对搜索引擎未来发展的探讨

Research for the Future Development of Search Engine

赵仲孟 张 蓓 沈均毅

(西安交通大学计算机系 西安710049)

Abstract Search engine has become an absolutely necessary tool of information retrieval on Internet, but in present, the search efficiency and quality for search engine are not very good. How to make search engine more effective, this paper analyses the questions in two ways and looks into the future direction for search engine.

Keywords Search engine, Information retrieval, Network

由于 Internet 网络的普及,网络已成为人们获得信息的必要途径和重要手段,但是网络上的信息浩繁,内容庞杂,组织松散,人们常常苦于查找有用信息,反而使网络丧失了信息获取的便捷性,所以搜索引擎成为人们在网检索信息的必需工具。

但是目前搜索引擎查询效率和查询质量并不高,本文就如何让搜索引擎更为有效入手,从两方面来分析问题所在,并展望搜索引擎未来可能的发展方向。

1 让搜索引擎更了解你的需求意图

当前的信息检索服务大多是基于关键字查找,并运用布尔运算。由用户给定需求信息中的关键词语,系统在数据库中查找包含关键词语的文件,并给出相应的网络地址连接,这种方式将词语从语义背景中提出,以孤立的关键字标引文档,割裂了原始文档中的逻辑语义,只用文档中的词组不能完全表达文档的含义,更重要的是在自然语言中,有很多方法可以表达一个询问的概念,同时一个词可以运用到多个领域,所以关键字方式脱离了用户本身意图的领域背景,使得“误解”概率增大,导致检索结果与用户意图不符。

我们认为搜索引擎不能很好地了解用户的意图是目前导致搜索引擎效率低的根本原因之一,解决的方法之一是让搜索引擎更了解用户。

1.1 让用户更好地表达需求

现有的关键字接口不符合人们的日常习惯。随着网络的普及,方便用户的使用也会成为一个系统成功的关键。自然语言接口的提供可以让用户以自然和熟悉的方式使用搜索引擎,而且自然语言能更好地表达用户的意图和需求,使检索的结果更加准确。

1.2 主动让搜索引擎了解用户

搜索引擎的定制:个性化是未来商业服务的一个发展趋势,用户有各种背景,虽然两个用户输入同一关键字,但他们想要的结果可能相差千里,要明确用户的意图,就应该了解用户的背景,包括他的专业、职业、爱好等等。让用户成为搜索引擎的注册用户,为他们量身定做搜索引擎不失为一种方法。注册用户可以提交自己的个性化信息,甚至自己认为合格的结果样本,搜索引擎则分析和学习这些信息和样本。在为进行检索服务时,运用这些用户背景信息,对结果进行相应的剪裁,或对用户提交的查询服务进行完善,从而使结果更符合用户的意图和要求。

通过注册机制,可以建立主动为用户提供有价值信息的服务,搜索引擎在每次更新后,如果认为有对用户有用的信息源,会主动通告用户,搜索引擎还可以为每一个用户维护一个“最有价值信息源”列表,方便用户使用。同时通过反馈,用户还可以修正系统的行为。

但是搜索引擎的定制涉及机器学习和人工智能方面的知识,只有这方面的技术有所突破,这种搜索引擎才可能实现,并能达到一定的效果。

2 让搜索引擎更好地理解信息

目前搜索引擎对信息的筛选,主要依赖关键字,它认为含有特定关键字的文档是有效的。这就好比一个卖挂毯的商人,他在进货进行筛选时,只看见挂毯是否有某种颜色,而看不见图案,而用户往往注重的是有某种颜色的图案,他们更注重图案。如果挂毯商人不能看到用户能够看到的東西,他如何满足市场?搜索引擎的

赵仲孟 副教授,在职博士生,主要从事计算机网络、数据库系统、多媒体技术的研究。

情况正相同,我们看到一个文档,不仅仅知道它含有的词语,还了解这些词语正用来说明什么,而搜索引擎只知道文档中含有的词语,它不能知道这些词语正被用来说明什么。由于搜索引擎与用户的这种差异,使得用户与搜索引擎之间存在理解障碍,为改善这种情况,就要让搜索引擎了解更多的信息,而下仅仅是关键字。解决这种问题的途径有两个:

提供特定领域的信息检索服务:由搜索引擎决定信息背景,这就是专业搜索引擎,你得根据自己的需要对搜索引擎进行选择。专业搜索引擎针对某一领域,主要满足特定专业人员的需求,所处理的信息语义背景明确,系统易于建立相关专业的知识库,便于进行页面的内容分析,而且用户群的需求也相对明确,这些都能提高搜索引擎的质量和效率。通过反馈和多种服务,也便于用户之间的交流和协作。由于只针对一个专业,系统会更加实用、更经济、利用率更高。

将站点归类:比如在一个商业站点,"CAD"很可能是用来说明一种产品,而在另一个研究站点,它很可能用来说明目前的研究方向。如果用户正在查询有关"CAD"的技术资料,那第二个站点更符合用户的需求。

通过挖掘站点的特征,能够发展特定站点类型的搜索引擎,比如高校搜索引擎,商业站点搜索引擎等专门的搜索引擎,这种搜索引擎在检索部分站点上的信息,内容上不会过于庞杂和松散。

3 搜索引擎之间的合作

搜索引擎之间需要相互协作。由于各个搜索引擎的覆盖范围有限,同时各个搜索引擎的覆盖面又不相

同,可以在各个搜索引擎之间建立协作关系,如果在一个搜索引擎中找不到满意的结果,可以通过此搜索引擎向其他搜索引擎提出请求,得到结果后,处理返回给用户,目前这种搜索引擎之间的合作已经开始,如用 Google 进行查询时,它会将用户的请求发给其他的搜索引擎,用户可以通过点击连接,直接进入其他搜索引擎的结果页面中。但是如果各个搜索引擎仍然只在有限的范围内进行合作,则效果不大。目前网络上的 Web 页面超过了 8 亿,而有调查显示现在最强的搜索引擎的覆盖率也不超过 16%。我们想如果几个搜索引擎建立了协作关系,就可以划分负责检索的 IP 范围或专题领域,这样总的覆盖面可以做到最大,充分利用资源。

网络的发展给搜索引擎提出了严峻的挑战,为了满足用户的需要,提高搜索引擎的质量和效率,要综合利用各项技术,并利用最新的技术,包括人工智能技术、信息检索技术、数据库技术等,开发出智能、高效的搜索引擎。总之搜索引擎是一个较新的领域,充满挑战和机会,有相当的开发潜力,这之中有很多技术等待我们进一步的创新和发展。

参考文献

- 1 翁惠玉,马范掇,等.网络搜索引擎的现状分析.情报学报,1999(3)
- 2 李名智.中文搜索引擎发展的现状、问题及对策.信息技术与产品,1999(2)
- 3 邹寿,王继成,等 WWW 上的信息挖掘技术及实现.计算机研究与发展,1999(8)
- 4 朱瑞勤,朱慧.对 Internet 上专题型搜索引擎的设计与实现.情报学报,1999(9)

(上接第 74 页)

从而可以使自动答疑机为学生提供回答问题的参考材料。

结论 Internet/Intranet XML 以及 C/S 技术的成熟及应用的日趋丰富为基于 XML 的标准课件应用平台的研究和设计提供了一个理想的计算环境,而 IEEE LTSC 标准的制定又为网上教学系统提供了一个规范标准。本文详细描述了一个基于 XML 的标准课件的定义,以及课件基于 XML 应用平台的设计,包括课件的管理和基于课件内容的检索等方面的研究与设计,其出发点是集成已有的技术和标准,提出规范和提出标准课件的概念和结合 XML 的设计,同时对 IEEE LTSC 标准中关于课件的 Metadata 的定义用 XML 技术实现。我们最终的目的是使远程教育系统发展成一个结合最新技术、符合国际标准、与具体系统平

台无关的协同应用的系统。

参考文献

- 1 IEEE LTSC. Draft Standard for Computer-Managed Instruction
- 2 IEEE LTSC. Learning Object Metadata Working Draft Document 3 (approved 1999-11-27)
- 3 Harold E R. XML Extensible Markup Language. IDG Books Worldwide, Inc.
- 4 Ciancarini P, Vitali F, Mascolo C. Managing Complex Documents Over the WWW. A Case Study for XML. IEEE Transactions on Knowledge and Data Engineering, 1999, 11(4)
- 5 Vouk M A, Bitzer D L, Klevans R L. Workflow and End-User Quality of Service Issues in Web-Based Education. IEEE Transactions on Knowledge and Data Engineering, 1999, 11(4)