语义检索在 Web2.0 环境下的应用探讨*

董 慧 唐 敏

摘 要 目前 Web2.0 服务虽能调动大众用户的力量,但所提供的检索方式单一,不能很好地解读用户的需求,也不利于信息的跨平台检索。语义检索是针对概念及概念之间关系的检索,能实现基于语义的匹配和推理,将语义检索应用于 Web2.0 服务中,能很好地调动用户互动参与,增强信息搜索的准确度,使服务更具个性化。而通过 Web2.0,非专业用户能为语义检索的实现贡献力量,用户使用反馈也能逐步提高语义检索的质量。本文在分析 Web2.0 服务所提供的检索和语义检索各自优劣势的基础上,从语义信息的生成、存储和检索三方面探讨两者结合的方法,并对其应用优势进行展望。参考文献 17。

关键词 语义检索 Web2.0 语义标注 本体
分类号 G354.4

ABSTRACT Although nowadays Web2.0 services have the advantage of mobilizing public power, the search methods that they provide are few. They cannot read user's needs correctly, nor are they conducive to cross-platform information retrieval. Based on the relationship of different concepts, semantic search can achieve semantic matching and reasoning. The application of semantic search in Web2.0 services can foster users' interaction and participation, enhance the accuracy of information search, and make the service more personalized. Through Web2.0, non-professional users can contribute to the implementation of semantic retrieval and improve its quality with their feedbacks. This paper analyzes the advantages and disadvantages of the semantic search and the search methods provided by Web2.0 services, then discusses the combination of the two methods from three aspects: semantic information generation, storing and retrieval, and finally presents an application prospect of the semantic search. 17 refs.

KEY WORDS Semantic search. Web2.0. Semantic annotation. Ontology.
CLASS NUMBER G354.4

随着维基、博客、社交网站(SNS)的日益盛行,Web2.0 环境下用户互动所产生的信息成为了互联网的另一重要数据来源。各 Web2.0网站针对这些数据的新特征采用了一些新的检索方法,但是检索性能仍不尽如人意。语义检索是针对概念及概念之间关系的检索,由于采用了计算机可理解、可处理的技术,能使计算机更好地理解用户的需求,从而使检索结果更准确。本文将结合 Web2.0 和语义检索各自的优劣势,探讨 Web2.0 环境下进行语义检索的潜力。

1 Web2.0 检索及语义检索的优劣势分析

1.1 Web2.0 检索的优劣势

Web2.0 环境下产生的微内容可以分为三类:围绕人的——人与人之间的链接;围绕物的——新闻资讯、博客、图书等;交互的——人和人之间虚拟的或真实的讨论[1]。这些内容的检索除提供诸如标题、作者、全文等传统检索形式外,还提供了基于标签的检索。此处的标签相当于索引,其产生主要是以用户主导为主。这些方式使非专业用户能组织检索信息。但是由于这类标签的定义掺入了过多的主观意识,用户很难判定所需信息已被定义为何种标签。对此,一些网站采取了一系列机制以保证标签更具专业性,如向用户推荐标签、淘汰使用频率不高的标签、制

* 本文系国家自然科学基金委资助项目"基于数字图书馆的本体演化与知识管理研究"(项目编号:70773087)研究成果之一。

作标签导航等。此类大众标签是 Web2.0 产品进行信息组织检索的一种创新性方式,但是对标签的检索也仅限于词型匹配,结果反馈较差,很难突出其优势。而各网站的标签信息很难集成,不利于信息的跨平台检索。

1.2 语义检索的优劣势

为了进一步提高检索质量,我们需要基于主题、概念、语义推理的检索结果,而不仅仅是简单的基于关键字,这便是语义检索[2]。目前语义检索研究主要集中在对信息资源的语义处理以实现效率更高的检索上,语义信息的提取和处理可以是基于语义网方法和技术的,也可以是基于自然语言处理技术的,前者在语义检索研究中相对更为普遍[3]。目前互联网基本上使用的是 html 语言来表示数据,而语义网则需要 xml、rdf、owl 等语言来表示数据。要实现计算机的自动推理,各个领域需要具有代表性的领域本体。这些理论及技术已获得了较快的发展,但其推行却较为缓慢。究其原因可发现:语义网实现需要大量的人力物力,专家的力量毕竟有限,而非专业用户却没法为其发展做出贡献。如果语义检索能利用 Web2.0 草根阶层的力量,必将促进语义网的发展,且能对 Web2.0 产品产生的海量信息提供有效的检索方式。

2 Web2.0 环境下语义检索的研究现状

目前,对在 Web2.0 环境下运用语义检索已有相关研究。文献[4]运用小世界原理分析了对等网环境(P2P)下文档中词与词之间的语义结构关系;文献[5-6]分别基于服务注册和本体提出了 P2P 环境下的语义检索框架;文献[7]探讨了语义网环境下博文整合的方法和潜在优势;文献[8]通过创建语义标注系统来实现博客中的语义检索;文献[9]对语义维基进行了扩展;文献[10]将 Web2.0 理念运用到 Web1.0 中,两者交互以期提高检索结果的可信度;文献[11]从过程处理的角度分析了语义检索,以语义模型描述了过程处理中所涉及的元素,并融入了用户的参与。本文将在这些思想的基础之上,探讨语义检索在 Web2.0 环境下的实现。

3 Web2.0 环境下语义检索的实现探讨

要使语义检索充分利用 Web2.0 大众用户的力量,可以从以下几个方面着手。

3.1 语义信息的生成

信息的存在是检索结果的基础保证,因此,充分利用语义标注信息获得语义网数据资源是十分重要的。目前已有研究语义标注的辅助生成工具,如奥地利维也纳大学设计的 MapFace 编辑工具,可对 MMTx 语义标注系统产生的本体标注信息进行再次修改[12]。另一重要实践领域便是语义维基,它通过对超链接进行属性标注,使得计算机可以自动处理导航链接。这两者都是对非专业用户进行语义知识创建的探索,给大众参与语义数据创建提供了思路。在 Web2.0 环境下,结合现有的工具技术,可采取以下措施促进语义信息的生成。

3.1.1 显性语义信息的创建

这里的显性语义信息指全文信息标注、超链接的标注及关系的定义等。目前的 Web2.0 产品仅仅提供简单的标签输入界面,对于标签的选择没有强有力的指导。由于标签的杂乱无章,我们面对标签检索工具也无从下手。虽然我们无法获知用户将要发布的信息,但是从目前博客等 Web2.0 产品信息发布方式来看,用户的信息总会被归到相应的大类下,这些大类可以理解为不同的领域,可以算是通用本体,这些通用本体只能由专家用户创建,而普通用户能在此大类下创建自身所需的语义标注。对用户应提供更具指导意义的标注界面,鼓励用户使用已经存在的标注,此时可利用树状列表提供已存在的二级或三级类目。用户也可在其上创建修改标注及定义标注之间的关联属性,标签属性及其关系的创建是语义检索的重要基础,可提供典型事例及微型效果展示图来指导用户。如余秋雨创建在作家的列表下,他有作品和所属年代及地区等属性。在微型的效果展示中可展示与余秋雨合著作品或同时代作家的关系联系图,通过类似体验方式加强用户对标注的兴趣。此外,超链接的标注也

是不能忽略的,除文档链接外,视频、音乐等的链接也逐渐增多,Web2.0 产品如能吸收语义维基标注超链接的思想,将使此类产品的检索更加方便。

3.1.2 隐性语义信息的提取

隐性语义指未加工的网络内容和结构,包括网络中人与人之间的交互[13](当前被称为群体智慧)。Web2.0 环境中最大的优势便是用户互动,因此可利用用户之间的交互行为提取隐性语义,判断显性语义信息的准确性。事实上,豆瓣等网站对不常用标签的摒弃也是基于此思想,但是目前对这一隐性语义的挖掘比较单一,不能调动用户参与的积极性。在对各标签赋予语义关系的前提下,对任一标签的修改都会带来连锁效应,因此对标签的删除或修改需要十分谨慎。我们可通过一系列的激励机制将这种谨慎意识传达到用户。当一个用户将要修改或删除标签时,应显示与此标签关联的其他标签;当用户确认修改或删除标签后可将此操作行为提示给用户的好友和其他使用此标签的用户,可发起评论和投票,最后确定最终的标签。为鼓励用户创建准确的语义标注,可定期发起比赛,谁的语义标注的认可度越高,便可在一定时间内享受更多更好的服务。

3.2 语义信息的存储

要充分利用语义标注的等价、传递等关系信息,需要将数据存储为 OWL/RDF 格式,这样才能使计算机具有强大的推理功能。除 AllegroGraph 等专门存储语义数据的数据库外,传统数据库如 Oracle 已开始支持存储 RDF 数据。语义维基也支持将文章以 OWL/RDF 形式导出,以及导出此文章所涉及的实例、属性及类;如本体以其指定的格式创建,语义维基也接受外界的本体导入。这些技术的发展可推动 Web2.0 产品中的语义数据的格式转换、整合及存储。

3.2.1 语义信息的格式转换

目前除了博客和新闻站点所提供的 RSS 阅读器使用 XML 表示信息外, Web2.0 产品多用 HTML 来表示信息。目前 Web2.0 产品所采用的大众标签的表现形式并不复杂,通常是直接存储在数据库中,各数据表即为标签所属的一级类目,各子标签即为表中的数据。此类信息结构化程度较高,提取较容易。其过程包括:分析数据库的结构特征和数据库的概念模型;分析本体库的概念模型,实现数据库 ER 图向本体库的类属性结果的映射;创建实例;检测与提炼[14]。RSS 中的新闻列表都处于相应的分类目录下,具体信息包括题名、摘要、链接及时间等,这些信息规律性强,利用 DOM4j 和 Jena 接口可以很好地映射为 OWL/RDF 格式信息。除了转换已有的数据形式外,也可提供编辑窗口让用户编辑简单的 RDF 文件,可提供标准格式供用户参考或自动生成必须的数据项。RDF 是以属性为中心的一种表达方法,通过定义拥有这个属性的主体范围及属性的取值范围来表现词汇之间的相互关系,以使计算机更加智能。因此在编辑界面上需提供常见属性定义的生成按钮。OWL 语言所表示的关系相对复杂,用户很难快速学习,可通过对 RDF 数据的抽取来得到。

3.2.2 语义信息整合及存储

各 Web2.0 产品有自己特定的用户群,会产生不同的语义数据源,为防止这些数据成为信息孤岛,对数据的整合是必不可少的。以语义标签为例,可利用本体的语义相似度计算的方法避免数据的冗余存储,这些方法有:基于距离的方法,即通过两个概念词在本体树状分类体系中的路径长度量化它们之间的语义距离;基于内容的方法,即比较两个概念词共享的信息量;基于属性的方法,即比较公共属性项的多少[15]。目前比较流行的 Web2.0 产品都能找到共同的表达领域,如人物、音乐等。因此语义集成重用是可行的。

语义检索的一个重要障碍便是本体的构建。上文所讨论的语义标注要成为更加通用的本体,需要经过各个领域用户的检验。目前的云计算技术可为这一问题的解决提供新思路。“云计算”一词用来描述一个系统平台或者一种类型的应用程序,一个云计算的平台按需进行动态的部署、配置、重新配置以及取消服务等[16]。如各个 Web2.0 社区能将自己产生的语义标注等信息存储在“云端”,为所需的用户提供使用,这必将减少本体构建的开发时间。

3.3 语义信息的检索

语义信息不单是一个个简单的文档,还包括各语义数据间的大量关系,对语义信息的检索应充分利用其语义关系。这种优点可充分体现在语义检索的条件扩展和结果反馈上。

3.3.1 检索条件的扩展

目前大众用户较为熟悉基于关键字的检索,这种方法不需要用户熟悉检索式的语法结构,允许他们使用自然语言表达信息需求,是目前搜索的惯用方法。因此语义检索要为大众所接受,也应该用关键字查询界面作为起始页面。要使计算机能处理用户的非结构化的自然语言信息,需使此类信息转化为结构化的条件信息。目前自然语义处理技术能比较准确地切分出需要的关键词,但是由于词义的模糊性,其相关的解释不唯一,如"苹果"一词可以理解为水果也可意为电脑品牌。针对此情况,首先需扩展出查询词的同义词,再采用图的最短路径算法,将语义标注中的类作为节点,关联属性作为边,并赋予边一定的权值,以此得到从某一上位类到查询词的 N 条最短路径,再选择相关路径的类名、属性、条件等作为检索条件输入。计算机自动处理的这一结果应提供给用户进行最终的选择,可提供一个界面以节点图的形式展示前 N 条最短路径的推理,用户可剔除其中的一条路径。如用户剔除路径,计算机可自动加入另一条较短路径,直到用户点击确认为止,待确定最后的关键字后,可展示出这些关键字所对应的属性及关系供用户选择。目前已有 SPARQL 等语言能很好地支持对 RDF 的属性及关系进行检索。如查找某一部电影,输入电影名字,通过计算机的算法选择可确定此查询关键字属于电影类别,在扩展界面中可看到该类别含有播放影院、评价指数等子类,用户可根据自己的需要选择。

3.3.2 检索结果的反馈

Web2.0 环境下用户之间有了更多的交互,这种交互不单体现在好友列表上,也体现在所发信息的相关度上,而目前 Web2.0 产品中的搜索往往将各类信息孤立起来,如按标签搜索只能提供标签中包含相应关键字的网页,综合搜索也是在各特殊位置含有关键词的页面链接的整合,这种信息简单累积的方式加重了用户查找信息的负担。即使有些网站在检索结果中提供了相应标签的链接以利于再次搜索,检索结果也很难契合用户的查询需求。针对此情况,可在检索结果旁提供剔除标志让用户选择最需要的信息,而且对相应的标签也提供类似的操作,计算机可针对用户的选择重新解析检索式,进行自动有效的二次重检。由于语义信息的支持,用户点击感兴趣的标签时,可提示用户有哪些与此标签有关联的标签可供选择,进一步提高检索的准确度,如检索书籍时,再点击该书作者的标签,可显示与该作者风格或时代相同的其他作者标签。

4 Web2.0 环境下语义检索的应用优势

语义信息的产生、存储和检索等语义检索实现过程都给 Web2.0 技术提供了新的资源,Web2.0 服务可充分利用这些资源,形成语义检索的独特优势。

4.1 个性化服务

随着 Web2.0 技术的较快发展,各 Web2.0 网站所提供的新的服务及其设想很容易被抄袭,利用个性化服务吸引用户是解决这一难题的关键。目前博客、社区网站等 Web2.0 产品都以人与人之间的互动作为发展的一大基础,而语义检索可为这一互动提供更加个性化的推荐。这种推荐可体现在:①上传用户之间。如存在信息与用户上传的信息具有相同的语义标签,且这些标签具有的属性和关系基本一致,那么在用户上传该文时,系统便可显示此类信息。用户可点击查看此类信息,也可查看信息发布者的其他信息。此方式也有利于信息的查重。②上传用户与检索用户之间。在上传用户所上传的信息的标题和标签与检索用户的检索词(式)存在较大数量的相似的时候,可以建立该上传用户与检索用户之间的社会语义关系,并对这些上传用户今后要上传的信息从某个检索用户的角度给予更多关注[17]。③检索用户之间。对相关主题进行检索的用户行为进行相似性分析,挖掘出各行为之间的语义关联,可对好

友及日志推荐提供新的思路。

4.2 新的获益方式

除了用户可从语义检索中获得更好的体验外,网站也可从中获益。这种收益包括:①获得更高的人气。较高的人气是网站生存的基础,只有抓住广大用户,企业才能获得长久发展。而语义检索所带来的智能个性化的服务给用户带来了方便,也增加了乐趣。②获得新的经济收益。前文所述的语义标注等信息集合了大众的智慧,并且是可重用的信息资源,这些资源的形成需长久积累才能日趋完美,很难模仿,可将此以有偿的方式提供给需要的用户。对一些事物评论的有力挖掘、检索,可生成能预测事物发展趋势的有价值的信息。

5 结语

语义检索能处理信息的语义内容,实现基于语义的匹配和推理。面对日益膨胀的信息及激烈的行业竞争,如能注重对信息深层次的语义构建及挖掘,实现语义检索,将给用户带来更多更好的新鲜体验,对 Web2.0 产品的进一步推广也会有帮助。本文介绍了 Web2.0 环境中语义检索实现的思路,但在实际构建中仍有许多问题有待进一步研究和解决,如如何激励各企业采取 OWL/RDF 这一新的格式组织数据,如何有效地实现语义检索的功能,这些都有待在具体实施过程中进一步探索。

参考文献:

[1] 郑钧. Web2.0 的信息组织需要引入语义的新思路[OL]. [2010-05-20]. http://blog.csdn.net/zhengyun_ustc/archive/2007/10/16/1827467.aspx.
[2] Li Lizhen, Dong Zhifeng, Xie Keming. Ontology of general concept for Semantic Searching [C]. 2010 second international conferences on computer modeling and simulation. IEEE Press, 2010:81-84.
[3] 黄敏,赖茂生. 语义检索研究综述[J]. 图书情报工作,2008(6):63-65.
[4] Bordag S, Heyer G, Quasthoff U. Small worlds of concepts and other principles of semantic search [C]. IICS2003, LNCS2877. Springer Berlin/Heidelberg, 2003:10-19.
[5] Bianchini D, De Antonellis V, Melchiori M. Service-based semantic search in P2P systems [C]. 2009 seventh IEEE European conference on web services. IEEE Press, 2009:7-16.
[6] 赵捧未,王亚楠,赵飞. 对等网环境下的语义检索研究[J]. 情报杂志,2009(6):159-162.
[7] Karger D R, Quan D. What would it mean to blog on the semantic web [C]. ISWC 2004, LNCS 3298. Springer Berlin/Heidelberg, 2004:214-228.
[8] Hope G, Wang Taehyung, Barkataki S. Convergence of Web2.0 and semantic web: a semantic tagging and searching system for creating and searching blogs [C]. International conference on semantic computing. IEEE Press, 2007:201-208.
[9] Peter H, Daniel H, Mark M, et al. Semantic wiki search [C]. ESWC 2009, LNCS 5554. Springer Berlin/Heidelberg, 2009:445-460.
[10] Katsumi T, Satoshi N, Hiroaki O, et al. Improving search and information credibility analysis from interaction between Web1.0 and Web2.0 content [J]. Journal of software, 2010:154-159.
[11] Thanh T, Peter H, Rudi S. Semantic search—using graph-structured semantic models for supporting the search process [C]. ICCS 2009, LNAI 5662. Springer Berlin/Heidelberg, 2009:48-65.
[12] Theresia G, Katharina K, Patrick M, et al. Easing semantically enriched information retrieval-an interactive semi-automatic annotation system for medical documents [J]. Human-Computer Studies, 2010:370-385.
[13] Ricardo B, Massimiliano C, Peter M, et al. Towards semantic search [C]. NLDB 2008, LNCS 5039. Springer Berlin/Heidelberg, 2008:4-11.
[14] 董慧,聂曼曼. 中文本体的半自动化构建研究[J]. 情报杂志,2009(11):145-149.
[15] 孙海霞,钱庆,成颖. 基于本体的语义相似度计算方法研究综述[J]. 现代图书情报技术,2010(1):51-56.
[16] 陈康,郑纬民. 云计算:系统实例与研究现状[J]. 软件学报,2009(5):1337—1348.
[17] 刘春茂,米国伟. Web2.0 环境下面向社会语义的信息构建的新认识[J]. 情报理论与实践,2010(2):89-92.

董 慧 武汉大学信息资源研究中心教授、博士生导师。通讯地址:武汉大学信息管理学院。邮编:430072。
唐 敏 武汉大学信息管理学院硕士研究生。通讯地址同上。

(收稿日期:2010-06-28)