

●金燕 张玉峰

基于中文自然语言理解的知识检索模型^{*}

摘要 基于中文自然语言理解的知识检索模型的设计思路是:通过对用户提问及 Web 文档信息进行语义层次的自然语言处理,构建概念和概念网络,针对用户真实查询需求与概念网络及其映射的源文档进行推理匹配,然后对检索结果进行排序处理,提交给用户。图 2。参考文献 5。

关键词 知识检索 中文自然语言 检索模型 概念网络

分类号 G354

ABSTRACT In this paper, the authors introduce the conception for a knowledge retrieval model based on Chinese natural language understanding, which includes the natural language processing of user's queries and Web documents at the semantic level, the construction of concepts and concept networks, the inferential matching of user's real queries with source documents of concept networks and their maps, the sorting of search results and the submission of results to users. 2 figs. 5 refs.

KEY WORDS Knowledge retrieval. Chinese natural language. Retrieval model. Concept network.

CLASS NUMBER G354

目前 WWW 上流行的基于关键词匹配的检索方法,仅仅靠目标文档中是否出现用户查询所用的关键词来判断文档的相关性。这种以词为中心建立关联的方式,缺乏语义理解能力,容易割裂文档内容间的知识关联,也未能妥善解决一词多义、一义多词问题。更为重要的是,由于 WWW 检索面向的是终端普通用户,而普通用户不具备专门的检索知识,在表达其需求尤其是构造基于关键词的标准检索表达式时存在较大困难。因而,即使不考虑其他因素,仅是检索需求的表示不准确,也会造成检索结果“真实相关度”较低的状况。针对这些问题,本文提出一种基于中文自然语言理解的知识检索模型,提供基于自然语言和概念的知识检索机制,使信息需求表示符合用户的习惯,也增强了系统的知识处理能力,在减轻终端检索用户的认知负担的同时,把检索从基于关键词层面提升到基于知识的层面上来,弥补基于关键词检索的不足,提高信息检索的效率。

1 自然语言理解与知识检索

知识检索是为了解决 WWW 信息检索中存在的问题、信息很多但知识很少、检索效率很低的问题而提出的一种新的信息检索理念。在对蕴含于信息中

的知识和知识关联进行分析的基础上,在知识处理技术和知识组织技术的支持下,实现信息查询深入到语义理解的智能化阶段。也就是说,知识检索综合应用信息科学、人工智能、认知科学及语言学等多学科的先进理论与技术,基于知识和知识组织,融合知识处理和多媒体信息处理等多种方法与技术,是一种能充分表达和优化用户需求,高效存取所有媒体类型的知识源(文本、图像、视频、声音等)并能准确精选用户需要的结果的高级信息检索方法^[1]。一般而言,知识检索具有如下基本特征:(1)支持自然语言检索。知识检索应该具有分析和理解自然语言的能力,能分析和处理自然语言形式的用户提问和文档信息内容。(2)支持词语、语义内容的处理,实现同义词扩展检索和关联检索。(3)具有概念推理和学习功能,利用概念网络的多维认知结构,实现多维语义推理和动态连接学习。(4)具有强大的人机交互功能,能够通过自然语言和知识语言进行人机交互,并利用各种反馈机制向用户学习。

自然语言是人们日常交流所使用的语言,与人工受控语言相比,灵活方便,为大众所熟知。自然语言用于信息检索,优点比较明显:(1)减轻了用户的认知负担,易于用户准确、真实地表达其信息需求。

^{*} 本文为教育部人文社会科学研究重大资助项目“信息可视化与知识检索”(项目编号 02JAZJD870004)研究成果之一。

用户不需要具备专门的检索知识,便可以用自己熟悉的语言形式准确表达其信息需求,避免了构造复杂检索表达式的过程。(2)对于新出现的事物和很少文献论述但其名称可以确定的事物,检索效果相当好^[2]。(3)降低文献处理成本,加快文献处理速度,减轻乃至消除了文献处理难度,增强了系统的易用性^[3]。但其缺陷也同样存在,如词义模糊、词间关系不清等易造成漏检误检,以及自然语言抽词、切分词技术,短语识别技术,同义词处理技术等尚未得到有效解决,给自然语言的利用也带来一定困难。但必须承认,随着信息技术、计算机技术的迅速发展,以及语言学研究的突破,自然语言对信息检索的作用会日益增加。

一般而言,自然语言应用于信息检索有多种方式,有的用于设计人机接口,有的用于信息源的分析和检索过程中,如信息源的自动分词、自动聚类。本文提出的知识检索模型是基于自然语言理解的,自然语言处理技术既用于设计人机接口,又用于检索过程。

基于中文自然语言理解的知识检索模型包括 3 个核心功能:(1)语义理解分析。通过中文自然语言理解与处理技术实现。(2)知识库组织。包括构建概念网络、各种词典和规则库等。(3)知识检索机制,即问题的求解与反馈。模型采用分布式结构。其简单结构如图 1 所示。

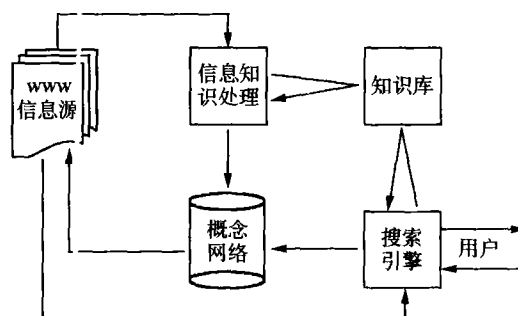


图 1 知识检索模型简图

该模型主要由 3 个模块构成,即自然语言接口模块,信息源采集与知识处理模块,知识检索求解与处理模块(如图 2 所示)。各主要组成部分功能如下:

2 基于中文自然语言理解的知识检索模型

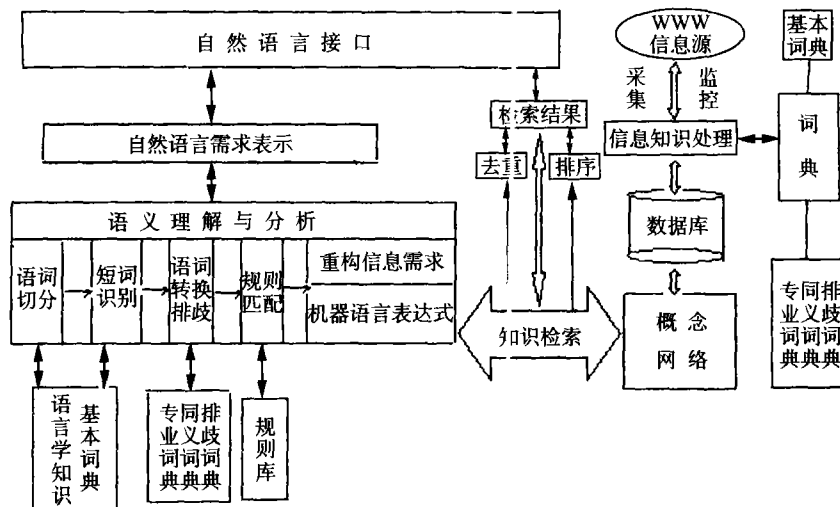


图 2 基于中文自然语言理解的知识检索模型

(1)自然语言接口。它是一种智能接口。其主要任务是向系统发出检索请求并接受系统的服务,具体包括:接受自然语言请求,并对其进行语法、语义分析,然后生成机器可识别的查询逻辑即概念和概念逻辑;与用户交互,接受并反馈用户信息;向用户发送系统服务,包括提供检索结果等。

(2)语义理解分析。是自然语言接口执行的主要任务之一,因它是整个检索系统查询的起点,故在

此单列说明。主要是对自然语言提问进行语法、语义分析,推导其内容属性,以充分理解用户的真实请求,提取有价值的查询概念。这个任务要借助语言学知识以及各类词典、规则等实现。由于中文自然语言的特殊性,如词间没有空格自动分隔,汉语语词无法自动表达语气、语调、轻重音、停顿等,短语词组的边界难以自动界定等,因而,中文自然语言的处理也比较复杂,如汉语的自动分词、短语识别、语义排

歧等。构造准确、全面的词典有利于弥补汉语自然语言理解中存在的问题。目前中文信息处理通常按如下步骤进行:借助停(禁)用词、非用词词典,去除无关字词;对留下的语词进行识别、切分、抽词(借助基本词词典、主题词词典、专业词词典等);再借助同义词词典、反义词词典等对相关特征词进行扩展,或根据情景、上下文和领域利用专业词典进行限定;或利用检索交互等对相关词进行扩展、排歧,实现对汉语语词的一词多义、一义多词等易造语义理解模糊现象的控制。

(3)信息源采集与处理。由于 WWW 信息资源丰富,对信息源的采集借助 Robot 自动程序进行。Robot 肩负双重任务:自动采集目标网站信息,并提交给服务器进行处理;自动监控目标网站的变化,并及时更新变化。对采集到文档信息进行处理,包括提取文档中有意义的能够表示文档特征的词,利用已建立的各类词典,识别概念和概念间语义关联,对文档语义内容进行理解和表达。概念与源文档之间通过概念索引相互关联。

(4)概念网络的构建。概念是人对客观事物本质特征的概括,通过字、词、词组等概念描述元素表达出来^[4]。概念有上位概念、下位概念、同义概念之分。概念并非孤立存在的,总是与其他概念之间存在各种各样的联系,这种联系实际上反映的是知识因子之间的关系。一般而言,概念间的关系包括如下几类:①蕴含关系。即概念之间是等级种属关系,如计算机和微型计算机。②同一关系。即等同关系,指具有同一外延而又有不同内涵的概念之间的关系,多是同义词之间的关系,如电子计算机和电脑。③相关关系。即相反、交错、因果、矛盾等多种关系,如烟与肺癌、西服与领带等。在对自然语言信息进行语词、语义分析之后,抽取相关概念,根据概念间的关联关系,构建多维概念网络。概念网络的构建按如下步骤进行。

第一,识别概念。

第二,建立概念关联。构建概念关联有多种途径:借助多种词典,在各种规范词和非规范词间建立联系;借助模糊推理学习、语义网络等技术,构建语义关联;借助语词共现技术,建立特定关联,如足球与韩国,足球与日本,这些词原本没有语义联系,但在 2002 年世界杯期间的新闻报道中,这些词间有很强的语义关联。

第三,选用合适的形式构建概念网络,如概念

图、概念树、语义网、本体等。

第四,对新的文档依此顺序进行处理,借助自动索引、自动分类聚类技术,把新概念归入概念网络中。

(5)知识检索。在对用户查询请求和文档内容语义分析理解的基础上,把析出概念与概念库或概念网络进行逻辑匹配,查找出源文档。逻辑匹配机制如下:

若用户(输入概念即“被”)检索概念与概念网络中的某一概念一致,且概念网络中无同义概念,则根据概念网络与源文档的映射,直接求出源文档。

若检索概念与概念网络中某一概念一致,且概念网络中有同一概念,则把该概念与其同义概念一同激活,推导出其映射的源文档。同义词扩展检索得以实现。

上述两种情况下,系统还可以根据用户检索概念,把概念网络中与其具有蕴涵和相关关系的概念析出,根据概念与源文档之间的相互关联,检索出相关文档,实现概念关联检索。

若用户检索概念与概念网络中的概念均不一致,系统会与用户进行交互学习,根据交互信息重复上述过程。交互包括引导用户完善、精炼其需求或给出相关概念向用户求证。

(6)检索结果的处理。仅仅把源文档析出并未完成全部检索任务,还需要对检索结果进行去重和排序处理。检索结果的处理不仅仅发生在结果检出之后,在信息源收集阶段利用网络数据挖掘技术,通过超链分析,检测路径相同的相似文档和镜像网站,即检索结果的前处理也可以帮助去除重复文档^[5]。

3 模型评价

该模型的设计思路是:通过对用户提问及 Web 文档信息进行语义层次的自然语言处理,构建概念和概念网络,针对用户真实查询需求与概念网络及其映射的源文档进行推理匹配,最后对检索结果进行排序处理后提交给用户。自然语言处理在语义层次上分析、理解文档信息和用户提问,融入了知识处理因子,利用概念网络,辅之以同义词词典、专业词典、排歧词典等后控词典,实现同义词扩展检索和关联检索,可以提高检索的知识性和真实相关性,检全率和检准率也由此得到改善。和基于关键词的检索相比,该模型具有如下优势:(1)允许用户以自然语言提出检索需求,符合人们日常的思维习惯,减轻了用户寻找关键词、构建检索表达式等带来的认知负

●董 慧 杜文华

基于本体和多代理的数字图书馆信息检索模型*

摘 要 在分析图书馆传统的信息检索机制的局限性的基础上,提出了基于本体和多代理的数字图书馆信息检索模型,并介绍了该模型各部分的作用和功能。图3。参考文献3。

关键词 数字图书馆 信息检索模型 本体 多代理

分类号 G250.76

ABSTRACT After analyzing the limitation of traditional library information retrieval mechanisms, the authors propose a model of digital library information retrieval based on ontology and multiple agents, and introduce the functions of components parts of the model. 3 figs. 3 refs.

KEY WORDS Digital library. Information retrieval model. Ontology. Multiple agents.

CLASS NUMBER G250.76

1 图书馆传统的信息检索机制的局限性

一般来说,信息检索机制有两方面的含义:一是检索技术,二是检索效率的评价。

传统检索技术实现的方法多采用词切分、单汉字以及词切分和单汉字相结合;检索主要借助目录、索引和关键词等方法来实现。此技术的优点是简单快捷;但缺点是无法挖掘信息之间的内在联系,检索结果不能准确、全面反映用户的需求。

传统检索效率的理想要求是快、准、全。在保证
担,使用方便,易用性好。(2)通过学习、反馈等多种形式进行全过程的人机交互,有利于准确了解用户的真实检索需求,同时,在检索过程中,通过与用户进行交互,修正和调整检索策略。(3)突破了传统检索的关键词局限。仅靠关键词的出现与否进行文献取舍,易造成检出文献过多或检出文献过少的局面。(4)尤其适用目标文档中无确切查询用词,但有该词的同义词、近义词和关联词的情况。

当然,该模型也存在一定的缺陷,如后控概念词典需要人工辅助进行定期更新,新词与新的概念也需要借助人工辅助才能准确地加入概念网络等。以人工辅助方式进行知识库的更新,虽然能保证较高的质量,但耗时大,效率低,在一定程度上影响检索效率。同时,如果目标文档中有确切查询词时,查询效率不如基于关键词的检索方式。因而,下一步的改进是在

查全率(Recall)与查准率(Precision)前提下的快速3项指标作为对检索效果进行量化的评价标准,但是在海量的互联网上的数字图书馆信息检索上用查全率与查准率来衡量检索效率是否合适?在某些场合,高的查全率带来的成千上万条命中记录对用户实在是一个沉重的负担。

总之,传统的信息检索机制在数字图书馆中存在3个深层次的问题。这3个问题都与词汇紧密相关。这3个深层次的问题是:第一,“忠实表达”问题。很多情况下,用户很难简单地用关键词或关键词串

检索系统设计中把基于自然语言的概念检索与基于关键词的匹配检索有机结合起来,提高检索性能。

参考文献

- 1 张玉峰,晏创业.基于机器学习的知识检索模型研究.图书情报知识,2002(4)
- 2,3 张琪玉.情报语言学基础.武汉:武汉大学出版社,2001
- 4 李蕾等.基于语义网络的概念检索研究.情报学报,2000(5)
- 5 陈定权.web结构挖掘.情报理论与实践,2003(1)

金 燕 武汉大学信息资源中心博士生。通讯地址:武汉。邮编 430072。

张玉峰 武汉大学信息资源中心教授,博士生导师。通讯地址同上。(来稿时间:2003-07-18)

* 本文属国家社会科学基金项目“数字图书馆相关关键技术研究”(批准号:00BTQ004)的成果。