基于概念的中文搜索引擎技术*)

张秋余 张 红 马彦宏

(兰州理工大学电气工程与信息工程学院 兰州 730050)

摘 要 本文提出了一种基于概念的中文搜索引擎,给出了它的理论模型和软件设计过程,其核心是应用知识库中的语义、语法、词法等知识,提高了切句、分词的准确性,把词从本意上升到概念层次,并采用一种新的HTML标记权值统计算法提高了词的索引度,从而从根本上提高了搜索引擎的查全率与查准率。

关键词 概念,知识库,搜索引擎,权值统计

Concept-Based Chinese Search Engine Technology

ZHANG Qiu-Yu ZHANG Hong MA Yan-Hong

(College of Electrical and Information Engineering, Lanzhou Univ. of Tech., Lanzhou 730050)

Abstract This paper puts forward a new kind of concept-based Chinese search engine and gives its theory model and software design procedure. Its kernel is to use phraseological, semantic and lexical knowledge in the knowledge database to improve exactitude of cutting sentences and words and make words meanings raised to concept layer. And it uses a new weighting algorithm on calculating weight of HTML tags. Thus it has raised index degree of words. So it improves the recall ratio and precision ratio of search engine.

Keywords Concept, Knowledge base, Search engine, Calculate weight

1 引言

互联网的迅速发展和广泛普及带动着搜索引擎的更新换代,但是,大多数的搜索引擎都是基于关键词的,它不能区分同形异义也不能联想到关键字的同义词。搜索引擎已成为一个新的研究、开发领域,尤其是中文搜索引擎,这主要是由于中文本身的复杂性及其语义的多变性。本文就此提出了一种基于概念的智能中文搜索引擎[1],它的核心是借助于知识库[2]对用户请求或网页文件进行词法、句法、语义分析去掉禁用词得到能准确表达用户输入或网页的信息概念集,并应用新的HTML标记加权算法,同时增加了同义词扩展,提高了搜索引擎的智能化与查全率和查准率。

2 基于概念搜索引擎的理论模型

该搜索引擎从理论上来说也是三部分:搜索软件、索引软件和检索软件,如图1所示。

(1)搜索软件

网络搜索软件通常称为Web"蜘蛛"(Spider)、"爬虫"(Crawler)或"机器人"(Robots)。它以一个初始的URL列表为起点,利用标准协议遍历WWW空间,包括Web页面里的所有链接(link),进行网页信息采集,并将其存储在网页数据库中,以备索引模块进行标引处理,提取概念。这里的初始URL列表可以由网络用户通过一个特定格式主动向搜索引擎提交注册,也可由搜索引擎自身制定一定采集策略来确定,该系统同时采用这两种采集方法。

(2)索引软件

索引软件主要是用于对网络搜索软件采集到的网页信息借助于知识库进行词法及句法分析,并提取概念进行网页标引,矢量化后存储在索引数据库中,建立可供检索的Web索引数据库。知识库的出现使这一切都成为可能,因为知识库中存储了大量的词法、句法知识,语义、语用知识,常识,语料库,词典库,禁用词表,反向词频统计表等等,随着时间的推移,知识库还会更加强大与丰富,那么搜索引擎将真正达到智能化、准确化。通常的索引软件主要通过从网页中自动提取能表达网页主题意义的分类或特征信息作为标引词来构建网页标引记录,如网页标题、网址、链接、人名、机构名、地名和网页前面若干个词等。本系统抽词的依据主要有词频、按照一定算法计算出的权重以及词语在页面中出现的位置及类型,依据知识库的禁用词表取道分词结果中的无意

*)基金项目:甘肃省科技攻关项目(GS021-A52-54)。张秋余 副研究员,主要研究方向为数据仓库与数据挖掘、网络与信息系统,软件总线。张 红 硕士研究生,主要研究方向为数据挖掘、网络信息检索。马彦宏 硕士研究生,主要研究方向为人工智能,智能代理,嵌入式系统。

义词，然后让该词序列的权重乘以知识库中反向词频表中的各词的权重得到其最终的权重，再把它们索引后以矢量表示存储在索引数据库中，以备检索。

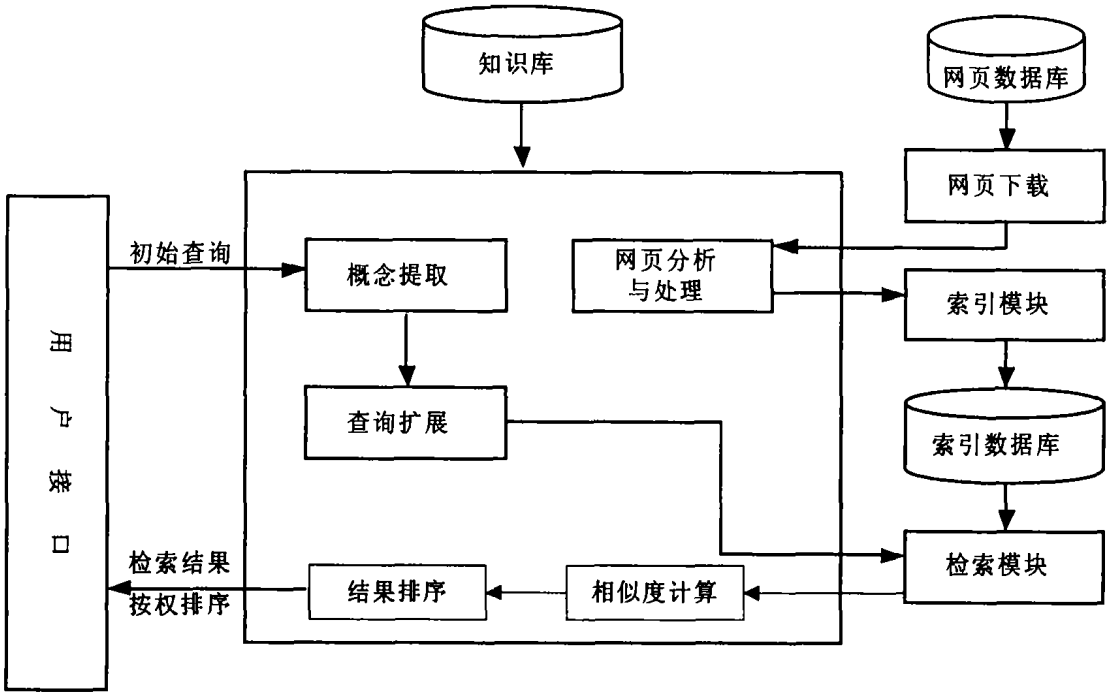

图 1　基于概念智能搜索引擎的理论模型

(3)检索软件

检索软件主要从索引数据库中检索满足用户要求的 Web 网页。该系统中首先对用户输入利用知识库中的词法及语法分析，借助于反向词典及正向词典进行切句切词，对照知识库中的禁用词表去掉概念信息不明确的词，如一些助词，再进行同义词扩展，这样得到用户输入的概念信息，然后从索引数据库中去检索，将返回结果进行相似度排序后返回用户界面。

3　系统设计[3]

该系统采用了 Java——一个广泛使用的网络编程语言作为开发工具，各个部分的设计如下：

(1)搜索器(Robots)[4]

搜索器用了 Java 中的 BOT 包，主要实现该功能的类有 Spider 类、SpiderInternalWorkload 类、SpiderWorker 类、SpiderDone 类，SpiderSQLWorkload 类，主要的接口为：SpiderReportable 接口、IWorkloadStorable 接口。

由于硬件限制，调试时并没有把搜索的网页存储在网页数据库中而只是存储了网页的 URL 地址，这样不仅大大节省了空间，又使得在学校有限资源条件下开发该项目成为可能。

(2)索引器(Indexer)

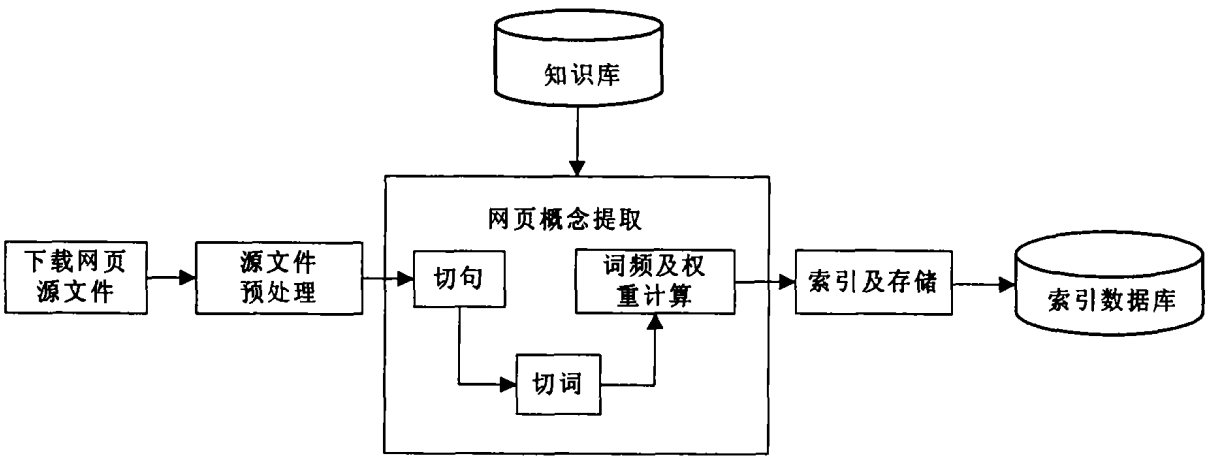

图 2　索引器设计结构图

索引器的设计过程如图 2 所示，具体细节如下：

①得到网页源文件　它主要用到了 Java 中的 net 包，从网页库中取出 URLs，对每一个 URLs 创建 URL，URLConnecion 及 HttpURLConnection 对象，通过 URLConnection 的 connect 方法连接到服务器找到网页源文件，并将网页源文件读入缓冲区等待下一步处理。

②预处理模块　由于该系统要用到网页中的标记，好多网页又不标准，有的是大写，有的是小写，而 JAVA 采用 unicode 编码并区分大小写，因此在进行网页处理之前首先把网页源文件中的标记均变为大写，并去掉标记中的空格（如〈b〉和〈b〉），以便于查找与处理，它主要用到了 String 及 Stringbuffer 对象。

③语句分割模块　语句切分是将预处理后的 HTML 源码文档根据标记类型及中文标点符号找出中文字串（不含中文标点符号）。其中 Latin 字符和中文标点在 Unicode 布局中位于以下几个范围（十六进制表示）：

Latin 字符：0000…007F；

Latin 扩展字符：0080…024F；

中文标点：3000…303F 和 FF00…FFFF

所以用下面的逻辑关系就可将中文字串找出，chineseInt ＞ 0x024F&&! ((chineseInt ＞ ＝ 0x3000&&chineseInt＜＝0x303f) ‖ (chineseInt＞＝0xff00&&chineseInt＜＝0xffff) ‖ (chineseInt＝0x2000&&chineseInt＜＝0x206F))，根据知识库中的词法、句法及语义知识确定该字串是否保留，如“today. getYear()，“年”，”经过切句后只得到“年”，对这样无意义的字串直接去掉，这样就得到一串串的能够表达网页含义的中文字串，并将它存储在链表 Linkedlist 中。

③切词模块　对于 Linkedlist 中的每一串中文字串，依据知识库中的正向词典和反向词典，采用双向匹配算法进行切词，结果保存在链表 LkL1 和 LkL2 中，对照知识库中的禁用词表，去掉 LkL1 和 LkL2 种的禁用词，如“的、是、是的”等等。经过对大量网页文档的验证，对结果的取舍采用以下的算法：判断 LkL1 和 LkL2 的长度。若相等，则取反向切词结果；若不等，则取长度较小的为结果。

④词频及权值统计　权值计算时该系统采用的不同于其他搜索引擎的加权策略[5]如表 1 所示，经过多次试验，具有一定的准确性。

该系统中索引词权重统计既应用了最有影响的索引词加权公式 tf×idf，又考虑了词在具体网页的中的位置及类型属性，也就是要结合表 1 进行计算，具体计算如下：

词 k_i 在文档 d_j 中的权重是：$tf^w_{i,j}=\frac{freq^w_{i,j}}{\max_i freq^w_{i,j}}$，式中 $freq^w_{i,j}$ 为词 k_i 在文档 d_j 中出现的加权频率，$\max_i freq^w_{i,j}$ 为文档 d_j 中所有关键词的最高加权频率，$freq^w_{i,j}=\sum_{k=1}^{m} w_k \cdot f_k$，$w_k$ 表示 HTML 标记权重，其值取表 1 中的数据，f_k 表示词出现 k 次时的加权系统数，取 $k=1$ 时 $f_k=1$，$k>1$ 时 $f_k=1.5$。$tf^w_{i,j}$ 只是词 k_i 在单个文当中的权重，没有考虑在文档间的关系，为了能更准确地表达该文档 $tf^w_{i,j}$ 再乘以 w_i，得到最终的权重 $tf^{w*}_{i,j}$。w_j 是词 k_i 在知识库反向词频表中的权值，它是对大量文档统计的结果，计算如下：

表 1　HTML 标记加权一览表

HTML 标记	权值
〈TITLE〉…〈/TIT〉	8
〈H1〉…〈/H1〉	6
〈H2〉…〈/H2〉	5
〈H3〉…〈/H3〉	4
〈H4〉…〈/H4〉	3
〈H5〉…〈/H5〉	2
〈B〉…〈/B〉	3
〈EM〉…〈/EM〉	3
〈I〉…〈/I〉	2
〈IMG alt＝…	2
Others	1

设整个文献集合中共有 N 个文档且 n_i 是索引词 k_i 出现的文档数，$freq_{i,j}$ 为词 k_i 在文档 d_j 中出现的频率，则有

$$tf_{i,j}=\frac{freq_{i,j}}{\max_i freq_{i,j}}$$

式中 $\max_i freq_{i,j}$ 为文档 d_j 中所有关键词的最高频率。

$idf_i=\log\frac{N}{n_i}$，则 k_i 在文档 d_j 中的权重为：

$$w_{i,j}=tf_{i,j}\times idf_i \quad w_j=\sum_{j=1}^{N} w_{i,j}。$$

⑤存储模块　经过上述几步后得到词的最终权重，并将分词结果进行矢量化后保存在索引数据库中，存储格式为：($word_1$ $freq^{w*}_{1,j}$ $word_2$ $freq^{w*}_{2,j}$ …… $word_n$ $freq^{w*}_{n,j}$))

(3)检索器(Searcher)

检索器用 JSP 设计，界面友好，检索法方法灵活，还提供查询修正，及相关度反馈技术，具体的就是反馈和用户输入词相关的网页中权值最高的概念

集。其设计过程是对用户输入切句切词(即调用搜引器设计过程中的语句分割模块和切词模块),再对切词结果依据知识库进行同义词扩展,得到一组用向量表示的概念集,该概念集与索引库中的网页向量进行向量内积运算得到相似度,以相似度大小的顺序排序输出给用户界面。

结论 本文给出了基于概念的智能搜索引擎的理论模型,并从编程与算法角度详细介绍了决定搜索引擎好坏的索引器的建立过程,整个设计中借助了知识库的各种知识,把词从表面含义提升到概念层次,提高了搜索引擎的查全率与查准率。但是该知识库有待进一步扩展与提高,这还要依靠人工智能、推理机、自然语言理解等诸多科技研究的结果,待知识库内容逐渐丰富壮大后,不仅能够提高查全率和查准率,还能从真正意义上使搜索引擎达到智能化,提高网络利用率。

参 考 文 献

1 何绍义. 概念信息检索的理论与实践[J]. 情报学报,1995,14(2)

2 袁占亭,张爱民,张秋余. 基于概念的 Web 信息检索[J]. 计算机工程与应用,2003,39(23)

3 (美)Ayers D 等,著. 王辉,等译. Java 数据编程指南[M]. 北京:电子工业出版社,2002

4 (美)Heaton J 著. 童兆丰,等译. 网络机器人 Java 编程指南[M]. 北京:电子工业出版社,2002

5 (加拿大)Patterson L 著,徐征,等译. HTML 4 编程指南[M]. 杭州:浙江科学技术出版社,1999

(上接第 2 页)

其中,$P_1,P_2,\cdots,P_5$属于新闻类网页,因此更新速度较快,$P_6,P_7,\cdots,P_{10}$属于天气类网页,因此更新较慢。

若 Crawler 以 1/6(次/小时)统一更新所有网页,即 $f_0=1/6$(次/小时),$f'=1/3$(次/小时),根据分类更新方法,将网页分为两类 $F_1=\{P_1,P_2,P_3,P_4,P_5\}$,$F_2=\{P_6,P_7,P_8,P_9,P_{10}\}$,以 $f=1$(次/小时)访问 F1,同时以 $f_0=1/6$(次/小时)统一更新所有网页。

下面通过计算更新度来比较分类更新方法和统一更新方法的有效性。个体更新方法在此不做考虑,上文已说明统一更新方法优于个体更新方法。在网络容量巨大的情况下,系统无法实现个体更新方法。

各网页在不同更新方法下所获得的更新度如表 2 所示。

表 2 各网页在不同更新方法下所获得的更新度

网页	统一更新方法	分类更新方法
P_1	$1\times(1/6)$	$1\times(1/2)+1\times(1/6)$
P_2	$(1/0.8)\times(1/6)$	$(1/0.8)\times(1/2)+(1/0.8)\times(1/6)$
P_3	$(1/0.9)\times(1/6)$	$(1/0.9)\times(1/2)+(1/0.9)\times(1/6)$
P_4	$(1/1.1)\times(1/6)$	$(1/1.1)\times(1/2)+(1/1.1)\times(1/6)$
P_5	$(1/1.2)\times(1/6)$	$(1/1.2)\times(1/2)+(1/1.2)\times(1/6)$
P_6	$(1/12)\times(1/6)$	$(1/12)\times(1/6)$
P_7	$(1/10)\times(1/6)$	$(1/10)\times(1/6)$
P_8	$(1/9)\times(1/6)$	$(1/9)\times(1/6)$
P_9	$(1/11)\times(1/6)$	$(1/11)\times(1/6)$
P_{10}	$(1/8)\times(1/6)$	$(1/8)\times(1/6)$

由更新度计算公式(1)可得:

统一更新法所得更新度 $U_1=U_1(P_1)+U_1(P_2)+\cdots+U_1(P_{10})=0.7875$

分类更新法所得更新度 $U_2=U_2(P_1)+U_2(P_2)+\cdots+U_2(P_{10})=3.3393>0.7875$

可知 $U_2>U_1$,因此,分类更新方法能获得比统一更新方法更大的更新度。

结束语 分类更新方法结合前两种方法的优点,不会将系统资源耗费在过度更新改变过于频繁的网页上,也不会过多访问改变缓慢的网页,而是均衡的分配系统资源。实际上,分类更新方法的基本思想可以继续扩展,在实际应用中,网络容量日益膨胀,网页改变速度各不相同,统一更新方法不能适应用户对信息更新度的要求,因此,可以根据网页的改变速度把网络化分为不同子集,再根据各子集的改变频率来更新网页集合,这也可以满足用户对某些网页集合的特殊要求,如实时更新。

参 考 文 献

1 宋聚平,王永成,尹中航,滕伟. 面向主题的网页搜索系统. 上海交通大学学报,2003,37(3):401~403

2 张领,叶允明,等. 一种高性能、分布式 WEB CRAWLER 的设计与实现. 上海交通大学学报,2004,38(1):59~61

3 Cho J, Garcia-Molina H. Synchronizing a database to Improve Freshness. In: Proc. of 2000 ACM Intl. Conf. on Management of Data (SIGMOD) Conf. Dallas, Texas, United States. 2000. 117~128

4 Arasu A, Cho J, Garcia-Molina H, Paepcke A, et al. Searching the Web. ACM Trans. on Internet Technology, 2001, 1(1): 2~43