基于语义的搜索引擎探讨

项　珍
(浙江工商大学图书馆　杭州　310018)

摘　要　本文主要介绍了基于语义的搜索引擎的研究现状,分析了基于语义的搜索引擎的应用前景,探讨了基于语义的搜索引擎对图书馆的影响和图书馆的应对策略。

关键词　语义　搜索引擎　应用

随着互联网走进千家万户,搜索引擎在人们的工作、学习及生活中扮演着越来越重要的角色。综观整个搜索引擎发展史,可以看到自搜索引擎的鼻祖—Archie 创始以来,就一直不断地在改进,甚至更新换代。基于语义的搜索引擎的课题也随之应运而生。基于语义的搜索引擎以自然语言理解技术为基础,并且它将信息检索从目前基于关键词层面提高到基于知识(或概念)层面,对知识有一定的理解与处理能力。这就是所谓的概念检索。概念检索又包括同义扩展检索和相关概念联想检索两个方面。例如查询"实验动物"时,也能查询"小白鼠"、"兔子";查询"非典"时,一也能查询"SARS"、"传染病";查询"茅盾"时,也能查询"子夜"。概念扩展检索能够提高查全率,而相关概念联想检索则加强搜索引擎与用户间的交互。

1　基于语义的搜索引擎的研究现状

世界各国对基于语义的搜索引擎及相关理论作了大量的研究。在语义网络等语义表示理论、自然语言处理研究领域的语法理论等提出之后,关于自然语言信息检索,Jose Perez - Carballo 在其 natural language information retrivel : Progress Report 一文中,就对该领域的新进展进行了总结。在本体领域,出现了(Onto) 2Agent、OntoBroker、OntoSeek 等技术。而近年来概念检索领域较为突出的研究成果,如美国伊利诺大学与亚利桑那大学已经开发出基于美国国防高级研究署信息技术办公室(ITO)的国防科技项目研究报告摘要信息的主题概念空间(ITO Space)及其概念图(ITO Map),以及基于美国癌症医学数据库的癌症概念空间(Cancer Space)及其癌症概念图(Cancer Map)。国内的研究主要集中在对基于语义搜索引擎的特征和体系结构模型的探讨,以及基于概念检索的实现方法,如唐培丽就提出采用"以网对网"的方法来实现概念检索。

不过基于语义的搜索引擎的研究现状并不容乐观。在理论上,多数研究者认为自然语言处理技术、语义理解技术在信息检索领域应用的效果并不理想。绝大多数的文本内容的自动分析系统是基于统计模型。而这些方法不能胜任获取文本中存在的潜在的丰富的语义知识。在实践上,已有的搜索引擎还远远没有达到能够像人一样分析与理解自然语言语义的水平,而且在今后短时期内也达不到这样的水平。国外虽然有一些公司做出了基于概念的产品(如日本的 Justreseaarch 公司),但仅仅做到了语用层面,语义层面尚未

涉及。而对于中文搜索引擎来说,因为有着中文处理方面的问题,这方面的工作才刚刚开始。

2 基于语义的搜索引擎的应用

2.1 在信息检索领域的应用

搜索引擎是人们获取知识和信息的一个重要途径,其在信息检索领域有着极广泛的运用。搜索引擎的每一次更新换代,都大力地推动地信息检索技术的发展,对信息检索理论也有着深远的影响。迄今在信息检索领域有较大影响力的搜索引擎有百度和 google 等。通过对百度和 google 进行检索实验发现,检索结果不甚另人满意。结果如下表所示:

表 1 两者的检索用时及返回结果数量比较

检索用时及结果数量 搜索内容	百度		Google	
	用时	返回结果数	用时	返回结果数
桌面	0.001 秒	33,100,000 篇	0.03 秒	29,600,000 篇
苹果机	0.001 秒	556,000 篇	0.05 秒	573,000 篇
飞鸟集	0.001 秒	387,000 篇	0.04 秒	132,000 篇
泛蓝	0.001 秒	640,000 篇	0.08 秒	983,000 篇
brown	0.058 秒	5,800,000 篇	0.11 秒	363,000,000 篇
挪威的森林	0.060 秒	1,430,000 篇	0.03 秒	588,000 篇
清彩釉六方壶	0.001 秒	692 篇	0.12 秒	7,020 篇
通知到了	0.001 秒	43,600 篇	0.54 秒	1,950,000 篇
学习文件	0.001 秒	1,590,000 篇	0.19 秒	7,070,000 篇
养父母	0.001 秒	539,000 篇	0.41 秒	317,000 篇
你把手抬高点	0.001 秒	95 篇	0.34 秒	1,190,000 篇
调查污染严重的企业	0.001 秒	165,000 篇	0.33 秒	2,920,000 篇
天津大学排名	0.046 秒	180,000 篇	0.23 秒	1,250,000 篇
杭州最近三天的天气	0.093 秒	1,630 篇	0.23 秒	1,300,000 篇
开放系统功能理论的基础	0.232 秒	83,900 篇	0.38 秒	1,930,000 篇

表 2 两者的检索返回结果的相关度比较(top20)

搜索结果 top20 / 搜索内容	百度	google
桌面	总相关度为 90 %	总的相关度为 85 %
苹果机	总相关度为 50 %	总的相关度为 65 %
飞鸟集	总相关度为 65 %	总的相关度为 60 %
泛蓝	总相关度为 80 %	总的相关度为 85 %
brown	相关度为 75 %	相关度为 70 %
挪威的森林	总相关度为 85 %	总的相关度为 80 %
清彩釉六方壶	总的相关度为 55 %	总的相关度为 20 %
通知到了	总的相关度为 90 %	总的相关度 75 %
学习文件	总的相关度为 80 %	总的相关度为 75 %
养父母	总的相关度为 100 %	总的相关度为 100 %
你把手抬高点	总的相关度为 10 %	总的相关度为 55 %
调查污染严重的企业	总的相关度为 45 %	总的相关度为 50 %
天津大学排名	总的相关度为 50 %	总的相关度为 85 %
杭州最近三天的天气	总的相关度为 5 %	总的相关度为 40 %
开放系统功能理论的基础	总的相关度为 5 %	总的相关度为 5 %

对表 1 的小结:

(1) 英文词及包含关键字较多的句子查询的检索用时较长。

(2) 输入同一查询内容,google 的检索用时相对较长。

(3) 输入单个的中文检索词和短语作为查询内容,百度和 google 的返回结果的数量不分上下。输入英文词,则 google 的返回结果的数量比百度要多。

(4) 对于句子查询,google 的返回结果的数量比百度要多一些。

(5) 上述的检索结果是在特定的电脑配置及网络配置条件下取得的。

对表 2 的小结:

(1) 表中提及总的相关度,这里的相关包括检索结果与检索词/短语/句子的不同义项的相关。搜索引擎(包括百度和 google) 将不同义项的相关结果都返回,可见其无法准确识别用户查询的具体需求。

(2) 检索结果中所提供的相关网站的链接,并不能直接显示与查询相关的内容,有时需多次点击网站上的链接,方能找到相关内

容,甚至也可能找不到。

(3)句子查询的结果相关度低,有时网页只含有检索内容中的若干个字或词,就被收进检索结果中来。在句子查询方面,google 的检索结果明显优于百度。

(4)两者的检索都出现不同程度的有若干条返回结果打不开的现象。

(5)两者的检索结果中都会出现重复现象。由第 4 条及第 5 条可知搜索引擎未能将检索结果做很好的结果处理,至少结果去重工作做得不够。

(6)检索结果并不是完全按照相关度的大小来排序。往往检索结果中是用户点击量较大的网页排在前面。

(7)检索结果不够清晰和直接,几乎只是"裸数据",结果的进一步利用还需用户进行选择和判断。

(8)在检索结果的 top20 中几乎没有出现检索输入在内涵上的同义词或近义词,尤其是句子检索。相关搜索基本只是字面上的相关。

而最新一代的搜索引擎中较典型的有两个,都是基于本体的语义检索系统,分别是:SWOOGLE——语义网中的基于蜘蛛网的检索系统,系统从每个搜索到的文本中抽取本体,根据本体之间的相关度来比较文本之间的关系;TUCUXI(InTelligent Hunter Agent for Concept Understanding and LeXical ChaIning),该系统根据查找的本体在网页上爬行,决定哪种网页最满足需求[1]。其他的搜索引擎中智能化程度较高的搜索引擎还有国内的尤里卡、问一问,国外的如美国的 AskJeeves 等。

搜索引擎不断往智能化方向发展给信息检索用户们带来了惊喜:搜索引擎更"聪明"了;搜索引擎使用起来更方便了;检索结果更令人满意了。而基于语义的搜索引擎一旦实现,信息检索的效果会空前优化,人们的信息需求也会得到空前满足,甚至人们的信息检索行为和检索方式也会发生变化。

2.2 在自动问答系统领域的应用

自动问答系统也是快速准确获取信息的最好途径之一。自动问答系统能从大量数据中找出问题的答案。其答案往往是一个短语或一个句子,而不是文本。对于自动问答系统的研究,国外开展得比较早,国内才刚刚开始,研究的机构也很少,目前还没有成型的中文自动问答系统。自动问答系统包括三个主要部分:问题理解、信息检索和答案抽取。信息检索部分的任务就是用前面提取出来的关键字到文档库中查找相关的文档并返回一些最相关的文档。该部分就需要使用问题的搜寻匹配技术,这与搜索引擎的相关性算法是一致的。那么在自动问答系统中的信息检索模块可以直接调用已有检索系统或者也可调用搜索引擎。由此可见,自动问答系统领域不仅可以借鉴基于语义搜索引擎的发展,甚至可以将基于语义搜索引擎直接运用于自动问答系统中。如有一个简单的专业领域的自动问答系统就直接使用了 lucene 这样一个搜索引擎。

2.3 在信息服务中的应用

首先,基于语义搜索引擎与信息服务的若干阶段都不无关联,不可否认它促进了信息处理技术往自动化、智能化方向发展,其所蕴涵的知识库中的丰富而完善的知识更为信息提供服务创造了一个"源头活水"。由此也提升了信息服务的层次的提高:由信息层面转向知识层面。其次,基于语义搜索引擎推动了个性化信息服务的发展。个性化服务表现为两个层次:第一层次为按用户要求进行信息订制,可以让用户根据自己的需要订制专门的信息。第二层次则是挖掘用户兴趣模

式,主动提供服务,使信息服务机构成为一个智能型、主动性的信息提供者。这第二层次就离不开将基于语义搜索引擎中的用户模型应用于个性化信息服务。

2.4 在自然语言处理和机器翻译领域中的应用

搜索引擎和机器翻译这两个领域本来是风马牛不相及的,但由于高质量的机器翻译系统必须结合语言学知识以及常识(客观世界中性知识),而常识恰好是搜索引擎知识库的重要内容,并且搜索引擎领域对知识库的研究已经相当深入。知识库中有着丰富的概念和层次化、结构化的概念联系。那么通过把某种语言中的词汇映射到知识库中的概念,可以支持在源语言分析时进行歧义消解和目标语言生成时的词汇选择,并可作为源语言和目标语言之间中介表示的概念来源。这就是搜索引擎系统在机器翻译领域的应用。

另外,自然语言处理的相关理论同为搜索引擎领域和机器翻译领域的理论基础之一,自然语言处理技术的成熟与否也直接影响了这两个领域研究成果的效果。虽然它们对于自然语言处理的研究都已经有一定成效,但由于汉语语言的复杂性和多样性,对于汉语语句及文档理解的准确性还很不够,还需要这两个领域在自然语言处理研究方面共同努力。这两个领域的研究是相辅相成的,可以说基于语义搜索引擎领域对于自然语言处理的研究成果可以直接渗透、应用于机器翻译领域。

3 基于语义搜索引擎对图书馆的影响、及图书馆的应对策略

3.1 基于语义搜索引擎对图书馆的影响

随着搜索引擎的出现,信息搜索已经逐渐成为竞争性的商业市场,以往图书馆在其中只扮演了参与者的角色。用户完全有机会去选择使用搜索引擎之类的检索工具。用户喜欢用新的工具做各种类型的信息检索,特别是成长起来的年轻一代。根据 Bielefeld University 在 2002 年 11 月所做的一项调查,学生们仍会使用图书馆的在线目录,但他们更喜欢通过类似于 google 的界面进行检索。之所以还使用图书馆在线目录,是因为他们找不到可以替代的工具。毋庸置疑,搜索引擎的使用非常流行[2]。

搜索引擎的流行使用给图书馆带来严重的冲击,不仅取代了图书馆的部分信息职能,甚至连图书馆向来较有优势的学术搜索服务也受到威胁。以前,由于知识产权及技术方面的原因,搜索引擎不能覆盖深层 web(“deep web”or“invisible web”)。所以书籍、电子期刊及数据库的检索是它的“软肋”。而目前若干优秀的搜索引擎也纷纷推出学术搜索服务。如 google 就推出了免费的 Google 学术搜索的服务,其搜索范围包括专家评审文献、论文、书籍、预印本、摘要以及技术报告等内容。作为这项服务扩展的一部分,Google 学术搜索在索引中涵盖了来自多方面的信息,信息来源包括万方数据资源系统、维普资讯、公开的学术期刊、中国大学的论文以及网上可以搜索到的各类文章等。同时 Google Scholar 还提供中文版界面。

这样看来,用户各种类型的信息需求很大程度上通过使用搜索引擎都能得到满足,并且搜索引擎有其自身优势,信息丰富、更新及时,使用起来快捷方便。随着基于语义的搜索引擎的发展,搜索引擎的功能会更强大,服务会更人性化,它的优势还在不断扩大。站在图书馆的立场看,竞争形势不容乐观。图书馆的信息用户在流失,图书馆的信息职

能似乎可以被搜索引擎所替代,图书馆在信息产业中的地位也受到动摇,甚至有人开始怀疑图书馆还有无存在的必要。这种情况不能不引起图书馆的重视和反思。

3.2 图书馆的应对策略

面对搜索引擎带来的冲击和影响,激烈的竞争形势逼迫图书馆去积极应对这种影响。同时,图书馆有必要抓住当今社会信息技术和网络技术突飞猛进的机遇,增加信息服务的知识含量,提升自身在信息搜索市场中的综合竞争力,用实力证明"图书馆衰退论"及"图书馆消亡论"是不符实际的。

3.1.1 扬己所长,用户至上。图书馆经过长年的积累,其信息资源优势是其他部门所无法企及的。图书馆要好好管理和深加工这些资源,将它们的作用尽可能发挥到极至。面对读者新的需求,图书馆要站在用户的角度看问题。因此,以用户为中心,个性化、特色化的主动信息服务方式势在必行。馆藏数字化建设比较好的图书馆,可以把"mylibrary"系统及服务平台发扬光大,给用户创建一个合意的个人信息门户。相对数字化实力较弱的,也要尽可能多培养些学科馆员,把信息虚拟咨询的服务工作做好。

3.1.2 将研究成果应用于实践。把基于语义的搜索引擎的研究成果应用到图书馆的信息检索中去。图书馆一直参与着语义检索的研究,因此图书馆没有理由不把研究成果加以应用,开发属于自己的基于语义词表的语义检索系统,甚至是更高级别的基于本体、知识库的语义检索系统。事实上这不是设想,而是已经有了成功的例子。如美国国立医学图书馆就有一个被称之为"癌症数字图书馆"的检索系统,它就是基于 MESH 词表和 UMLS 的语义检索系统。该系统能正确地表述出现在文献中的与查询词相关的词语,并能把它们分成不同类型的术语。如果说这还不是完全的语义检索,相信离真正的语义检索系统的出现也不会太远了。

图书馆要构建自己的语义检索系统,首先要明确基于语义的搜索引擎(检索系统)并不等于基于语义网的搜索引擎。图书馆在构建本体知识库的时候不一定要依赖或利用语义网的本体。其次,图书馆在实施的时候要量力而行,不妨先从"垂直型"的语义检索系统做起。因为一来主题范围小,实施起来相对较为容易;二来着眼于"窄口径"也容易创出自己的特色。再次,语义检索系统开发的关键——知识库的构建有多种模式,例如本体构建是要参考现有本体还是重新构建新的,知识库构建要达到哪一个层次,图书馆必须要找准定位,选择适合自己的本体、知识库开发模式。

基于语义的搜索引擎代表了搜索引擎发展的一个方向,其基于语义的优势和作用在今后还会逐步显现,并切实地影响着人们的工作、学习和生活。当然,基于语义的搜索引擎的研究还有很多工作要做,并且是项涉及多学科多领域的浩大工程,需要多方的共同努力。特别是我国在这方面的研究比较薄弱,作为重要信息传播机构的图书馆更应责无旁贷地积极参与其中,为我国的基于语义的搜索引擎的研究和发展尽绵薄之力。

参考文献

[1] 语义搜索引擎综述. [EB/OL]. [2008 - 09 - 22]. http://xchuspace.spaces.live.com/blog/cns!3f3f394f3a76da53!115.entry, accessed in Oct. 11,2005.

[2] Norbert Lossau. Search Engine Technology and Digital Libraries[J]. D - Lib Magazine. 2004,(6).

(责任编辑:张根彬)