doi:10.3969/j.issn.1007-855x.2010.04.016

基于 NLP技术和相似度计算的智能搜索引擎研究

梁晓诚[1],岳晓光[1],麦范金[1],赵子强[2],路 英[3],王 挺[4,5]
(1.桂林理工大学 信息科学与工程学院,广西 桂林 541004;2.太原科技大学 机械电子工程学院,山西 太原 030024;
3.莫纳什大学 管理学院,澳大利亚 维多利亚州 墨尔本 3800;4.利物浦大学 计算机科学系,英国 利物浦 L69 7ZF;
5.西交利物浦大学 计算机科学与软件工程系,江苏 苏州 215123)

摘要:针对传统的搜索引擎对于自然语言理解方面存在的问题,文章研究了一种新的基于自然语言处理技术和相似度计算的智能搜索引擎的模型.其核心技术是基于自然语言处理的中文分词技术、语义相似度和对立度等理论,将这些概念理论结合起来,从用户习惯的思考角度,结合DotLucene开源全文搜索引擎建立一个智能的搜索引擎.研究表明,该模型在对已经收录的文档有着86.1%的查准率.该智能搜索引擎较好的对查询语句的实现了理解,能够对用户的提问做出正确的回答.

关键词:自然语言处理;中文分词;相似度;DotLucene;智能搜索引擎
中图分类号:TP39 文献标识码:A 文章编号:1007-855X(2010)04-0076-04

Research on Intelligent Search Engine Based on NLP Technology and Similarity Calculation

LIANG Xiao-cheng[1], YUE Xiao-guang[1], MAI Fan-jin[1],
ZHAO Zi-qiang[2], LU Ying[3], WANG Ting[4,5]
(1. School of Information Science and Engineering, Guilin University of Technology, Guilin, Guangxi 541004, China;
2. Mechanical and Electronic Engineering College, Taiyuan University of Science and Technology, Taiyuan 030024, China;
3. Department of Management, Monash University, Melbourne 3800, Australia;
4. Department of Computer Science, University of Liverpool, Liverpool, L69 7ZF, UK;
5. Department of Computer Science and Software Engineering, Xi'an Jiaotong-Liverpool University, Suzhou, Jiangsu 215123, China)

Abstract: To deal with the problems of traditional search engine in understanding natural language, this article proposes a new intelligent search engine model which is based on the natural language processing and similarity calculation. Its core technology is Chinese word segmentation technique based on natural language processing, semantic similarity and contrary degree. Thinking from the users' view, the model combines DotLucene with those concepts. The precision of the intelligent search engine is about 86.1%. The intelligent search engine can understand the natural languages to query and offer the right answer to users.

Key words: natural language progressing; Chinese word segmentation; similarity; DotLucene; intelligent search engine

0 引 言

随着互联网不断发展,导致信息日益增长,面对海量的信息,如何使中文搜索引擎能够快速方便地帮

收稿日期:2009-07-06.基金项目:广西自然科学基金资助项目(桂科自 0991254).
作者简介:梁晓诚(1958-),男,高级工程师.主要研究方向:计算机应用技术.E-mail:lxc@glite.edu.cn
通讯作者简介:岳晓光(1986-),男,在读硕士.主要研究方向:人工智能.Email:yxg@glite.edu.cn

助用户寻找到有用的信息已经成为当今国际国内研究中文搜索引擎的科研机构的研究热点[1]. 搜索引擎是指在万维网中主动搜索信息并能自动索引且提供查询服务的一类软件系统. 它通过使用网络搜索软件(又称网络搜索机器人)或网站登录等方式, 将互联网上大量网站的页面收集到本地, 经过加工处理而建成数据库, 从而能够对用户提出的各种查询做出响应, 提供用户所需的信息[2].

Google、Yahoo、百度、新浪、天网等搜索引擎采用以关键词检索为基础的检索技术, 用户输入检索关键词向搜索引擎提出请求, 而不是以自然语言形式提供的[3]. 以百度知道 (http //zhidao baidu com)为例, 虽然现在的百度知道看起来是一个智能化的搜索引擎. 但其基础还是以用户问答这种形式, 然后百度通过收集这些信息, 给新提问者一些以往的问答列表. 所以说, 这些传统的搜索引擎还是基于传统的技术, 不能够对自然语言进行真正的理解.

传统的搜索引擎存在的问题是: 不符合人们提出问题的惯常思维, 并且采用关键词为基础的检索技术并不能尽如人意. 自然语言处理可以认为是一种交叉学科的语言文字的处理技术. 运用自然语言处理技术可以发现语言文字所表达的"义"[4]. 只有真正理解了语言文字的"义", 才能改进搜索引擎的效果和使用的便捷性. 基于自然语言处理的智能搜索引擎可以理解为接受用户以自然语言形式表述的提问, 并能从大量的数据中查找或推断出用户问题答案的智能信息检索系统.

1 搜索引擎的相关理论

1. 1 中文分词技术

中分分词理论是研究解决将汉字中的词与词分隔开来的问题的理论. 中文分词的主要方法可以总结为: 基于规则的分词方法、基于统计的分词方法、基于规则和基于统计相结合的方法等等. 其中最常见的分词方法之一是最大匹配算法. 最大匹配算法可以分为正向最大匹配算法、逆向最大匹配算法和双向匹配法等. 其主要原理都是切分出单字串, 然后和词表进行比对, 如果是一个词就记录下来, 否则通过增加或者减少一个单字, 继续比较, 一直还剩下一个单字则终止, 如果该单字无法切分, 则作为未登录词处理[5].

前面讲到词表, 词表属于语料库知识的一部分. 语料 (corpus), 又称为语言素材, 是自然发生的语言材料的集合. 语料库, 顾名思义, 就是存放语言的仓库[6]. 而文献[7]中指出: 要构建高质量的分词语料库当前亟需在以往的分词规范中补充以下 3方面的内容: ①命名实体(人名、地名、机构名)标注细则; ②表义字串(数字、时间、日期、电子邮箱等)标注细则; ③歧义字串消解细则. 虽然是讲述语料库建设的, 但是也从另一角度阐述了中文分词技术的需要解决的要点.

分词是智能搜索引擎的第一步, 通过分词, 为系统获取到关键词打下了基础.

1. 2 相似度计算[8]

相似度计算主要讨论的问题就是语义相似度和对立度. 这里我们把相似度的计算的概念拓展了, 相似度为 1, 则认为相似度很高, 或者认为对立度很低; 相似度为 - 1, 则认为相似度很低, 或者认为对立度很高.

引入文献 8中的几个概念, 再来讲述语义相似度和对立度的概念.

定义 1: 词语全集为 I, $a \in I$, $b \in I$, 记 $Dist(a, b)$ 为 a与 b的语义距离, $Dist(a, b) \in [0, +\infty)$. 若 a与 b为相邻节点, 则 $Dist(a, b) = 1$; 若 a与 b为完全对立节点, 则 $Dist(a, b)$ 为一个无穷小量.

定义 2: 词语全集为 I, A 为阳系词语集合, B为阴系词语集合, $A < I$, $B < I$, 若 $a \in A$, $b \in A$(或 $a \in B$, $b \in B$), 则称词语 a与词语 b同层; 若 $a \in A$, $b \in B$(或 $a \in B$, $b \in A$), 则称词语 a与词语 b不同层.

定义 3: 记 $Disp(a, b)$ 为词语 a与词语 b之间的语义位移, 当 a与 b同层时, $Disp(a, b) = Dist(a, b)$; 当 a与 b不同层时, $Disp(a, b) = -Dist(a, b)$; 显然, $Disp(a, b) \in (-\infty, +\infty)$.

定义 4: 词语全集为 I, $a \in I$, $b \in I$, 记 $Sim(a, b)$ 为 a与 b的相似度, $Sim(a, b) \in [-1, 1]$. 当 $Sim(a, b) = 1$时, a, b为同义词语, 即 $a = b$; 当 $Sim(a, b) \in (0, 1)$ 时, a, b为近义词语; 当 $Sim(a, b) = 0$时, a,

b无相似性; 当 $Sim(a, b) \in (-1, 0)$时, a, b为意义不完全相对的反义词语; 当 $Sim(a, b) = -1$时, a, b为意义完全相对的反义词语. 特别地, 当 $Sim(a, b) \in [-1, 0)$时, $Sim(a, b)$可以称为词语 a和 b的对立度, 记为 $Con(a, b)$, 此时 $Con(a, b) = Sim(a, b)$.

关于对立度的理解, 通俗地说, 当 $Sim(a, b) \in (-1, 0)$时, 说明 a与 b的反义词相似; 当 $Sim(a, b) = -1$时, 说明 a等价于 b的反义词, 也就是 a, b互为反义词. 例如"男人"和"女人"这 2个词通过计算, 得到的语义相似度是 -0.902 5, 而"男人"和"父亲"这 2个词通过计算, 得到的语义相似度是 0.814 5.

语义相似度和对立度计算为拓展系统检索范围打下了基础, 因为人的思维是具有相似、相对性的, 思考一个问题, 自然而然会考虑到对立面的问题, 例如前面所讲述的, 思考到"男人"方面, 很自然的会考虑到"女人"、"父亲"等方面.

1.3 DotLucene简介

先讲述一下 Lucene, Lucene是一个用 Java写的全文检索引擎工具包, 可以方便地嵌入到各种应用中实现针对应用的全文索引/检索功能, 用户可以基于它开发出各种全文搜索的应用[9].

DotLucene是从 Apache的 Lucene(java)项目移植到. NET(C#)上的. DotLucene相率非常高, 而且它和 Lucene的索引兼容, 所以可以不同的平台间迁移而不会丢失任何索引数据, 其系统结构如图 1所示[10].

2 搜索引擎关键处理过程分析

2.1 提取关键词

为了能够搜索到相应的网页, 需要事先提取关键词. 搜索引擎不是在接到用户查询请求时才做的. 这主要是从 3个方面来考虑: 响应时间, 即便网络比较畅通的情况下, 下载一篇网页大概也需要 1 s钟; 搜索的使用效率, 比如全球最大的中文搜索引擎百度每天要响应来自 138个国家超过数亿次的搜索请求; 考虑到系统负担过重[11]. 网页对应着不同的关键词, 就能够检索到相应的网页. 提取关键词的时候也要考虑到 2个问题:

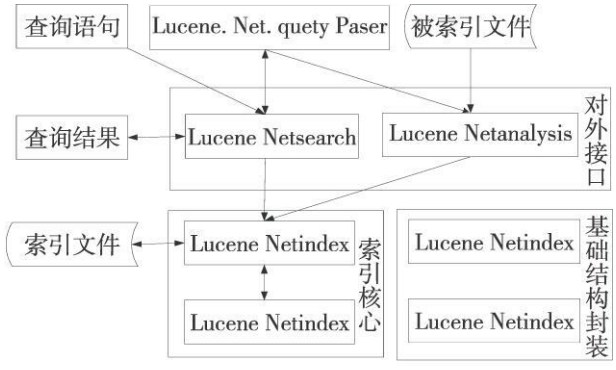

图1 DotLucene系统结构图

Fig.1 System structure map of DotLucene

1)多语种对应问题, 一个智能的搜索引擎应当能够意识到用户查询的多种语言对应的本地化网页. 例如用户查询"computer"和"计算机"都是对应着相似的网页.

2)用户提问的要点. 用户提问的时候心情语气, 计算机不得而知, 计算机不知道用户是高兴或者生气, 但是计算机可以判断用户提问的要点. 通过标点符号判断, 是一个相对容易处理的问题. 但是, 我们研究认为大多数用户不会在搜索的时候打上标点的, 因为用户已经习惯了传统的搜索引擎中的关键字的方法去求解. 例如, 用户想知道北京离上海有多远, 可能会输入"北京 上海 多远"此类的"词或词组 + 空格"的组合查询方式. 所以, 智能搜索引擎可以参照这个思路, 在(1)的基础上, 对应着多语言, 按照用户提问的要点的思路进行提取关键词, 并对用户的查询做出正确响应.

2.2 相似度和对立度计算

在建立词库的时候要统计词之间的"远近亲疏"程度, 计算时可以参考文献[8]中的公式和概念. 首先讲述语义位移和语义相似度的概念. 语义位移与语义相似度之间有着密切的关系: 2个语义位移的绝对值越大, 其相似度的绝对值越小; 反之, 2个语义位移的绝对值越小, 其相似度的绝对值越大.

定义 5 词语全集为 I, $a \in I$, $b \in I$, 语义相似度与语义位移之间的关系如下:

$$Sim(a, b) = \begin{cases} \lambda^{Disp(a, b)}, Disp(a, b) \in [0, +\infty) \text{时} \\ Con(a, b) = -\lambda^{-Disp(a, b)}, Disp(a, b) \in (-\infty, 0) \text{时} \end{cases}$$

其中, λ为一个常量参数, λ∈ (0, 1), 表示当语义距离为 1时的语义相似度. 特别地, 当 $Con(a, b) = -1$时, $Disp(a, b) = -\varepsilon$, $Dist(a, b) = \varepsilon$, 其中 ε为一个无穷小量.

通过比对搜索关键字, 计算词库中的名词、量词、动名词的相似度和对立度, 可以为查询者提供更加全面的查询. 例如: 查询“山大在哪? ”和查询“山东大学在哪? ”、“山西大学在哪? ”、“中山大学在哪? ”都是相关联的. 用户查询完“山大在哪? ”, 从而也得知“山大 ”这个概念并不是一定唯一的, 不同的地域有不同的“山大”, 所以说对用户也是一个学习的过程, 反过来, 用户也会觉得这个搜索引擎是智能的.

2 3 检索结果排序

检索结果的排序就是一个综合评价的问题. 对于排序结果不可能有一个十全十美的结果, 因为每个人的偏好喜欢是不同的[2]. 网络中的数据每天都在不断增加, 在查询界面中随便输入一个关键词, 返回的结果都有几百甚至更多. 按每次页面能显示 10条记录来算查询响应较快的搜索引擎往往更能得到用户的青睐, 而用户往往只会浏览到前 3, 4页. 所以查询结果应将更有用、更准确的信息放置在前面[11].

3 实验及其结果分析

系统在. NET 平台下, 使用 C#和 ASP. NET 技术实现, 整合主题网络机器人和 W eb服务器, 建立一个智能搜索引擎. 其原理图如图 1所示.

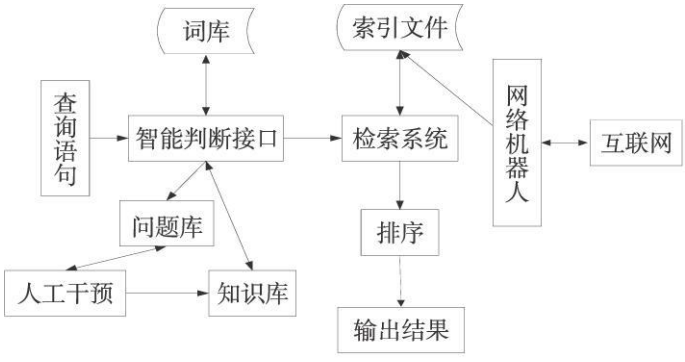

图2　智能搜索引擎原理图

Fig.2　Principle map of intelligent search engine

其主要工作原理是: 系统利用网络机器人对指定的网站进行信息搜索, 提取网站所有的 URL, 并获取关键字, 然后建立索引. 为了保持页面的更新率, 系统每周进行一次搜索. 然后进行检索的时候, 先进行智能分析, 如果分析出现问题, 先进行知识库查询判断, 如果仍然无法解决, 则存放到问题库, 由日后管理员人工干预转化为知识库. 系统调用 DotLucene分析查询器分析智能判断接口提交的查询, 然后调用 IndexSearcher类进行搜索, 最后返回的是 H its类, 分页显示. 在页面排序的时候, 根据计算的该页面的关键词的相似度和对立度的高低进行排序. 一般的, 我们认为, 相似度高的应该拍在前面, 对立度高的可以有选择的拍在前面一部分. 但是, 并不代表着相似度高的网页一定排在前面, 这个需要和对立度结合在一块考虑.

经过学校网页、学校论坛和一些指定网页的网页测试, 其索引测试数据如表 1所示.

表 1　索引测试数据

Tab 1　Index test data

抓取时间 /m in	总的页面个数 /页	平均下载速度 /(页• m in^{-1})
100	20 361	203. 61

把上面的页面存储下来, 进行搜索, 搜索结果和查准率如表 2所示.

表 2　搜索结果和查准率

Tab 2　Search results and precision

关键词	大学	公告	Computer	在哪
页面数目 /页	820	133	95	158
时间 /s	2 90	1 05	1 01	1 47
查准率 /%	97. 2	98. 5	85. 0	83. 9

经过对常用的 30个关键词组成的 155句自然语句进行查询, 平均查准率达到 86 1%, 实现查询的要求.

4 结 论

文章里面讨论的基于自然语言处理的智能搜索引擎, 以及和 DotLucene结合起来的研究. 中文分词技术可以实现对输入的查询语句的切分, 这也是人工智能将要突破的一个重大问题, 通过自然语言处理技术可

(下转第 88页)

3 结 论

经实验分析: 采用溶剂法回收废弃印刷线路板中环氧树脂是可行的. 其中从线路板中得到环氧树脂的比率为 15.37%. 回收环氧树脂的最佳反应条件是: 反应温度 80℃, 反应时间 3 h, 硝酸浓度 8 mol/L, 投料比为废弃线路板质量 $M_{板}$:硝酸体积 $M_{酸}$ = 10 g:50 mL.

参考文献:

[1] Jiuyong Guo, Jie Guo, Zhenming Xu. Recycling of Non-Metallic Fractions from Waste Printed Circuit Boards: A Review. Journal of Hazardous Materials, 2009, 168: 567-590

[2] Jae-Min Yoo, Jinki Jeong, Kyoungkeun, et al. Enrichment of the Metallic Components from Waste Printed Circuit Boards by a Mechanical Separation Process Using a Stamp mill[J]. Waste Management, 2009, 29(3): 1132-1137.

[3] Guan Hua Xing, Janet Kit Yan Chan, Anna Oi Wah Leung, et al. Environmental Impact and Human Exposure to PBCs in Guiyu, an Electronic Waste Recycling Site in China[J]. Environment International, 2009, 35(1): 76-82

[4] 程果锋, 路书玉, 罗丽娟. 废弃印刷线路板中环氧树脂的资源化技术[J]. 再生资源与循环经济, 2008, 1(9): 30-33;

[5] 久保内, 昌敏, 党伟荣, 等. 胺类固化剂固化的双酚环氧树脂回收再利用的研究[J]纤维复合材料, 2002, 1(19): 58-60

[6] 庄燕, 陆文雄, 李小亮, 等. 废弃线路板中非金属材料的回收和利用[J]. 上海化工, 2008, 6(33): 1-5.

[7] 周翠红, 路迈西. 废旧电路板的组成与解离特性研究[J]. 环境污染治理与设备, 2005, 4(4): 28-31

(上接第 79页)

以理解用户的自然语言, 用户可以计算机进行真正的人机交流, 从而使搜索引擎明白检索怎样的文档. 而引入相似度和对立度的概念, 可以增加检索信息, 同时为网页文档排序提供了一个更好的检索结果. 所以, 从用户的角度考虑问题, 从人的思维方式的角度考虑问题, 结合人工智能的研究成果, 是智能搜索引擎的必经之路.

参考文献:

[1] 吴晓辉, 宋萍萍. 中文智能搜索引擎分析和框架模型的设计[J]. 情报科学, 2008, 26(12): 1814-1815.

[2] 陈魁. 智能搜索引擎系统的分析设计与开发[D]. 大连: 大连理工大学, 2004

[3] 陈林, 杨丹, 赵俊芹. 基于语义理解的智能搜索引擎研究[J]. 计算机科学, 2008, 35(6): 152-154

[4] 吴晨, 张全, 缪建明, 等. 语义理解下的自然语言处理及信息检索模型[J]. 计算机科学, 2008, 35(5): 113-117.

[5] 麦范金, 赵子强, 岳晓光. 基于语义相似度的主观题阅卷系统模型设计[J]. 微计算机信息, 2009(6-3): 255-256.

[6] 蔡雷. 语料库技术在英语教学中的应用与研究[J]. 宿州学院学报, 2008, 23(5): 159-161

[7] 李玉梅, 陈晓, 姜自霞, 等. 分词规范亟需补充的三个方面内容[J]. 中文信息学报, 2007, 21(5): 3-7.

[8] 麦范金, 王挺. 语义对立度及其计算模型的研究[J]. 中文信息学报, 2008, 22(4): 39-42.

[9] 严良达. 基于 Lucene搜索引擎的设计与实现[J]. 宁波职业技术学院学报, 2009, 13(2): 57-60.

[10] 李占波, 廖继东, 李华. 基于 DotLucene的垂直搜索引擎的研究[J]. 微计算机信息, 2007, 23(8): 194-195.

[11] 杨倩晨. 浅析搜索引擎的运行机制[J]. 大众科技, 2009, 117(5): 41-42