结合本体筛选和文本挖掘的垂直搜索引擎研究*)

赫建营[1] 晏海华[2] 金茂忠[3] 刘 超[4]
(北京航空航天大学计算机学院 北京 100083)

摘 要 针对垂直搜索引擎研究领域的关键技术问题,提出了一个结合本体筛选和文本挖掘的垂直搜索引擎构建思想。首先探讨了作为研究基础的本体和文本挖掘技术,讨论了两者的作用;之后阐述了垂直搜索引擎构建的关键技术,包括基于本体筛选的智能搜索器、结合文本挖掘的网页信息分析及抽取、索引器及查询处理器的构造;最后,对提出的思想进行了实现验证,构造一个面向高校毕业生招聘的垂直搜索引擎原型。

关键词 垂直搜索,本体,本体筛选,文本挖掘

Research of Vertical Search Engine Incorporating with Ontology Filtering and Text Mining

HE Jian-Ying[1] YAN Hai-Hua[2] JIN Mao-Zhong[3] LIU Chao[4]
(School of Computer Science & Engineering, Beihang University, Beijing 100083)

Abstract This paper presents a construction method for vertical search engine utilizing ontology filtering and text mining towards existing problems in the domain. Firstly, it discusses ontology and text mining as well as their applications. Then, we provide a set of key techniques for the construction of vertical search engine which include ontology-based Web crawling, Web page analyzing combined with text mining, indexer and searcher constructing. Finally, an evaluation of our proposed ideas is presented by implementing a prototype of job hunting search engine towards college students.

Keywords Vertical search, Ontology, Ontology filtering, Text mining

Internet是一把双刃剑,一方面人们把越来越多的信息推送到网上,极大地推动了信息的共享,另一方面过多的垃圾信息湮灭了用户想要真正获取的"真知灼见"。如何从呈指数级增长的Internet资源库中迅速准确地获取所需信息成为一个亟待解决的问题。搜索引擎以其对Web信息强有力的检索能力成为目前人们从浩如烟海的Internet中获取所需信息的重要途径[1]。然而,即使技术先进如Google和百度这样的通用搜索引擎巨头仍然面临诸如下述的棘手问题尚未解决[2]:

(1) 低查准率:查准率和查全率犹如"矛"和"盾"一样难以协和兼得,通用搜索引擎往往以牺牲查准率来获得较高的查全率,而这种做法对有特定信息需求的人群越来越显得无能为力;

(2) 搜索的"垂直度"问题:如何针对专业领域的行业需求,更精确地、深入地挖掘和获取用户所需信息既是一个难度很高也是一个亟待解决的现实问题;

(3) 可定制性问题:目前已经存在一些垂直搜索引擎,如Google公司的Froogle、Ucloo搜人引擎等,但其所采用的技术与行业应用结合过于紧密,如何快速定制出面向新领域的垂直引擎则需要进一步的探讨。

针对上述问题,本文通过引入本体技术和文本挖掘技术来构造一个面向领域的垂直搜索引擎。首先探讨了本体管理和文本挖掘技术,这是我们进一步研究的技术基础;之后提出了一个结合本体筛选和文本挖掘的垂直搜索引擎构建思想,具体包括基于本体筛选的智能搜索器(Spider/Crawler)、结合文本挖掘的网页信息抽取及分析、分类器和查询处理器的构建等;最后,我们给出了一个基于此思想的原型系统的实现。

1 基础技术的研究

搜索引擎的前身是信息检索(Information Retrieval),主要通过对文本信息进行系统性的操作(索引),以方便快速地从大量文档中通过查询(搜索)获取相关信息,其基本流程包含数据收集、特征选择、模型选择、训练、测试、评估等活动[1]。主要部分可以划分为模型和模式结构、评分函数、优化和搜索算法、数据管理策略等几个部分。信息检索在其发展过程中,先后产生了多种信息资源检索工具,其中基于Web的搜索引擎以其界面友好、使用方便成为目前全球最流行的检索工具,为广大用户快速、准确地查询与获取网上信息创造了便利[3]。

本文的主要特点在于采用本体技术来筛选与领域相关的Web页面,通过文本挖掘技术来对筛选出的Web页面进行结构化数据自动分析和提取。因而,本体和文本挖掘技术是本文研究的基础,下面就此两项技术进行深入探讨。

1.1 本体的作用及其构造

本体被定义为"概念模型的明确的规范说明"[4],可以用来描述某个领域甚至更广范围内的概念以及概念之间的关系,使得这些概念和关系在共享的范围内具有大家共同认可的、明确的、唯一的定义,这样,人机之间以及机器之间就可以进行交流。

尽管本体的研究日趋成熟,但是很少关注本体的实际应用。Riichiro和Mitsuru提出本体的应用可以分为八个层次[5],其中前三个层次只是作为一个公共的受控词典来为它所索引的知识库内容提供信息骨架,而后五个层次由于涉及到内容,因而更多地和人工智能技术联系在一起。在本文的

*)基金项目:国家自然科学基金资助项目(编号60573084)和武器装备预研基金(9140A15050106HK0114)。赫建营 博士研究生,主要研究领域为软件工程、软件测试技术和知识管理;晏海华 硕士,副教授,主要研究领域为软件工程、软件测试技术和面向对象技术;金茂忠 教授,博导,研究方向为软件工程和编译技术;刘 超 教授,博导,CCF高级会员,主要研究领域为软件工程。

研究中，本体的应用也只限于前三个层级，其具体作用包括：

• 用以过滤领域相关 Web 页面的特征描述（详见 2.1 节）；

• 进行网页分析及信息抽取时参照引用的基础（详见 2.2 节）。

对于本体的具体构建，我们采用基于 Web 的双语本体管理系统：WBBOMS[6]，它可被用于类 WordNet 结构的本体的管理和维护，避免了传统系统的难维护性，并适用于大规模的本体的构建。同时，提供语义信息到 Web 本体语言（OWL）的输出，以利于其它本体相关应用（如 Protégé）以程序方式直接利用本体库。借助相关词提取算法，还可以从本体库中获取语义相关的词用于辅助用户进行查询扩展。其体系结构如图 1 所示。

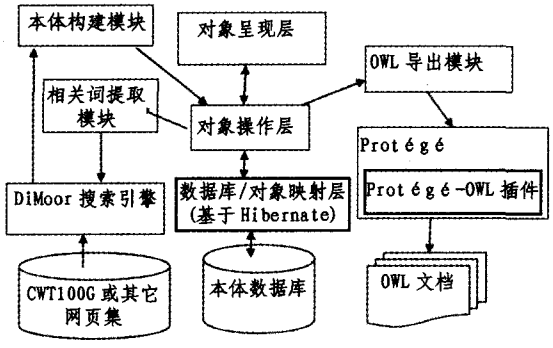

图 1　WBBOMS 的系统结构

1.2　文本挖掘技术及其作用

文本挖掘被用于发现文本中的模式、规则或趋势，目标是从非结构化或半结构化的文本信息中提取需要的（感兴趣的、非平凡的）信息和知识[7]，其挖掘的流程有很多种，一般都是遵循管道（pipeline）模式，即前一个阶段的输出作为后一个阶段的输入[7]。由于本文的侧重点在于使用自然语言处理技术对原始文本切分后进行索引，然后在索引的基础上进一步构造与垂直搜索相关的应用，故遵循图 2 所示的基本流程。它主要由五部分组成：文本提取器、元信息提取器、文本分析器、索引程序和检索分析程序，其中分词器我们采用北航软件所开发的中文分词系统 BUAASEISEG[8]，它在稳定性和新词识别能力上具有一定的优势；索引程序和检索分析程序将在第 2 节详细介绍。

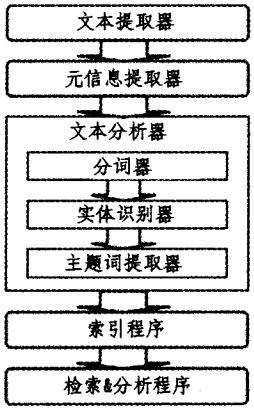

图 2　文本挖掘流程

• 文本提取器：针对不同类型的文档提供不同的解析器（Parser），将文本信息和元信息提取出来；

• 元信息提取器：对网页文本中具有特定含义的标签所标识的信息进行提取；

• 文本分析器：针对由于中英文词法的不同特点，采用相应的分词算法对前一阶段提取出的文本进行分词，并在分词的基础上进行进一步的实体识别和主题词提取[9]。

2　结合本体筛选和文本挖掘的垂直搜索引擎构造

无论是通用搜索引擎还是垂直搜索引擎，其基础体系结构及运行原理大体相同，一般包括搜索器（Spider/Crawler）、索引器（Indexer）和检索器（Searcher）[1]。搜索引擎利用 Spider/Crawler 获取网页，用 Indexer 解析和索引页面，用 Searcher 利用 Web 服务器（Web Server）来响应用户的查询请求进行检索[9]。本文的研究在此基础架构之上，针对垂直搜索引擎“专”、“精”、“深”的应用需求及技术特点，引入本体过滤技术和自然语言处理中的文本挖掘技术，形成如图 3 所示的垂直搜索引擎体系结构，其具体内容详述如后。

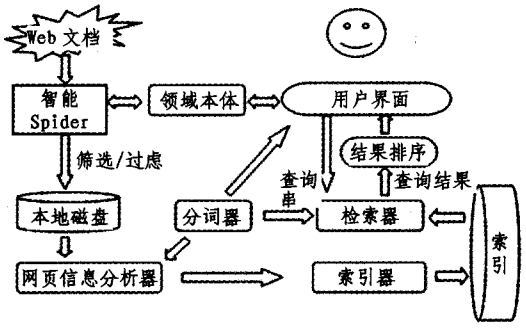

图 3　垂直搜索引擎总体架构图

2.1　基于本体筛选的智能搜索器

垂直搜索面向的是行业应用，如何保证搜索器抓取的页面都是领域相关的页面是一个关键问题。事实上，它是一个页面过滤的问题，也是垂直搜索引擎的核心，我们采用以下方式来从 Internet 上过滤出面向某一具体应用领域的 Web 页面：

（1）使用本体定义出某一领域的具体特征，以 OWL 格式表示；

（2）参照本体定义出的领域特征，在搜索器抓取 Web 页面之前，搜索器结合 2.2 节中的页面分析技术对 Web 页面内容进行初步的过滤，通过滤掉一部分和目标领域相关度低的 URL，从而既减少了下载这些 Web 页面的网络开销，又从源头上保证了搜索结果的领域相关性和准确性；

（3）利用 BUAASEISEG 对网页链接的锚文本进行分词，之后根据分出的关键词查询领域本体库以分析此关键词是否和领域相关，根据分析结果决定网页链接的取舍。

2.2　结合文本挖掘的网页分析

网页分析就是对搜索器抓取的 Web 页面进行抽取特征，同时还将此网页中的超链接提取出来，返回给搜索器进一步深入搜集信息[10]。一方面利用网页分析结果作为搜索器取舍抓取的 URL 的依据，另一方面利用其分析得出的主题词建立起与 URL 的关联以作为下一步索引的基础。其具体过程为：

（1）对经过过滤后的领域 Web 页面进行分词及主题词提取，提取出的主题词被反馈到搜索器中充当页面取舍的依据；

(2) 利用机器学习算法,对含有同类信息的 Web 页面进行学习从而得出同类 Web 页面的信息抽取规则;

(3) 利用(3)中抽取的对其他 Web 页面进行大规模获取,并通过不断的训练和学习提高提取器的精度和成熟度;

(4) 在进行规则抽取的同时提取搜索器已经过滤后的 Web 页面的元数据。

2.3 索引器的构造

在信息检索中,首先要解决的问题是如何快速地通过匹配文本找到对应的文档。在面对大规模 Web 文本的情况下,对每个文本都查询一遍,再快的匹配算法也会消耗大量的时间。因此,预先索引就成为了必由之路[9]。早先的研究是把文本串切分为单词后,建立 Hash 表以此作为预先索引的方法。当然,所有用于数据库索引的算法同样可用作文本索引。由于词汇数量相对文档数量要小,故以词汇为索引的倒排表逐渐成为目前主流的索引方式。倒排表即以词为索引,文档为项建立的一个链表集合。也可以看作是以词为行,文档为列的庞大稀疏矩阵。为了快速查找,其中的词和文档编号列表都是预先排序的。

在本文的研究中,我们不但使用倒排表存放了每个词所对应的文档编号列表,还存储了该词对应的文档的数目、在某文档中的词频,甚至该词在某文档中的位置等,以方便快速获取。这样,词和它所对应的文档数目实质上构造了一个以大规模语料为基础的词频词典。在索引时,为了建立对文档的描述结构以便应用程序使用,我们不仅存储和索引网页内容本身,还按字段存储一些文档相关的描述信息,如标题、URL、站点(Host)、抓取日期等。这样就可以支持按字段查询,如在指定站点内查询就需要利用站点(Host)字段的信息。

2.4 查询处理器及用户界面

查询处理器和用户界面属于用户呈现的层面。就用户界面来说,我们提供两种模式以方便用户使用:普通检索和结构化的检索,这和普通搜索引擎并无差异之处。而在查询处理器上,则引入了分词能力,先通过 BUAASEISEG 对用户输入自然语言文本进行分词,之后再从索引库中检索出匹配的文档,其好处包括:(1) 提高检索速度;(2) 提高搜索的查准率。

3 验证及实现

我们在开源搜索引擎 Nutch[11] 的基础上已经验证了本文所提出的思想和方法。Nutch 是一个典型的开源搜索引擎实现,具有松耦合的体系结构(如图 4 所示),通过替换其中某些模块并不会影响整体的协同,但是能够彻底改变系统的行为。此外,它还具有可扩展的基于类似于 Eclipse 的插件机制的分层体系结构,并实现了最基本的搜索引擎的相关模块,因此是构建新的搜索引擎的一个良好起点。

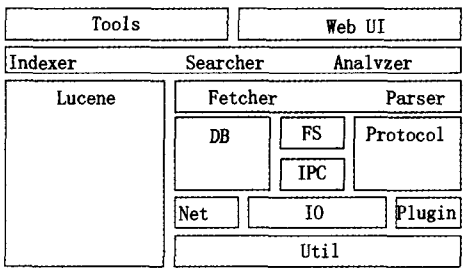

图 4 Nutch 的体系结构

在具体的实现过程中,我们采用北航软件所已有的 WBBOMS 来进行本体的创建和管理,采用 BUAASEISEG 来进行分词,在扩展 Nutch 体系结构的基础上,通过替换和整补模块来构建了一个面向高校毕业生招聘的垂直搜索原型系统(如图 5 所示),此项目得到了教育部的支持和资助。

结论 有数字表明这样一个现实:搜索引擎的行业化、专业化是一个必然的发展趋势。垂直搜索引擎技术区别于通用搜索的核心技术就在于 Web 页面的过滤,亦即:Web 页面的资源构成了垂直搜索引擎进入的门槛。然而观察目前已经出现的行业搜索引擎,它们是在各自领域内长期的积累造就了其 Web 页面资源的专业化,而并未从技术上解决面向整个 Internet 的页面过滤问题。

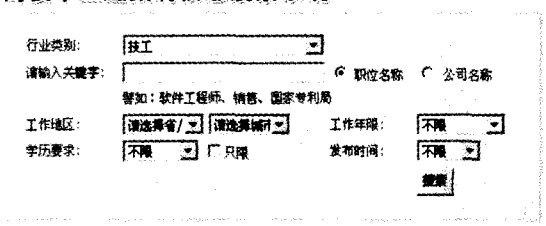

图 5 面向高校毕业生招聘的垂直搜索系统

本文的研究正是针对这个问题及与此相关的主要技术难题,提出一个结合本体筛选和文本挖掘的垂直搜索思想,并以 Nutch 为基础构建了一个面向探索并构建一个面向高校毕业生招聘的垂直搜索原型。事实上,只要修改其特征过滤规则即可构造面向不同行业的垂直搜索引擎。当然,本文构建的垂直搜索引擎还处于原型阶段,其成熟和完善还有不少工作要做,这也是我们下一步努力的方向。

参 考 文 献

1 Baeza-Yates R, Ribeiro-Neto B. Modern Information Retrieval. China Machine process, 2004

2 Hearst M A. Next Generation Web Search: Setting Our Sites. In: Luis Gravano ,ed. IEEE Data Engineering Bulletin, Special issue on Next Generation Web Search, September 2000

3 Broder A. A taxonomy of Web search. In: SIGIR Forum, 2002, 36(2): 3~10

4 Gruber TR. A translation approach to portable ontology specification. Knowledge Acquisition, 1993, 5:199~220

5 MIZOGUCHI R, IKEDA M. Towards Ontology Engineering. The Institute of Scientific and Industrial Research, Osaka University, 1998

6 曹勇刚,曹羽中,金茂忠,刘 超. 基于 Web 的双语本体构建系统. 计算机科学,2005,32(9A):60~63

7 Tan Ah-Hwee. Text mining: The state of the art and the challenges. In: Proceedings, PAKDD'99 Workshop on Knowledge discovery from Advanced Databases (KDAD'99), Beijing, April 1999. 71~76

8 曹勇刚,曹羽中,金茂忠,刘超. 面向信息检索的自适应中文分词系统. 软件学报,2006,17(3):356~363

9 曹勇刚,曹羽中,金茂忠,刘 超. 提取、索引和挖掘中文学术论文. 南京大学学报(自然科学版), 2005,41:845~852

10 Chang C, Hsu C, Lui S. Automatic information extraction from semi-structured Web pages by pattern discovery. Decis. Support Syst, 2003,35(1):129~147

11 Khare R, Cutting D, Sitaker K, Rifkin A. Nutch: A Flexible and Scalable Open-Source Web Search Engine. CommerceNet Labs:[CN-TR-04-04]. November 2004. 1~12