语义技术在搜索引擎算法中的应用研究

王清飞

(郑州牧业工程高等专科学校 图书馆,河南 郑州 450011)

摘 要: 针对海量信息的冲击,专家提出了语义技术的概念,作者在文章中主要讨论了语义技术在分词算法、知识库建设、输出排序算法中的应用。

关键词: 语义技术; 搜索算法; 语义搜索
中图分类号: G354.4 文献标识码: A 文章编号: 1004 - 1680(2012) 01 - 0007 - 04

由于第二代互联网技术的发展,很多传统的信息使用者转换成了信息的发布者和使用者,这样导致了网络信息量的激增,在海量的信息面前,人们逐渐对传统的搜索引擎能否满足人们的需要产生了质疑,关键字搜索引擎在这样的大环境下也逐渐改进了自己的服务技术,引用了诸如历史检索、联想检索、自然语言检索和建立用户使用历史库来应对海量信息的冲击。针对继续成几何倍数增加的信息,专家学者提出了语义技术的概念,将语义技术引用到搜索引擎中,增加标引的精准性和相关概念的语义联系性,这样可以极大地提高检索的检全率和检准率,可以将用户最关心的部分优先呈现给用户。

1 语义技术在分词算法中的应用

对于分词算法,目前主要有两种,一种是类似于西方的语言,另外一种是类似于亚洲语言系统。前者对其进行分词主要是依据空格分隔符,还有就是几种特殊的情况,如数字、连词符、标点和字母的大小写; 而后者则没有明显的分隔符,并且很多的中文单词还有多义性跟语境关系,在机器操作识别过程中有不小的困难,学术界对其研究颇下功夫,已有成熟的分词技术。通常主要有两类分词方法:

(1) 主要是基于语法、句法并结合语义进行分析,通过上下文内容所提供的内容信息进行分析,并界定该单元词的语义。

(2) 主要是基于词的出现频率以及系统提供的词典进行匹配来实现的[1]。

这两种方法比较起来各有特点,前者具有一定的语义理解能力,对检索的精度以及准确度都有很好的优势,但是算法复杂,不容易实现; 而后者方法简单实用,并且容易实现,但是由于仅仅是简单地匹配,缺乏语义效果,这样的检索结果很大程度上取决于检索式的构建,检索的效率不是很令人满意。

常用的分词算法如下:

(1) 逆向最大匹配法(Reverse Maximum Matching Method) 。我们常称其为 RMM 法。RMM 法的基本原理是设 A 为字典,MAX 表示 A 中的最大词长,STRING 表示为待切分的字符串,RMM 法就是从后面开始,每次从字符串中读取出长度等于 MAX 的子字符串与字典 A 中的词语进行匹配,若成功,则该子串为词,指针前移 MAX 的子串与 A 中的词继续匹配,直到全部匹配完毕。

(2) 正向最大匹配法(Maximum Matching Method) 。我们常称其为 MM 法。其基本思想与逆向最大匹配法相似,不同的是它是从前向后进行逐个匹配比较的。

(3) 专家系统分词法。此种方法主要是根据专家的经验进行组合,构建专业的词典及字库,供检索工具进行分词判断。

(4) 高频优先法。高频优先法就是在字典中对每个词出现的概率进行统计与分析,根据其出现次数的不同分别赋予不同的优先级,然后在进行匹配的时候优先匹配高频词汇,这就是高频优先法。此外还有全自动词典分词、扩充转移网络分词法、神经网络等分词方法。但是以上所有的分词方法都会存在一定的错误率。

我们必须要用一种新的思维方式来进行汉语自动分词的研究,来应对现有的分词方法和技术不能解决现在机器实现自动分词的困难。这种思维方式就是根据汉语的特点以及其自身的规律,可以考虑从汉语的构词规则如笔画来试图解决这样的问题,

这是一种新的自动分词研究方向。另外再对数据或者信息资源进行处理和标引时,在使用受控语言和自然语言的同时,还可以运用其他的人工构造语言,像程序设计语言、逻辑语言、数学公式等。这几种人工语言跟自然语言一样,都由一套自己完整的语法规则与相当数量的符号组成,我们可以根据他们的特点,对其组成规律进行总结,并且并入到语义字典中,并利用本体对其相互关系进行描述。同时,这些人工语言在很大程度上都适应了计算机技术的发展,并且各种程序设计语言和数学逻辑语言中的符号更是十分齐全,完全能够对信息进行表述、传递、组织、存储、检索并实现信息的交流。

为了使汉语便于计算机自动分词处理,就需要对现在的分词语法作以下三方面的改造: 一是要在待分词的汉语文本的词与词之间增加适当的间隔符,即设立分词标志。通过设立分词标志,可以方便地对文本进行分词处理,大大地提高效率; 二是要对汉语的词语进行明确界定,即规定什么是词,这样主要可以解决两方面的问题,单字词与字元素之间的区别,以及短语或者成语与词(词组) 之间的区别; 三是要制定完整的汉语书写规则。这些都是应该在汉语文本生成之前完成,并且从技术实现的条件来看是可行的,方便实现,这样就省去了很大一部分精力对预处理文本进行分词了[2]。

分词技术发展到现在已经比较成熟,已经出现了不少的分词算法,具体的几种分词算法比较见表 1。同时,对分词系统的研究也取得了不小的成绩,国内外已经有不少的成功软件系统,其中有一定代表性的主要有: Microsoft Research 多国语言处理平台 NLPWm 中的中文词语分析词系统、中国科学院计算技术研究所基于层叠隐马模型的汉语词法分析系统 ICTCLAS(Institute of Computing Technology, Chinese Lexical Analysis System) 、中科院自动化所三元统计模型、哈尔滨工业大学统计分词系统、清华大学 SEGTAG 系统、北京大学计算语言所分词和词类标注系统。它们的分词准确率有的已经达到 99% 以上。

表 1 分词系统比较

算法	时间(字每 min)	准确度(%)
基于词表的分词_最大匹配	66000	95.42
基于规则和基于统计相结合	200000	97.95
基于统计的分词	41000	96.25

2 语义技术在知识库构建中的应用

知识库(Knowledge Base) 是知识工程中结构化、易操作、易利用、全面有组织的知识集群,是针对某一领域问题求解的需要,采用某种(或若干) 知识表示方式在计算机存储器中存储、组织、管理和使用的互相联系的知识片集合。这些知识片包括与领域相关的理论知识、事实数据,由专家经验得到的启发式知识,如某领域内有关的定义、定理和运算法则以及常识性知识等。知识库是由类似于人类的背景知识和相应的推理规则所组成的,知识库使基于知识的系统具有智能性。并不是所有具有智能的程序都拥有知识库,只有基于知识的系统才拥有知识库。一般的应用程序与基于知识的系统之间的区别在于: 一般的应用程序是把问题求解的知识隐含地编码在程序中,而基于知识的系统则将应用领域的问题求解知识显式地表达,并单独地组成一个相对独立的程序实体[3]。

知识库就是人的认识库,它是由人对客观世界的认知过程所留下的经验组成的。其是否丰富,很大程度决定着检索效率的高低,它是实现智能信息搜索的核心和基础。知识库可以对输入的信息进行接受、判断、提取、分析和概括之后形成自己新的知识并保存,成为下一次分析、概括的依据和材料,这样知识库就始终处于一种持续不断的增长状态。知识库主要有以下特点:

(1) 知识库中的知识根据它们的应用领域特征、获取时的背景信息、使用特征、属性特征等而被构成便于利用的、有结构的组织形式。

(2) 知识库的知识是有层次的。最低层是“事实知识”,一些事物存在常识上的关联,如电灯与开关、相机与照相,在计算机进行处理时,应该使用语义网络来表示这些知识; 中间层是本体论层面,对概念的本体论的定义与解释,概念之间复杂的语义关系,也就是我们通常所讲的知识,是对具体事物描述的抽取; 最高层次是“策略”,它以中间层知识为控制对象。策略也常常被认为是规则的规则。因此知识库的基本结构是层次结构,是由其知识本身的特性所确定的。

(3) 典型方法库,如果对于某些问题的解决途径是肯定和必然的,就可以把其作为一部分相当肯定的问题解决途径直接存储在典型方法库中。这种宏观的存储将构成知识库的另一部分。在使用这部分时,机器推理将只限于选用典型方法库中的某一层体部分。

(4) 增广知识库,知识库中可有一种不只属于某一层次(或者说在任一层次都存在) 的特殊形式

的知识——可信度。因为在数据库的处理过程中一切都是确定的,不存在不确定性度量[3]。

知识库是新提出来的一个概念,但是它与存在已久的因特网信息资源的关系很是密切,知识库就是组织有序的、描述详尽的信息库,包括信息以及对信息之间的描述。而信息库就是我们经常用到的网络,其上面存在的信息是无序的、非结构的,并且好多是异构的信息,难以有效地利用。这样两者之间就存在密不可分的关系,知识库就是要对信息库进行组织、提取、分析、标引和检索,而信息库则是知识库的原材料库。信息库(因特网) 也是用户所要检索的最终目的地,智能搜索引擎所做的就是通过知识库把用户的问题或者检索式提高到知识(概念) 的层面,组成含有语义知识的检索式,然后利用这个知识(概念) 检索信息库,这样返回给用户需要的检索结果。实质上,利用知识库检索信息库与信息库映射到知识库的概念与概念关系,是一个知识获取过程。

想要实现语义检索,就必须将信息库与知识库相结合,两者的有机结合需要做到以下三个方面:

(1) 知识管理。知识管理主要就是应该实现知识库的自我增长和自我扩充,知识库增长的基础是对信息库中的信息资源的提取与组织,所以知识管理就是根据其管理的策略对信息进行分析和描述,然后将新增加的知识元素添加到知识库中对其进行扩充。

(2) 语义分析。语义分析就是对用户的检索提问进行分析,明确其具体的含义,其应该具有以下几个方面的功能: 第一,要具备分词功能,将待处理语句进行分词处理与匹配; 第二,具有处理同义词的功能,应该具有同义词的词典,有足够的能力处理意思相近的词语; 第三,根据知识库分析关键词,明确该关键词的概念与语义,并根据分析结果确定用户真正的用意,利用正确的检索式对目标库进行检索; 第四,一定程度的知识库。

(3) 知识检索。知识检索是实现智能搜索的最后一环,通过前面语义分析结果,明确用户用意,对信息库进行知识(概念) 层次的检索,在给出准确答案的同时,给出用户相关问题,从多方位对用户的问题进行解答。

目前人们对知识库系统研究已经有了一定的成果:

(1) Cycle 知识库。Cycle 系统是由美国 CYCorp 公司开发的一套最新的知识库系统,Cycle 知识库是一个经过形式化的海量的人类基本知识库,其目标是通过知识工程手段,让计算机建立起和人一样的常识系统,CYC 系统的知识服务器包含一个非常庞大的多语境知识库和推理引擎。推理引擎用于执行通用的逻辑推理[4]。因为 Cycle 知识库系统含有模糊或复杂语法的句子,所以它对自然语言处理有很大的帮助,并且,Cycle 可以利用自己对已经存入的常识知识的理解而对输入的语句进行正确的语法分析。Cycle 自然语言处理系统有三个组件[5]: 词典、句法分析器、语义解释器。这些在自然语言理解网页中有非常详细的描述。现在,我们集中关心三个部件的概貌。目前,我们能够正确分析许多不同的句型,包括歧义与语法复杂的输入。Cycle 能够处理否定,Modals,嵌套量词,并且正在开发用户接口。他们同时也正在做一个生成部件,它能够从 Cycle 表示的公式自动生成英文[6]。有了自然语言处理器,就可以用自然语言进行数据库的查询,更方便地进行搜索,并提供友好的搜索引擎人机交互界面。另外,由于网络中数据存储有多种格式包括结构化、半结构化以及非结构化,Cycle 可以将其转化为有用的知识并进行存储。这些知识通过语义综合总线联入 CYC 系统,CYC 将数据库中的记录和知识库中的断言相联系,在推理的过程中,对记录进行自然语言处理。然后到知识库中为它所涉及到的断言进行定位[7]。

(2) UMLS(Unified Medical Language System) 。UMLS 是美国国立医学图书馆主持的一项长期开发研究计划,目的是为了建立一个计算机化的可持续发展的生物医学检索语言集成系统和机读情报资源指南系统。其目的在于提高计算机程序“理解”用户提问中生物医学词汇涵义的能力,并利用这种理解帮助用户检索和获取相关性的有用信息资源。其包括三个部分: 超级叙词表(Metathesaurus) 、语义网络(Semantic Network) 和专家词典(SPECIALIST Lexicon and Lexical Programs) [8]。

3 语义技术在输出排序算法中的应用

检索结果的输出是直接呈现给用户的,这个环节的好坏直接关系到用户对该搜索引擎的评价,所以能否将用户最想要的结果首先呈现给用户,显得尤其重要。最直接最简单的相似度计算方法是直接计算查询语句与文档的点积,即:

$$\mathrm{sim}(a,b) = \sum_{i=1}^{n} W_i Q_i \qquad (3.1)$$

其中,Q_i 为 Q 在第 i 个词条的权重,W_i 为 Q 在

第 i 个词条在文档 A 中的权重。

目前比较流行的是余弦表示法,因为该方法考虑到文档的长度与字符串的长度问题,具体公式如下:

$$\mathrm{sim}(a,b) = \frac{\sum_{i=1}^{n} W_i Q_i}{|a| \cdot |b|} \qquad (3.2)$$

其中 Q_i 为 Q 在第 i 个词条的权重,W_i 为 Q 在第 i 个词条在文档 A 中的权重,$|a|$ 表示文档的长度,$|b|$ 表示查询串的长度。

检索结果的语义关联即对检索结果赋予语义,意思就是检索的结果不仅仅是纯碎的文档,而是语义对象及相关的实例。其中语义关联就是非常重要的一类语义对象,它所指的不仅仅是一个个的实体,也包含各个实体之间的关系。研究者可以通过背景指数、深度指数等来研究语义实体之间的关联[3]。其中背景指数反映了语义关联通过用户所感兴趣的区域情况; 深度指数反映了语义关联中的实体和关系规范化程度。

作者认为将语义关联模型引入到搜索引擎的输出排序算法中,可以通过体现各个本体之间的语义关联,在关键词与实际本体之间匹配的同时,通过相关本体之间的关联性,来进行检索的语义扩充,可以在很大程度上提高搜索的检全率,同时搜索引擎也可以智能地回答用户的提问。

参考文献:

[1] 印 鉴. 搜索引擎技术研究与发展[J]. 计算机工程,2005(7):54 - 56.

[2] 邱均平,文庭孝,周黎明. 汉语自动分词与内容分析法研究[J]. 情报学报,2005(6):310 - 317.

[3] 知识库[EB]. 2005. http://baike.baidu.com/view/7976.htm? fr = ala0_1_1

[4] 项 珍. 基于语义的搜索引擎研究[D]. 南京: 东南大学,2007.

[5] what is CyC? [EB]. http://www.cyc.com/cyc/technology/whatiscyc_dir/whatsincyc

[6] Minkov. 什么是 CYC? [EB]. 2009. http://blog.sina.com.cn/s/blog_620502150100fdk1.html

[7] 邱均平,余以胜. 基于知识库系统的智能搜索引攀研究[J]. 现代图书情报技术,2005(7):413 - 416.

[8] 龚 芳,耿 骞,王 洋. 自然语言检索中的概念控制[J]. 中国图书馆学报,2005(4):45 - 48.

[9] Aleman - Meza B. Context - aware Semantic Association Ranking [EB]. 2008. http://lsdis.cs.uga.edu/lib/download/AHASO3 - SWD - Workshop

The Application Research of Semantic Technology in the Search Engine Algorithm

WANG Qing - fei

(*Library*, *Zhengzhou College of Animal Husbandry Engineering*, *Zhengzhou* 450011, *China*)

Abstract: Because of the impact of the mass information, experts put forward the concept of semantic technology. The author in this paper mainly discusses the semantic technology used in words segmentation algorithm, knowledge base construction and output sort algorithm.

Key words: semantic technology; search algorithm; semantic search

作者简介: 王清飞(1983 -) ,男,硕士,郑州牧业工程高等专科学校图书馆助理馆员,研究方向: 网络信息资源管理与参考咨询。

收稿日期: 2011 - 09 - 24
(责任编辑 段麦英)