第三代搜索引擎研究

俞平[1] 肖南峰[1] 甘志刚[1]

摘要

搜索引擎是从互联网获取信息不可或缺的工具. 第一代搜索引擎的特征是目录式搜索,以雅虎为代表; 第二代搜索引擎的特征是关键字搜索,以 Google 为代表. 随着时代发展,信息急速膨胀,第二代搜索引擎已经逐渐不能满足人们的需求; 于是,人们开始期待第三代搜索引擎的到来. 什么是第三代搜索引擎? 这个问题目前还没有结论. 分析了与这个问题相关的有代表性的观点和尝试,并提出了 Search Engine Service (搜索引敬服务)的概念,认为它是第三代搜索引擎不可缺少的组成部分.

关键词

第三代搜索引擎;Search Engine Service

中图分类号 TP242.6

文献标志码 A

0 引言

Introduction

互联网一直以惊人的速度在膨胀,1999 年时约有 3.5 亿个网页,并以每天 100 万页的速度增长. 如今,主流的搜索引擎索引的网页数大都在十亿到百亿,而这个数字还不到总网页数的十分之一. 在“信息爆炸”的时代,如果没有合适的工具,很难从互联网中提取完整有用的信息,这个“合适的工具”就是搜索引擎. 调查[1]显示,有 68.2% 的人经常使用搜索引擎,是仅次于电子邮件的网络应用;有 41% 的人通过搜索引擎进入购物网站;84.6% 的新网站是通过搜索引擎被发现的.

1994 年,雅虎仿照图书馆管理的方式,推出目录式搜索. 目录式搜索通常以人工或半自动方式搜集信息,由编辑查看后,人工形成信息摘要,并将信息置于事先确定的分类框架中,信息大多面向网站,提供目录浏览服务和直接检索服务. 尽管在今天看来,这种方式简单、甚至有些笨拙,但在当时却被奉为至宝. 第一代搜索引擎以目录式搜索为主要特征,雅虎被认为是第一代搜索引擎的代表. 1998 年,Google 登场,令所有人眼前一亮. Google 的核心技术是基于关键字的搜索、网页自动抓取、页面重要性分析、超链分析,大大提高了搜索效率. Google 等搜索引擎中,一个被称为蜘蛛(spider)的机器人程序以某种策略自动地在互联网中搜索和发现信息,这些信息由索引器建立索引,检索器根据用户的查询输入检索索引库,并将查询结果返回给用户. 第二代搜索引擎以关键字搜索为主要特征,Google 被认为是第二代搜索引擎的代表.

今天,人们普遍感觉第二代搜索引擎已经不适应信息的急速膨胀. 一些观点认为,像 Google、百度等主流的第二代搜索引擎仅能搜索网页(以及部分文档,如 doc、ppt、pdf)中的信息,而面对同样包含大量信息的图像、视频、音频等多媒体数据则束手无策;此外,对于组织内部特定数据库、计算机辅助设计系统中的数据等[2],同样无能为力. 更多的观点则认为,目前搜索引擎普遍存在的问题是:返回的结果过多,许多结果“答非所问”;关键字的选择需要更多的技巧、反复的尝试;“查全”和“查准”是人们对于搜索引擎的基本要求,但已经越来越

收稿日期 2009-06-01

资助项目 国家自然科学基金与中国民用航空总局联合资助(60776816);广东省自然科学重点基金(251064101000005)

作者简介

俞平,男,硕士,工程师,研究方向为计算机应用技术.

肖南峰(通讯作者),男,博士,教授,研究方向为计算机应用. xiaonf@ scut. edu. cn

1 华南理工大学 计算机科学与工程学院, 广州,510006

得不到满足.

当第二代搜索引擎的弊端越来越明显的时候,下一代搜索引擎已经开始酝酿,针对第三代搜索引擎的研究正在展开,百度、中国搜索、Google、微软等厂商也进行了各种尝试. 那么,到底什么是第三代搜索引擎?

1 各方观点

Various views

针对什么是第三代搜索引擎这个问题,出现了各种各样的意见. 以下列举一些具有代表性的观点. 文献[3]认为,第三代搜索引擎应该具有以下 3 个特征:智能化、社区化、个性化. 其中,"智能化"是指搜索引擎必须要引入人工智能技术,尝试去理解用户的查询意图,并优先显示用户需要的结果. 目前比较可行的智能化技术包括智能纠错、分类和联想等. "个性化"是指搜索引擎必须要考虑到用户的个性化需求,不仅要给出符合不同用户需求的不同结果,连搜索结果的界面都应该有所区别. "社区化"则是指每个用户的搜索结果都可以存储并能和其他人分享.

文献[2]则认为,第三代搜索引擎的目标是从数据库、网页、文档或音频和视频剪辑中自动提取信息;识别人名、地点、组织、日期、金额并且寻找其中的关联性;同时随着企业拓展呼叫中心并转向基于 IP 电话系统、以及政府在智能化及国家安全信息技术上投入巨资,挖掘声音和图像的含义. 文献[4]介绍了第三代搜索引擎是"基于句子、短语的内容搜索. 这几乎是一个穷尽搜索需求的境界. 也就是说不可能再产生第四代搜索技术了."文献[5]从两个不同角度分析了第二代搜索引擎的缺陷,从而得出第三代搜索引擎的未来发展之路. 第一,从互联网资源组织的角度,第二代搜索引擎的核心概念"关键字",仅仅是出现在网页中的符号,它本身的语义没有被使用;同时,页面分析只是依据页面间的链接关系,没有得到页面本身包含的信息. 下一代搜索引擎必须能够表达和处理语义信息,其数据模型必须是语义数据模型. 下一代搜索引擎将是智能化的. 第二,从用户信息体验的角度,下一代搜索引擎应该克服"千人一面"的分类体系和页面内容,引入查询修正、查询结果聚类等技术,为用户营造个性化空间. 下一代搜索引擎将是个性化的.

文献[6]分析了第三代搜索引擎的智能化需求,提出了两方面的"智能检索":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,能够运用分词技术、同义词技术、概念搜索、短语识别以及机器翻译等技术;能够通过分析检索者的检索和浏览行为来学习检索者的需求,利用搜索引擎的现有服务有选择地为检索者提供个性化的检索服务. 文献[7]则着重分析了第三代搜索引擎的人性化发展趋势,认为应当根据用户知识水平、专业、爱好、心理倾向、行为方式等的不同,来提供多层次个性化的信息服务模板,实现"用户满意"和"用户快乐". 文献[8]提出第三代搜索引擎 5 点技术趋势.

1) 基于知识的搜索. 解决分词、同义词、概念搜索、短语识别以及机器翻译技术等问题,将基于关键词的搜索提高到基于知识的搜索层面上来,以提高搜索的智能水平;信息服务更智能化、人性化,允许自然语言检索.

2) P2P 对等网络. 大大提高搜索深度,不受信息文档格式和宿主设备限制.

3) 元搜索引擎. 对多个普通搜索引擎的结果进行分析综合.

4) 元数据. 利用 DC 元数据,有效识别有价值的可靠信息

5) 信息组织和检索规范化控制. 实现信息组织自动化、标准化、兼容化.

文献[9]指出,如何从庞大的资料库中精确地找到正确地资料,被公认为是下一代搜索技术的竞争要点. 而"智能搜索"可以通过对搜索内容相关性的自动学习,来提高搜索结果的准确度. 文中还提到另一个颇受瞩目的搜索技术——P2P,它通过共享硬盘上的文件、目录乃至整个硬盘,无需通过 Web 服务器,不受信息文档格式的限制,即可达到传统搜索引擎无可比拟的深度. 文献[1]的观点比较独特,认为第三代搜索引擎的最大特点是大量采取人工介入,实现人工和技术的完美结合,以提高搜索水平. 文献[10]相信,第三代搜索引擎即"智能搜索引擎". 它将基于关键字层面检索的传统搜索引擎提高到基于知识(或概念)层面来分析、处理检索提问,表现出较强的智能化与个性化特色,它以一定的知识库技术为基础,具有很高的自然语言理解与知识处理能力.

文献[11]认为,第三代搜索引擎应当在术语向量数据库之上添加词干(Word Stemming)和辞典,以保证搜索过程不脱离上下文. 同时关键字对(Keyword Pair)的自动提取有助于页面的自动分类,使得包含一些特殊关键字的搜索能根据上下文和用户意愿得到完全不同的结果. 此外,还应增加网络地图,以消除由于同义词、近义词导致的重复. 最后,第三代搜索引擎应当根据用户的搜索习惯,获得尽可能多的信息,通过一段时间的使用,搜索引擎应当"逐渐了解"用户. 文献[18]还提出了"主题引擎"的概念:通过术语向量数据库,权衡页面关键字密度以计算页面向量,页面向量将根据术语向量进行比较和储存. 接着通过链接等相关性计算页面权重(Reputation),以确保页面内容与权重吻合,最相关的吻合能得到最高的搜索排名.

2 业界动态

Trends in IT field

在IT领域,企业一直是技术发展和实用化的重要推动力量. 企业的观点,往往代表用户的切身需求、以及正确的发展趋势.

2.1 中国搜索

2003年,号称"第三代搜索引擎"的中国搜索宣告成立,业界普遍认为,中文搜索市场已形成百度、Google、中国搜索三足鼎立的局面. 但将中国搜索称为"第三代搜索引擎"还言之尚早. 在近年出现的所谓"第三代搜索引擎"中,还没有出现类似Google的"关键字搜索"那样明显的技术取代趋势,相对于第二代搜索引擎,还没有取得重大突破. 但从它们身上,人们的确看到了许多明显不同于第二代引擎的特征.

2004年12月,中搜推出的网络猪3.0被认为是新一代搜索引擎、个性化信息门户的开始. 2006年4月18日,中搜发布了客户端搜索软件IG(Internet Gateway),IG试图改变第二代搜索软件中用户获取搜索结果的方式——从每次都要通过关键词搜索自己想要的内容,到自己设置想要的内容目录,让这些内容每天自己送上门来. 中搜认为,这就是相对第二代搜索引擎发生质变的第三代搜索引擎. 用户可以通过"IG"的桌面软件定制内容,如新闻、股评、博客、天气预报和邮箱中尚未阅读的新邮件等,定制的内容也可以切换,白天用商务版,晚上换成娱乐版——不用再到每个网站去寻找,所有这些都会自动呈现在IG里. 简单的说,可以把IG看成一个听话的软件,它会根据设定的关键字,到互联网上把一切符合要求的东西都抓取过来,而不用你到每个网页上去辛辛苦苦地寻找. 目前这种技术或许还没有成熟到一定可以取代"第二代搜索引擎",但以个性化需求为本的想法已经体现. 在文献[5]一项针对"智能化"与"社区化"的测试中,中搜战胜了Google、百度、一搜、爱问和搜狗,成为其中"最像第三代搜索引擎"的一个.

2.2 微软

为对抗Google,微软决定把搜索技术"大脑"设到中国. 2005年10月28日,微软亚洲研究院成立互联网搜索技术中心,以加快第三代智能互联网搜索技术的研发. 微软亚洲研究院沈向洋博士指出,第三代搜索引擎要对整个网页做一种分析和数据挖掘,不仅是找到更多的结果,而要更加智能化、人性化、更加精确、能够理解用户需要什么结果,然后进行聚合和整理[12]. 微软认为:搜索引擎的用户界面将有重大变化,用户将不再仅在一个文本框中输入搜索关键字. 用户会提供更多信息,让搜索引擎更清楚的领会用户的意图,以便返回更准确的结果[13]. 在个性化方面,微软的官方网站上已经开始提供Windows Live测试版的两项中文服务,其中Windows Live Today功能可以搜集当天最新的消息[4].

2.3 Google

Google反对微软的看法,认为技术的进步意味着用户无需再提供更多的信息;完美的搜索引擎将准确地理解用户的用意,并提供用户所需要的信息[13]. 个性化方面,Google推出了个性化主页. 用户可以对Google的主页进行一定程度的定制,设定需要的服务(如天气、便条、代办事项)和感兴趣的新闻内容.

2.4 雅虎

雅虎认为,第三代搜索引擎的重大创新将是"社会搜索",它能够实现信息相关度的民主化,普通用户来决定对于他们和其他用户而言什么是重要的[13].

2.5 百度知道与新浪爱问

百度知道和新浪爱问的思想与雅虎的社会搜索有异曲同工之处,即让用户有越来越多的参与权,直至成为网络信息提供者,而不仅仅是被动的接受信息,这符合Web 2.0的核心思想. 与传统的搜索引擎不同,百度知道并非直接查询那些已经存在于互联

网上的内容,而是用户自己根据具体需求有针对性地提出问题,通过某种悬赏机制发动其他用户,来创造该问题的答案. 同时,这些问题的答案又会进一步作为搜索结果,提供给其他有类似疑问的用户,达到分享知识的效果.

2.6 Cgogo. com

Cgogo. com 提出"概念集群"和"动态分类",并认为这是"第三代搜索理念的具体体现"[14]. 其中,"概念集群"指的是第三代搜索引擎能够模仿人的一些思维和想法,是概念的模糊搜索,而不仅仅是"关键字"这样的文字符号. "动态分类"则通过分析网页之间的关联,建立一种类似人的思维的更智能化的概念分类方式,通过模仿人的思维模式,对要查找的概念进行关键字联想和分类.

2.7 Lexxe

Lexxe 将自然语言处理加入了自己的搜索引擎. 在 Lexxe 搜索引擎中实现了一种"双语计算":关键字首先作为语言进行匹配,然后才作为符号进行匹配. 通过这种技术,Lexxe 能够获得比传统第二代搜索引擎更准确的结果

3 总结

Summary

经过前面的分析,可以总结出第三代搜索引擎的一些发展趋势.

3.1 智能化

第三代搜索引擎将是智能化的、将融入更多人工智能的成分,这个观点已被普遍接受,"搜索技术领域的领先者都认为,最终的搜索引擎将是智能化的,能够理解世界上的所有事物"[10]. 智能化包括两方面的内容:搜索引擎在分析网页时,不应将网页仅仅作为一组字符,而应当理解其中的语义,提取如主题思想、类别等的自然语言层面的信息;搜索引擎在接收用户输入时,应当智能化处理,尝试精确理解用户的意图、纠正用户输入错误、分类导航、联想等,从而返回更准确的结果.

3.2 个性化

第三代搜索引擎将突破目前"千人一面"的用户界面:用户可以定制搜索引擎的用户界面,使之更符合个人的使用习惯;用户可以订阅信息,即搜索引擎根据用户的设定,主动获取所需信息(主要是新闻类)提供给用户;搜索引擎应当分析用户的行为,或根据用户的设定,给予页面不同的权重,即不同类型的用户,搜索结果排序不同.

3.3 社区化

社区化主要体现 Web 2.0 的思想:用户成为网络资源提供者. 用户将更多地参与页面权重的设定、共享搜索结果、优化搜索结论等.

3.4 信息源多样化、专业化

第三代搜索引擎信息的来源不仅仅是网页,还应当包括各种类型文档、图像、音频、视频等多媒体、特定数据库、各种 ERP、CRM、CAD、CAM 中的数据. 其中,多媒体信息的理解和搜索被认为在许多领域都具有重大意义. 除了通用搜索引擎,专业搜索引擎应当能够深入理解该领域的知识,从而为有特殊需求的用户提供更准确的搜索结果.

3.5 其他

P2P 模式的搜索引擎,搜索将不再通过集中式服务器,而是在用户的信息交互中实现. 听起来就像 BT 下载一样激动人心. 有人认为这种模式将取代目前 C/S 模式的搜索引擎,也有观点则认为将成为有益补充. 元搜索引擎,进一步处理其他搜索引擎的结果使之更准确,一些观点认为第三代搜索引擎将建立在元搜索技术基础之上,从而实现更高级的智能化和个性化.

3.6 结论

到底什么是第三代搜索引擎? 这个问题依然没有结论. 目前的各种观点和尝试,无论正确与否、成功与否,都将推动技术的发展. 现在要回答什么是第三代搜索引擎还为时过早,但可以相信,科技发展一日千里,得出结论的那一天并不久远.

4 搜索引擎服务

Search Engine Service

总结各种观点后,本文认为:第三代搜索引擎应当突破现有模式,不仅仅为"人"服务,同时也应当为"计算机应用程序"服务. Web Service 在近年逐渐兴起,并得到广泛的接受和应用. Web Service 突破了原有 http、www 的模式,利用 http 作为应用程序间通信的手段. 受到 Web Service 的启发,笔者提出 Search Engine Service(搜索引擎服务)的概念.

目前,能够为应用程序所用的搜索引擎,最著名的是 Google. 通过 Google API,应用程序可以和使用 www. google. com 一样,输入关键字,返回搜索结果.

但这种形式与传统搜索引擎模式并没有本质的不同,仅仅是改手工搜索为编写应用程序自动搜索.笔者认为,Search Engine Service 是第三代搜索引擎不可或缺的组成部分,它具备以下功能:

1)提供应用程序接口(API),使得程序能够以简洁的方式使用搜索引擎提供的服务,这一点与 Google API 类似;

2)搜索结果以"知识"的形式返回,这隐含了 Search Engine Service 的一个本质特征——智能.

这里的"知识"是指人类高层次的认知、技能、经验等在计算机系统中的表达.

许多人工智能的应用,如专家系统,往往由于"知识"的匮乏而效果不理想甚至失败;但与此同时,网络上的各种信息又是极其丰富.笔者认为,就像当今时代日常生活与网络已密不可分,人工智能与网络同样密不可分,一方面人工智能用于理解网络上的信息而获得"知识";另一方面这些"知识"又使得人工智能系统更加"智能".这就像现代的人类,通过网络不断学习、不断积累,而越来越聪明、能干.要实现 Search Engine Service,必需:

1)人工智能发展到一个较高的水平,能够理解人类自然语言并提取"知识",能够运用"知识"指导计算机的行为;

2)计算机能够在解决问题出现困难时发现"知识"的不足,并按照 Search Engine Service 的规范生成搜索串,从而使用 Search Engine Service 获取所需的"知识";

3)当 Search Engine Service 得到多个结果时计算机应当具有选择提炼能力,以及结果不理想时修改搜索串的能力;

4)计算机在解决问题后,能够提取"经验",并转换成自然语言撰写在博客、社区等中,使网上的"知识"越来越丰富.

显然,以目前的技术水平,这样的需求难以实现.但笔者相信,Search Engine Service 将是未来搜索引擎技术和人工智能技术发展的必然结果,将在搜索引擎和人工智能系统中扮演重要角色.

5 结尾

Concluding remarks

以 Google 为代表的第二代搜索引擎已经逐渐不能适应时代的发展,第三代搜索引擎的呼声渐高.对于什么是第三代搜索引擎这个问题,本文列举和分析了目前具有代表性的观点、看法、尝试,总结出第三代搜索引擎应当具备的特征,并提出一个全新的概念:Search Engine Service. Search Engine Service 需要高级的人工智能,用于在互联网上搜索,并转换成"知识"供人工智能系统使用.本文认为,Search Engine Service 是第三代搜索引擎不可缺少的组成部分.

参考文献

References

[1] 智能搜索:融合搜索技术与人类专业知识[J].计算机与网络,2003(10):24-25
Intelligent search:Combine search technology with human specialized knowledge[J]. China Computer & Network,2003(10):24-25

[2] 重磅炸弹:第三代搜索技术展望[EB/OL].[2006-07-24].http://dotnet.csdn.net/n/20060724/92887.html
Heavy bomb: Prospects for 3G-search engine technology[EB/OL].[2006-07-24]. http://dotnet.csdn.net/n/20060724/92887.html

[3] 胡坤.等待第三代搜索引擎[J].电子商务世界,2005,8:40-44
HU Kun. Waiting for the 3G-search engine[J]. Electronic Business World,2005,8:40-44

[4] 克丽.第三代搜索引擎现形[N].每周电脑报,2006-02-13(9)
KE Li. 3G-Search engine coming forth[N]. Computer Weekly,2006-02-13(9)

[5] 第三代搜索引擎初探:智能化、个性化[EB/OL].[2006-07-21].http://dotnet.csdn.net/n/20060721/92852.html
Preliminary probe into 3G-search engine:Individuation and Intellectualization[EB/OL].[2006-07-24].http://dotnet.csdn.net/n/20060721/92852.html

[6] 傅欣.第三代搜索引擎的智能化趋势研究[J].现代图书情报技术,2002,97(6):28-30
FU Xin. Studies on Intelligent trends in 3G-search engines[J]. Modern Technology of Library and Information Service,2002,97(6):28-30

[7] 李苏华,李建伟.论搜索引擎的人性化发展趋势[J].中山大学学报,2005,25(1):260-263
LI Suhua, LI Jianwei. On the humanistic development trend of search engine[J]. Supplement to Journal of Sun Yatsen University,2005,25(1):260-263

[8] 李建伟.试论第三代搜索引擎的技术发展趋势[J].广西右江民族师专学报,2004,17(3):55-57
LI Jianwei. On the development trend of 3G-search engine technology[J]. Journal of Guangxi Youjiang Teachers' College for Nationalities,2004,17(3):55-57

[9] 刘艳.搜索下一个 Google[J].互联网周刊,2003(4):38-41
LIU Yan. Search for the next google[J]. China Internet Weekly,2003(4):38-41

[10] 胡誉耀.智能搜索引擎与知识共享[J].中国信息导报,2003(11):52-55
HU Yuyao. Sharing smart search engine and knowledge[J]. China Information Review,2003(11):52-55

[11] The Future of search engine optimizing:theme engines,robinno-

bles[EB/OL]. (2009-03-24). http://www.searchengineworkshops.com/articles/se-optimization-future.html
[12] 微软卡位第三代搜索技术:Google 很快过时[EB/OL]. [2006-07-24]. http://dotnetcsdn.net/n/20060724/92889.html
Microsoft aims at 3G-Search engine, google will be soon outdated [EB/OL]. [2006-07-24]. http://dotnet.csdn.net/n/20060724/92889.html
[13] 搜索的未来:雅虎, Google 和微软的观点[EB/OL]. [2006-07-20]. http://news.csdn.net/n/20060720/92796.html
Search in the future: Viewpoints of yahoo, google and microsoft [EB/OL]. [2006-07-20]. http://news.csdn.net/n/20060720/92796.html
[14] 程天宇. 第三代搜索引擎在哪里[J]. IT Time Weekly, 2005, 15(8):71
CHENG Tianyu. Where is the 3G-search engine[J]. IT Time Weekly, 2005, 15(8):71

Research on 3-Generation Searching Engine

YU Ping[1] XIAO Nanfeng[1] GAN Zhigang[1]
1 School of Computer Sci and Eng, South China University of Technology, Guangzhou 510641

Abstract The search engine is an indispensable tool to get information from Internet The characteristic of 1-Generation Search Engine is based on catalogs, the typical example of which is Yahoo. The characteristic of 2-Generation Search Engineer is based on key words, of which Google is very typical. With the development of the era and expansion of information, the 2-Generation Search Engine nowadays can not meet people' requirements, so it is expected for the 3-Generation Search Engine to come into being. What the 3-Generation Searching Engine really is still remains a question without a conclusion. This paper analyzes typical opinions and attempts related to the problem and then proposes our points.

Key words 3-Generation Search Engine; Search Engine Service