搜索引擎技术及应用研究

李如平
(安徽工商职业学院 电子信息系,安徽 合肥 231100)

【摘 要】搜索引擎通过网页搜索软件查询互联网上的链接,访问网络公开域中的每一个站点,将它们的网址带回搜索引擎,同时给web页上的文本建立索引,从而创建出一个详尽的网络目录。由于网络文档的不断变化,搜索软件也不断地更新以前已经分类的目录。搜索引擎目前广泛应用于网络资源搜索和商业智能系统。

【关键词】搜索引擎;文本;信息
【中图分类号】TP393.09 【文献标识码】A 【文章编号】1673-1891(2010)03-0055-05

1 引言

当前,信息化产业革命正席卷全球,自从20世纪90年代以来,计算机网络技术的广泛应用带来了新经济革命,世界正走进信息社会。搜索引擎经过这些年的发展,在实现互联网信息搜索的同时,也在企业信息平台和商业智能中广泛应用。

2 搜索引擎技术

2.1 搜索引擎简介

搜索引擎(search engine)一般的定义是指根据一定的策略、运用特定的计算机程序搜集信息,在对信息进行组织和处理后,为用户提供检索服务的系统。

通常,搜索引擎分为三个部分:(1)在网上搜寻所有网页信息,并将它们带回服务器,加以处理;(2)将信息按主题组织索引,建立检索web文件的数据库;(3)通过web服务器软件,为用户提供浏览器界面下的信息查询。

用户在搜索引擎用户界面的检索文字框中,键入检索提问式,提交给搜索引擎检索系统,可以方便地从搜索引擎的数据库中得到满足检索要求的网页的网址,再通过浏览器的链接功能,便获取所需网页。搜索引擎使得网络信息检索变得简单,使人们利用网络信息资源成为可能。

2.2 文本信息检索技术

信息检索已发展多年,其中以文本为对象的文本信息检索是目前检索最主要的部分,它以非结构或半结构化数据为处理对象,研究大量文本的信息组织和检索问题。文本信息检索主要发现与用户检索要求(关键词)相关的文本。信息检索的主要对象是文档资源,文本是文档的一种典型形式,用户可以通过自然语言或者关键词表达检索需求,用户提交的检索需求称为一个查询。即计算机程序检索每一篇文档中的每一个词,建立以词为单位的倒排文件,检索程序根据检索词在每一篇文章中出现的频率和概率,对包含这些检索词的文章进行排序,最后输出排序的结果。

在检索开始以前,必须要先有一个定义好的文本数据库,它的功能为:(1)存放文档内容;(2)在该文档上可以进行操作;(3)文本模型。文本数据库中存放的信息是通过人工方式或网络爬虫方式采集到的,它可以通过数据管理模块对文档进行索引的建立。索引是关键的数据结构,它支持在大量数据中进行快速查找。在建好索引后,就可以进行检索了。用户首先详细说明用户需求,这个过程是个迭代的过程,用户会不断地对用户需求进行修正表达,系统也通过关联提示帮助用户对表达式进行描述。当用户表达式确定后,系统为用户的需求提供一个系统表达式。接着通过处理查询获得检出文档。在把文档送出前,将根据检出文档的类别以及相关度对检出文档进行排序。随后,用户查看经过排序的文档集合,查找有用的信息。

随着网络技术的普及,文本信息检索技术逐渐应用于网络信息检索常用的Web搜索引擎中,根据用户输入的查询条件检索出各类Web站点上的相关信息,并将搜索结果以条目的形式呈现给用户,当用户需要查看某条信息的详细内容时,只需单击该条目,系统便链接到相应的站点,将全文呈现给用户。搜索引擎本身就起源于传统的信息检索理论,文本信息检索技术已成为搜索引擎的核心支撑技术。

2.3 企业级搜索

企业级搜索引擎,它和一般的网络搜索引擎最大的不同在于它的信息主要来自于企业内部。企业搜索是个人搜索在企业内的延伸,能够在企业内部获取工作所需的各种最新最全面的信息,以便更好地为企业带来效益。企业用户对信息的需求不

收稿日期:2010-05-14
作者简介:李如平(1973-),男,安徽肥东人,硕士,讲师,主要研究方向:计算机应用技术、信息管理。

仅仅限于简单的查询结果，而是结合搜索、数据库查询、语义和句法分析、分类和聚类、相关性分析等技术，整合现有的信息资源，提炼出具有商业价值和社会价值的数据[1]。

企业搜索的目标具有很强的多样性和离散性。这是因为网页只是目标资料库中很小的一部分，企业各种有价值的信息大多存放在数据库或是Word、Excel、PDF等非结构化电子文档中。而且，这些资料的存放位置也不是惟一的，它们很可能分布在不同地点、不同操作系统的计算机里。另外，企业搜索对安全性的要求也比个人搜索严格得多。

在商业环境下，如何判断某一信息是否对企业工作有帮助，最终找到所有有用的信息，是企业搜索中的重要问题。如何将搜索到的内容整合为实际可用的信息，进而实现其商业价值，是企业搜索的目标所在。现在越来越多的公司开始重视企业搜索引擎的开发。例如IBM公司就有专门企业搜索引擎产品，为其他公司提供搭建内部搜索引擎的服务。可见企业搜索不仅是搜索公司关注的重点，同时也越来越受到传统IT企业的青睐。

3 Oracle全文本搜索技术

3.1 Oracle Text搜索技术简介

目前的搜索引擎普遍采用Oracle text作为核心开发工具。Oracle数据库服务器的文本管理组件提供一个文本搜索系统，可以使文本在Oracle数据库中与其它数据一样被快捷地搜索和管理。Oracle Text提供高级的文本搜索方法，将文本作为Oracle Server中的一种标准的数据类型。使用Oracle和Oracle Text，可以方便而有效地利用标准的SQL工具来构建基于文本的新的开发工具或对现有应用程序进行扩展[2]。

我们可以充分利用Oracle Text技术对使用文本的Oracle数据库应用程序进行搜索，搜索范围可以是现有应用系统中可搜索的注释字段，也可是实现涉及多种文档格式和复杂搜索标准的大型文档管理系统。Oracle Text支持Oracle数据库所支持的大多数语言的基本全文搜索功能。

3.2 Oracle Text的逻辑结构[3]

Oracle Text索引文档时所使用的主要逻辑步骤如下：

(1)数据存储逻辑：搜索信息存储表中的所有行，并读取列中的数据。

(2)过滤器：提取文档数据并将其转换为文本表示方式。

(3)分段器：提取过滤器的输出信息，并将其转换为纯文本。

(4)词法分析器：提取分段器中的纯文本，并将其拆分为不连续的标记。

(5)索引引擎：提取词法分析器中的所有标记、文档段在分段器中的偏移量以及被称为非索引字的低信息含量字列表，并建立反向索引。

Oracle Text应用实际上就是一个数据装载→索引数据→执行检索的一个过程。一个搜索引擎最重要的部分就是索引和检索。下面介绍Oracle text的索引和检索的原理。

3.3 Oracle text的索引

索引是搜索引擎最重要的组成部分，一般的搜索引擎都是采用倒排索引的方式，Oracle text同样也是采用倒排索引的方式。建立Oracle索引包括数据库中最基本的索引类型，索引的建立，索引的维护和优化等。

(1)索引类型

建立的Oracle Text索引被称为域索引(domain index)，包括4种索引类型：CONTEXT、CTXCAT、CTXRULE、CTXXPATH。依据应用程序和文本数据类型可以任意选择一种。可以利用Create Index建立这4种索引。在4种索引中，最常用的就是CONTEXT索引。

(2)索引的建立

在数据库中，用创建和插入这些索引的方法叫做索引管道。建立索引时，系统默认文档存储在数据库的文本列中。如果不显示指定索引参数，系统会自动探测文本语言、数据类型和文档格式。

(3)索引的维护

索引的维护包括查看是否有错误索引，删除索引，重建索引，索引的同步与优化等。

3.4 Oracle text的查询

查询是搜索引擎最关键的部分之一，搜索引擎性能的好坏一部分决定于查询的方式和算法。统一搜索引擎就采用了Oracle提供的查询API来实现检索功能，并有较高的效率。索引建好后，就可以通过SELECT语句中的CONTAINS运算符进行文本查询。通过CONTAINS可进行词查询和ABOUT查询。

(1)词查询

词查询是对输入到CONTAINS运算符中单引号间的精确单词或短语的查询。在查询表达式中，可以使用AND或OR等运算符来查询出不同的结果，也可以将结构性谓词添加到WHERE子句中。

(2)ABOUT查询

在所有语言中,ABOUT查询增加了某查询所返回的相关文档的数目。在英语中,ABOUT查询可以使用索引的主题词组件,该组件在默认情况下创建。这样,运算符将根据查询的概念返回文档,而不是仅依据所指定的精确单词或短语。使用Oracle Text查询应用程序后,用户可得到查询所返回的文档列表,用户选择一个文档,然后应用程序以某种形式显示该文档。通过Oracle Text,可以用不同的方式再现文档。例如,可以通过突出显示查询词来显示文档。突出显示的查询词可以是相关词查询中的词,也可以是英文ABOUT查询中的主题词。

在统一搜索系统中,采用HTML版本来高亮显示搜索结果中的关键词。

3.5 Oracle text的优点

选择Oracle text全文检索技术作为统一搜索引擎的核心,是因为它有以下的优点:

(1)简单、容易使用。

(2)性能好、检索速度快。

(3)集成于Oracle,功能更强大。

(4)能支持不同格式的文档。

目前,真正意义上的搜索引擎,基本都是对被检索对象的所有文字建立索引,然后再对这些索引进行检索。在这方面Oracle Text和它们基本相同,但是当用于对数据库的检索时,Oracle Text的优势就体现出来了,因为它是完全集成在Oracle数据库中的,所以文本索引的创建和管理变得更加容易,性能更好,并可通过SQL查询实现无缝搜索;Oracle还增加了许多额外的服务,使用户可以根据搜索条件,更方便快捷的访问数据库。当然Oracle Text也有一些局限性,比如:日期数字嵌套式列表以及对象列不能索引;不支持复合索引,而只能对其中的一列进行索引。

4 中文分词技术在搜索引擎中的应用

4.1 中文分词技术

中文分词是计算机人工智能技术的一种体现。众所周知实现计算机模拟人的理解方式需要有相应的算法提供支持。现有的中文分词方法通常有以下三类:基于词表的分词方法、基于统计的分词方法、基于规则和基于统计相结合。

(1)基于词表的分词方法[4]

这是一种有着广泛应用的机械分词方法,该方法依据一个分词词表和一个基本的切分评估原则,即"长词优先"原则,来进行分词。

(2)基于统计的分词方法

从语言学看,词是字的稳定组合,在上下文关系中,相邻的几个字在文章中出现的次数越多,组成词的可能性就越大。首先切分出与词表匹配的所有可能的词,然后对语料中相邻的各个字的组合的频度进行统计,计算这些组合的概率,并与设定的值相比较,超过一定的值范围,便可认为此字组合可能构成了一个词,这种方法称为统计分词法。

(3)基于规则和基于统计相结合

这种方法先运用最大匹配作为一种初步切分,再对切分的边界处进行歧义探测,发现歧义。再运用统计和规则结合的方法来判别正确的切分,运用不同的规则解决人名、地名、机构名识别,运用词法结构规则来生成词。

另外,还有一些其他分词方法,例如基于标记的分词算法、基于神经网络的分词方法等。对于上述提到的各种分词方法,按照目前各种分词系统的应用各有优势,目前暂时还无法判断哪一种分词方法准确度最高,对于中文词的识别,需要运用多种方法来综合处理不同的问题。

4.2 中文分词应用难点

从上个世纪九十年代开始,许多专家、学者、研究者都致力于中文语言的自动分词研究,目前已经取得了很多可用的分词技术,然而在实际应用的过程中,又遇到不少新问题。例如人名、地名、机构等未登录词以及歧义词的产生,于是中文分词技术当今仍存在以下几个难点有待进一步改善。

(1)中文词的概念限定

中文分词的最为首要困难是词的概念模糊不清。按照中文书写的习惯词与词之间无间隔很难在句子中区分一个词[5]。

(2)歧义词消除

歧义是指同样的一个句子,可能会出现两种或者更多的切分方法或者说是理解方法。例如:"市长春节致辞",因为"市长"和"长春"、"春节"都是词,那么这个短语就可以分成"市长/春节/致辞"和"市/长春/节/致辞"。这种称为交叉歧义。像这种交叉歧义十分常见,例如:"表面的"可以分成"表面/的"或者"表/面的"。交叉歧义和组合歧义相比还算是比较容易处理的,而组合歧义就必需结合整个句子来识别。组合歧义很难从词的层面上识别,需要根据上下文关系去理解。对于歧义词的消除目前也产生了一些研究成果,包括基于规则的歧义词消除方法是目前许多研究者采用的方法。

(3)未登录词识别

未登录词,主要是指中文分词的词典中未出现

的词，其大致包含两大类：①新词，即网络上流传的通用词或新的专业术语词等；②专有名词，如中国人名、外国译名、中文地名和机构名等。针对这些专有名词，可以收录专有名词建立资料库，并根据现有的资源统计出各姓氏、人名、地名用词出现的概率，在未登录词出现的句子中再以动态规划的方法求出可能最佳的专有名词。但是收集规模的语料库来支持该方法却有很大的困难。目前无须大语料库支持的未登录词识别解决方案主要有以下几种：利用上下文的限制成分识别未登录词；两趟分词，在“分词碎片”中计算单字成词概率和未登录词概率；有穷多层列举法，通过建立单字词和多字词表（不包括双字词），结合特殊的切分方法实现；通过特征词不断细分字串，随后逐步压缩含有未登录词的子串直至不能从中切分出已登录词。此外还有标记规则、决策树、分解与动态规划等方法。

（4）分词与理解不一致

即便是创建一个收集所有中文词汇的词典来完成实现匹配分词，计算机分词仍然面临知识短缺的大问题。从人的自身体验来看，人在阅读中文文章时，通常情况下是先理解后分词，或至少是边理解边分词。有时也会出现无法一次性断词的情况，则需要反复阅读。而计算机无法有人的理解能力，因此在词典充足的条件下也会出现分词不准确的现象。

5 搜索引擎的应用

5.1 搜索引擎在网络资源检索中的应用

搜索引擎通过软件根据深度优先和广度优先算法在网络上不断搜索相关网页来建立、维护、更新索引数据库，并根据检索规则和数据类型对数据进行加工处理，为用户提供基于布尔逻辑算符的检索方式，并通过检索接口为用户提供信息检索服务，根据用户的请求返回相应的结果。主要有以下几个方面的应用。

（1）网络浏览器

①检索查询功能：读入HTML文档，解释HTML所描述的图表、声音、动画、表格，一件进一步的链接信息，利用超文本传输协议（HTTP），可在任意WWW服务器上畅游。

②文件服务功能：能在下载文件时实时查阅该文件，并可利用HTTP去跟踪感兴趣的链接，可对正在查阅的文件随时保存、打印、前后浏览等

（2）多元搜索引擎

多元搜索引擎就是指在同意用户界面与信息反馈的形式下，共享多个搜索引擎的资源库，为用户提供信息服务的系统。

综合搜索引擎与搜索引擎最大不同在于它没有自己的资源库和软件。由于该引擎可以不考虑索引数据库的构建和维护，故而可以将精力放在查询上，设计出更加简明、方便的查询界面，提供更强的查询功能。

（3）高级智能代理器

智能代理器能够协助用户寻找、消化所需的网络信息，逐渐实现由“人找信息”过渡到“信息找人”的境界。代理的原意是一个人代表别人完成某件事情，在网罗信息资源检索领域里，它相当于用户的信息秘书，通过向用户和其他代理学习，逐渐了解用户的兴趣和爱好，不断为用户提供新的信息，提高了查询效率。目前智能信息检索已成为计算机科学和情报科学研究的热点。

5.2 搜索引擎在商业智能中的应用

商业智能（BI）是一种涉及公司客户、竞争对手、合作伙伴、竞争环境和企业内部业务的知识，可以通过它制定出有效的、重大的、通常是战略层面的企业决策，商业智能系统包括一些IT应用系统和工具[6]。商业智能的核心使命是帮助企业经理人员做出及时、正确、可行、有效的决定，从而改善企业经营效果，提高企业竞争力和获利性[7]。

商业智能中心系统中的搜索引擎，就是专为查询企业商业智能或者其他内部信息而出现的查询工具，对解决行业领域内的专业信息查询，这种企业级搜索引擎要比通用搜索引擎有效得多。

随着企业信息化的发展和历史原因，大多数企业同时运行着许多不同的管理信息系统和商务智能系统，众多的系统，对管理和使用公司资源造成了一定的困扰。如果一个经理不清楚每个系统的具体功能，他想得到一个产品一年销售报表，就需要经过多次周转，才能确定哪个系统的哪个报表才是他想要的。为了提高工作效率，整合公司资源，也更是为了探索商业智能在企业的应用，越来越多的企业着手建设企业“商业智能中心”。作为一个商业智能的探索项目，定位于一个利用数据仓储与数据挖掘技术，创造和累计商务知识和见解，改善商务决策水平，采取有效的商务行动，完善各种商务流程，提升各方面商务绩效，增强企业综合竞争力的面向于未来的系统。

商业智能中心将整合公司的多个商务智能系统，提供了统一的企业登录门户，实现了单点登录和统一搜索，提供“一站式”服务。商业智能中心应具有如下的功能：

(1)统一的登录门户

商业智能中心整合了多个子系统,提供了统一的登录门户平台,登录门户是重要的企业信息应用集成基础框架,实现企业信息应用的整合、集成、增值,帮助人们在获取特定的数据时不用再进入众多的应用系统,而是由门户就可以方便快捷的获取信息,人们能够快速地调用各种不同的后台应用,并完成对后台应用的各种操作。

(2)纯Web化的操作环境

客户端不需要安装任何程序,透过浏览器,不管身在何处只要透过网络即可执行。

(3)统一搜索引擎

一个基于Oracle text技术的全文本搜索引擎。它可以像google和百度一样满足用户现实和潜在的各类信息查询需求,提供信息导航作用。与一般的搜索引擎相比,它的特点是数据来源更多样化,对数据安全要求更高,更有针对性。

6 总结

搜索引擎已成为目前互联网应用的一个重要方面,随着网络技术和信息化水平的不断提高,人们对搜索引擎的要求也越来越高,希望能够更快更好地搜索到需要的信息。搜索引擎在互联网应用方面大力发展的同时,在企业信息平台及商业智能方面的企业级搜索将会有广阔的发展空间。

注释及参考文献:

[1]夏高强.论电子商务中的搜索引擎[J].新西部,2008(3):214-215.

[2]龚谷初.Oracle Text全文检索技术在信息管理中的应用[J].湖南电力,2004,24(6):23-24.

[3]陈天伟.基于Oracle Text电子政务全文检索技术的应用[J].办公自动化,2007(1):2-4.

[4]朱明.数据挖掘(第2版)[M].合肥:中国科学技术大学出版社,2008:331.

[5]温滔.自适应歧义切分的汉语分词系统的设计与实现[D].苏州大学,2005.

[6]斯蒂芬.哈格,梅芙.卡明斯,埃米.菲利普斯.信息时代的管理信息系统[M].严建援等译,第6版.北京:机械工业出版社,2007:67-168.

[7]赵红宇.基于DW、OLAP、DM的商业智能[J].商场现代化,2006(26):52-53.

Research of Search Engine Technology and Application

LI Ru-ping

(Department of Electronic and Information, Anhui Business Vocational College, Hefei, Anhui 231100)

Abstract: Search engine queries internet links through a web search software, and accesses each web site in public domain network, and then backs its address. At the same time search engine creates index for the text on the web page and creates a detailed network directory. As the the network documents are changed usually, the search software always update directory which has been classified previously. At present search engine is widely used in network resources search and business intelligence system.

Key words: Search engine; Text; Information