.蔡智，王煦法，于琨等.互联网中文信息获取研究[J].小型微型计算机系统，2003；(12)：2136~2141

6.吴立德等.大规模中文文本处理[M].上海：复旦大学出版社，1997

7.金翔宇，孙正兴，张福炎.一种中文文档的非受限无词典抽词方法[J].中文信息学报，2001；(6)：33~39

8.S Brin，L Page.The anatomy of a large-scale hypertextual Web search engine[C].In：7th International World Wide Web Conference，Brisbane，Australia，1998-04

9.J Kleinberg.Authoritative sources in a hyperlinked environment[J].Journal of ACM(JASM)，1999；46

10.D Rafiei，A Mendelzon.What is this page known for? Computing web page reputations[C].In：9th International World Wide Web Conference，Amsterdam，Netherlands，2000-05

11.潘谦红，王炬，史忠植.基于属性论的文本相似度计算[J].计算机学报，1999；(6)：651~655

12.王洪，贾惠波，徐端颐.基于人工标引的中文学术期刊文献自动分类算法[J].清华大学学报（自然科学版），2002；(6)：787~790

（上接 179 页）

表 2　两种不同方法的结果对比

分类方法	训练集上留一法评估		独立测试评估	
	正确分类	错误分类	正确分类	错误分类
k-近邻法	63	4	30	5
支持向量机	68	2	32	0

出了一种基于聚类和“信噪比”相结合的特征基因挑选方法，既有效地降低了样本表达数据的维数，又保留了原始数据所蕴涵的信息。支持向量机可以较好地处理高维数样本的分类问题，是一种有效的分析基因表达数据的工具。实验结果表明：利用此方法来挑选特征基因后，用支持向量机可以建立比较理想的癌症预测模型从而获得较高的分类准确率。

（收稿日期：2005 年 6 月）

参考文献

1.Sridhar Ramaswamy，Todd R Golub.DNA Microarrays in Clinical Oncology[J].Journal of Clinical Oncology，2002；20(7)：1932~1941

2.Javed Khan，Jun S Wei，Markus Ringner et al.Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks[J].Nature，2001；7(6)：673~679

3.Alizadeh A A，Eisen M B，Davis R E et al.Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling[J].Nature，2000；403(6769)：503~511

4.Welsh J B，Sapinoso L M，Su A I et al.Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer[J].Cancer Res，2001；61(16)：5974~5978

5.Terrence S Furey，Nello Cristianini，Nigel Duffy et al.Support vector machine classification and validation of cancer tissue samples using microarray expression data[J].Bioinformatics，2000；16(10)：906~914

6.Singh D，Febbo P，Ross K et al.Gene expression correlates of clinical prostate cancer behavior[J].Cancer Cell，2002；1(2)：203~209

7.T R Golub，D K Slonim，P Tamayo et al.Molecular classification of cancer：Class discovery and class prediction by gene expression monitoring[J].Science，1999；286(5439)：531~537