基于领域本体的文献智能检索模型研究

孟红伟 张志平 张晓丹

(中国科学技术信息研究所 北京 100038)

摘 要 针对文献检索的智能化发展趋势,提出基于领域本体的文献检索模型,对领域本体构建、语义标注计算和概念相似度计算进行研究,并把模型进行了实际应用。通过实验表明,基于领域本体的文献检索系统在检索结果上由于传统的检索方式,检索效率也有一定提高,具有研究的价值和意义。

关键词 领域本体 文献检索 语义标注 概念相似度

中图分类号 G354.4 **文献标识码** A **文章编号** 1002-1965(2013)09-0180-05

Research on Intelligent Information Retrieval Model Based on Domain Ontology

Meng Hongwei Zhang Zhiping Zhang Xiaodan

(Institute of Scientific and Technical Information of China, Beijing 100038)

Abstract This paper proposes a model of domain ontology-based intelligent information retrieval, and makes in-depth study on ontology construction and semantic annotation. The model is applied to practical application. Statistics indicate that the retrieval system based on domain ontology has an enormous potential of application.

Key words domain ontology information retrieval semantic annotation concept similarity

0 引 言

文献检索是伴随着科学研究的发展而发展的。随着文献检索的网络化发展,数字图书馆系统为数字化的文献资源提供了检索平台,方便用户快速而准确的查找。目前,大多数检索技术是基于关键词匹配和分类主题目录的形式进行的[1]。这两种检索技术各具有优缺点,都为用户检索文献提供了便利。关键词匹配,是在用户输入关键词后,系统自动根据关键词在索引信息中检索匹配的文献。但当关键词具有一词多义等现象时,检索结果与用户的需求就无法实现高度吻合。分类主题目录,是用户根据事先建立的分类目录,进行浏览和检索。但这种方式为信息更新和维护带来了不便。与此同时,在信息量剧增的时代背景下,能够快速而准确地找到所需信息显得尤为重要,因此,寻找基于语义的检索方式已经迫在眉睫。自2001年Tim Berners-Lee提出语义网的概念后,本体作为语义理解的工具广泛应用在人工智能领域,用于知识的表达、共享和推理,为文献检索形式开辟了新思路,推动了文献检索向智能化的方向发展。

1 基于领域本体的文献检索研究现状

智能检索的目标是为用户提供既相关又准确的信息,尽可能保证较高的查全率和查准率。本体是对概念的明确描述,可以把某个领域抽象为一组概念与概念间的关系,具有较好的概念层次结构。把本体融合到信息检索技术中,可以实现对概念关系的处理。总的来说,本体在信息检索中可以起到概念定义、查询模型和推理基础的作用[2],因此,基于本体的智能检索研究具有非常重要的意义。

目前,国外对本体在信息检索领域的应用研究较多。近十年里研究学者提出了许多基于本体的信息检索系统和模型。Dridi在文献[3]中概述了基于本体的信息检索技术以及相应的目前可以作为开发原型或者商业产品的本体工具。Castells等人[4]探讨了基于本体的信息检索模型,主要包括基于本体的文档标识方案模型和基于向量空间模型的信息检索模型,这些模型的目的在于提高大型数据库的检索效率。文献[5]

收稿日期:2013-06-02 修回日期:2013-07-19

作者简介:**孟红伟**(1987-),女,硕士研究生,研究方向:数字图书馆、智能检索;**张志平**(1963-),男,研究员,研究方向:数字图书馆、信息系统等研究;**张晓丹**(1975-),女,副研究员,研究方向:数字图书馆、智能检索、信息检索等研究。

和[6]探讨了两种使用模糊理论来实现信息检索的本体模型的方法。用户查询需求的分析与扩展是基于本体的智能化信息检索实现过程中非常重要的技术。文献[7]提出了一种自然语言处理方法,作者在文中实现了一种基于本体的查询处理方法,用以提高信息检索的性能。文献[8]提出了一种查询扩展方法,通过搜索查询主题相关本体概念的相似概念来实现查询的扩展。智能信息检索中另一个重要的问题就是查询结果相关性的排序问题。传统的结果排序是通过分析数据库中的关键词来实现的,而非语义层面上的。文献[9]提出了一种新颖的方法来决定结果的相关性排序。由于在基于本体的信息检索系统中,本体可以很好地支持查询的处理和分析,因而文章采用了基于本体的信息搜索方法来实现查询结果的排序。与传统的实现方法相比,基于本体的信息检索方法在查询结果相关性排序上充分体现了语义相关性,结果的相关性顺序也更为合理。

国内在这方面的研究起步较晚,但是目前也已经有很多学者正在开展基于本体的信息检索方面的研究,并取得了一定的研究成果。国内对基于本体的信息检索研究主要集中于:基于本体的信息检索系统[10]、基于本体的信息检索模型[11-12]、基于本体的信息检索技术[13-15]、基于本体的信息检索策略[16]等。在研究过程中,人们逐渐认识到使用语义进行检索,让检索拥有更好的查全率与查准率。本体在智能信息检索系统中提供了形成查询与信息描述所必需的概念,以本体技术为核心建立领域语义模型,为信息资源提供语义标注信息,使系统对领域内的概念、概念之间的关系及领域内的基本公理知识有一个统一的认识,从而能够显著地提高系统的联想能力和精确性,有望快速、高效、精确地检索出用户所需的有价值的信息,同时也提供给系统内所有用户对该领域的一个全面的共同视图。本体已逐渐成为一种智能信息检索系统的知识表示,是系统集成的核心部件[17]。因此,本文将着重探讨基于领域本体的文献检索模型方面的研究。

综上所述,国内外对基于本体的信息检索研究主要集中于 Web 信息检索的模型和方法上,并取得了进步。本文在前人研究的基础上,构建了基于领域本体的文献检索模型,把领域本体在 Web 应用中的方法应用到文献检索中,把领域本体应用到语义标注和查询扩展工程中,以期改善文献检索的效率。

2　基于领域本体的文献检索模型

基于领域本体的文献智能检索模型,利用领域本体对领域知识的组织优势以及自然语言处理技术、相似度算法等,最终实现智能化检索,提高检索效率和用户满意度。模型设计的基本思路:

a. 参考领域主题词表,并在相关领域研究人员的帮助下,建立合理的领域本体。

b. 利用本体对概念及概念间关系的描述,对用户输入的检索词按照概念间的相似度大小进行扩展,排除相似度较小的概念,获得检索词集合。

c. 为用户提供较高相关性和重要性的检索结果。

考虑到系统的完整性,该模型(见图 1)包括资源管理模块、智能检索模块和相关反馈模块。由于相关反馈技术本身可以作为单独的研究对象进行研究,因此本研究只对该模块进行了设想,并将在以后的研究工作中进行完善和实现。

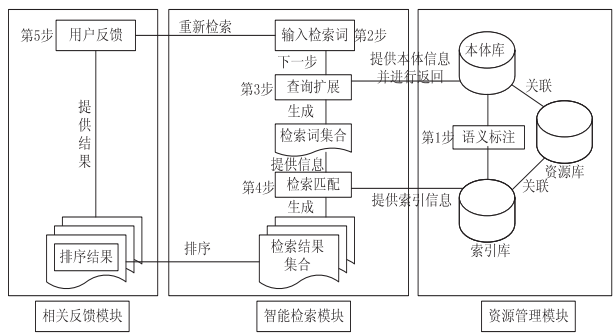

图 1　基于领域本体的文献智能检索模型

模型表示的检索过程为:

第一步,对文献进行语义标注,用领域本体对该领域的文献进行标引,构建索引库,为后续的智能检索模块提供资源。

第二步,用户输入检索词后,系统转换为可理解的检索关键词。

第三步,找出检索关键词相对应的本体概念,根据本体对概念及概念间关系的描述和概念间相似度算法,对用户的检索词进行扩展,生成检索词集合。

第四步,从检索词集合中选择出满足阈值的重要检索词,系统根据索引库中存储的文献索引信息进行匹配,获得与检索词相关的文献列表,然后系统采用智能排序算法对检索结果集合进行排序,并呈现给用户。

第五步,用户对检索结果进行评价,如果不满意检索结果或者产生新的检索词进行重新检索操作。

各个模块的功能如下:

2.1　**资源管理模块**　该模块主要是管理数字图书馆数据库中的科技文献,首先对文献进行语义标注,并构建索引库,为后续的智能检索模块提供资源。

用领域本体对该领域的文献进行标引,是根据领域本体中的概念及概念间的关系对文献内容进行标注。由于本研究拟应用于 NSTL(National Science and Technology Library,国家科技图书文献中心)中,因此选择二次文献作为标注对象,只对文献的题名、文摘、关键词进行语义标注。主要思路为:利用构建的领域

本体标注出文献的题名、文摘、关键词部分的领域概念,并为这些概念赋予相关度权值,然后提取出这些领域概念以及相关度权值,存储到索引库中。索引库中既包含文献资源的索引信息,又包括文献的标注信息,在查询扩展和检索匹配过程中起到重要的作用。

2.2 **智能检索模块** 用户输入检索词,系统对查询语句自动进行词法分析和句法分析,与专业术语相结合,生成系统可理解的检索词,然后找出检索词相对应的本体,根据本体对概念及概念间关系的描述和概念间相似度算法,对用户的检索词进行扩展,选择出满足阈值的重要检索词,系统根据索引库中存储的文献索引信息进行匹配,获得与检索词相关的文献列表,然后系统采用智能排序算法对检索结果集合进行排序,并呈现给用户[18]。该模块实现智能检索的流程图如图 2 所示。

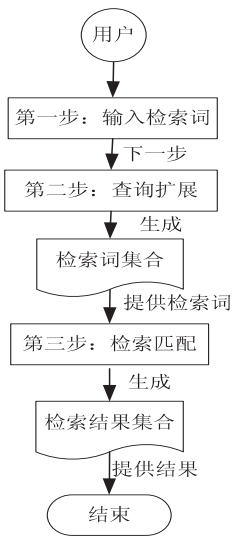

图 2 文献智能检索流程图

2.3 **相关反馈模块** 该模块的功能是保证为用户提供满意的检索结果,并记录下用户的偏好,为以后的检索节省时间。当用户给予正反馈时,系统记录下用户的偏好。当用户给予负反馈时,系统为用户提供相关检索词,并提示用户重新输入检索词。同时,该模块有利于资源管理模块和智能检索模块的优化。

3 模型实现的关键技术

3.1 **领域本体构建** 构建合理的领域本体,供参考的领域术语和相关领域研究人员的指导是必不可少的。本研究选择电力系统领域为对象,参考了《电力主题词表》和部分电力系统领域的文献关键词,并在该领域研究人员的指导下,使用本体构建工具 protégé 构建了电力系统领域本体(见图 3),生成 OWL 文件,共包含专业术语 110 个,非术语概念 45 个,概念间关系为 is_a 和 part_of。

3.2 **语义标注** 用电力系统领域本体对文献进行语义标注,主要是建立对应的概念–文献向量模型,同时量化相应的概念权重。其步骤如下:

第一步,首先对文献进行预处理,提取特征词汇并将其转化为领域本体中的概念,构建概念(C)–文献(D)相关度模型,采用 VSM 模型存储, $W_{cd} = \log(\mathrm{TF}(c,d)+1)\cdot\log\frac{m}{\mathrm{DF}(c)}$。

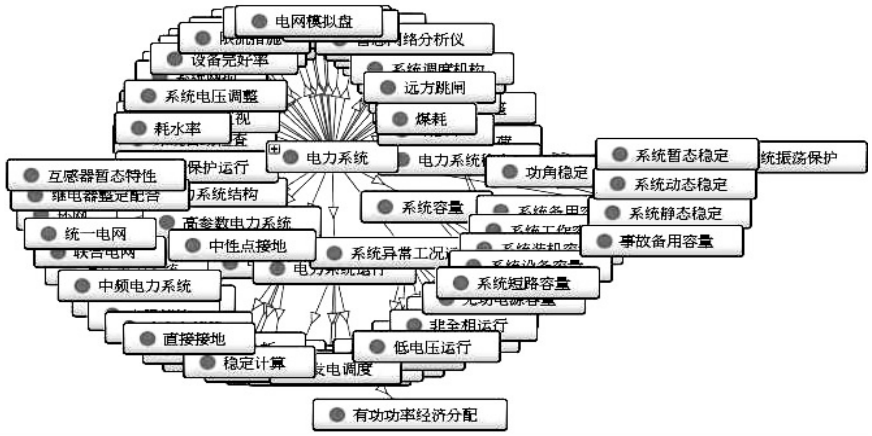

图 3 电力系统领域本体

第二步,采用改进的 TF–IDF 算法,如式(1)所示。计算概念在文档中权重值 $w_{c_i d}$(表示概念 c_i 在文献 d 的权重)(未计及概念与概念之间的语义关系,相应的矩阵记为 W),根据 c_i 在 d 中的重要程度,K 取 0 到 1 间的数,越重要 K 越大;

第三步,根据公式(2),计算概念与概念间的语义相关度 RC 矩阵,$c_{i\mathrm{super}}$ 和 $c_{j\mathrm{super}}$ 分别表示 c_i、c_j 的上位概念集合;

第四步,采用公式(3)计算,考虑语义推理之后的概念权重,并形成新的本体概念权重矩阵 W′;

第五步,存储领域本体概念与文献的映射关系及权重,保存至索引库。

$$w_{c_i d} = \frac{K * \log(TF(c_i,d)+1) * \log\frac{m}{DF(c_i)}}{\|\vec{w}\|} \quad (1)$$

$$rc_{ij} = \frac{c_{i\mathrm{super}} \cap c_{j\mathrm{super}}}{c_{i\mathrm{super}} \cup c_{j\mathrm{super}}} \quad (2)$$

$$W' = W \cdot RC \quad (3)$$

3.3 **查询扩展** 查询处理与扩展是对用户的查询请求进行处理,使其可被计算机所理解同时对用户的检索关键词进行语义扩展提高检索的查全率。本文研究了利用领域本体进行查询处理与语义扩展的方法,其核心即查询关键词与领域概念的相关度计算,这样再借助语义标注环节建好的概念–文献相关度矩阵,就可以构建关键词–概念–文献的所示关系了。

基于领域本体的查询扩展实现算法的实现步骤如下描述,具体实现流程图如图 4 所示。

$$Q = [q_1, q_2, \cdots, q_s] \quad (4)$$

$$QC = [\mu_1, \mu_2, \cdots, \mu_n]^T \quad (5)$$

$$W_{DQ}=\begin{bmatrix} w'_{d_1c_1} & w'_{d_1c_2} & \cdots & w'_{d_1c_n} \\ w'_{d_2c_1} & w'_{d_2c_2} & \cdots & w'_{d_1c_n} \\ \vdots & \vdots & \ddots & \vdots \\ w'_{d_mc_1} & w'_{d_mc_2} & \cdots & w'_{d_mc_n} \end{bmatrix} \cdot \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix} = W' \cdot QC \tag{6}$$

$$\text{diff} = W_{DQ} - ts \cdot [1,1,\cdots,1]_{m\times 1} \geqslant 0 \tag{7}$$

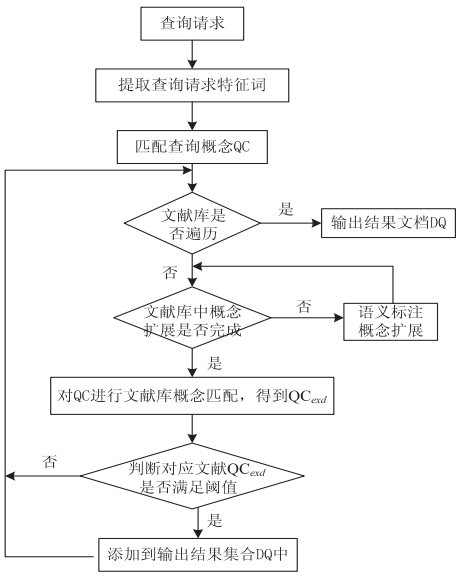

图 4　基于领域本体的查询扩展算法实现流程

算法步骤:

第一步:用户输入查询请求;

第二步:提取查询请求特征词集合 Q,写成向量空间模型,如式(4)所示;

第三步:将查询特征词转化为查询概念,形成查询概念集合 QC,分配概念的权重,如式(5)所示;

第四步:将查询概念集合中的概念逐一与文献-概念权重库中的概念进行匹配,形成匹配后的查询概念集合 QC_{exd}。文献-概念权重数据库在离线信息处理部分语义标注建模过程中已完成,概念的扩展亦通过语义扩展完成;

第五步:根据公式(6)计算文献-向量权重向量 W_{DQ};

第六步:根据公式(7)判定满足不小于设定阈值 ts 的文献集合 DQ;

第七步:返回文献集合 DQ,然后传至结果排序模块。

3.4　**检索结果排序**　首先用检索词集合中的概念作为构建检索式向量的元素,并以概念间的相似度作为权重,然后应用向量空间模型计算文献与检索式的语义相似度,按相似度大小排列检索结果,其相关性得到了保证。在此基础上,根据检索结果的被引次数和下载次数,按照公式(8)综合计算后并按降序排列,此时所得到的检索结果可以理解为是按重要性和相关性高低的次序排列的。

$$\text{Value} = 0.2 * a/100 + 0.1 * b/100 + 0.7 * c \tag{8}$$

a 为文献被引次数,b 为文献下载次数,c 为文献相似度。由于文献相似度(表示文献相关性)的计算结果为小于或等于 1 的数,所以被引次数和下载次数都要除以 100,且被引次数和相关性所能体现出的文献重要性均要高于下载次数,因此被引次数和相关性所占比重较大。

4　实验验证及分析

通过构建电力系统领域本体,把领域本体应用到语义标注和查询扩展中,从而提高文献检索的查全率和查准率。把上述方法用计算机语言表示,在 Java 运行环境下,把研究成果应用于 NSTL 的中文期刊检索中。通过对基于本体的检索系统与 NSTL 原系统的基于关键词的检索系统进行比较,来验证基于本体的检索系统对文献检索的查全率和查准率的影响。

从 NSTL 的日志记录中抽取了电力系统领域较为关注的 10 个查询语句,在加入本体后的中文期刊检索系统中进行检索,得到如下检索结果(见表 1):

表 1　电子系统领域前 10 个查询语句表

编号	检索式	基于本体的检索系统		原检索系统	
		结果总数	相关文献数	结果总数	相关文献数
1	电力系统自动化	1458	847	1309	735
2	电力系统状态估计	213	145	155	108
3	电力电子系统	334	208	226	154
4	电力系统分析	739	613	656	518
5	电力系统潮流计算	230	165	199	130
6	供电系统	15781	9107	11825	7844
7	配电系统	10808	7353	8929	5753
8	负荷预测	5399	3627	4087	2811
9	无功优化	2046	1179	1748	1076
10	电力监控系统	750	446	569	312
平均		3775.8	2369	2970.3	1944.1

分析检索结果,发现由于根据领域本体对用户的检索词进行了扩展,所以检索的查全率提高了。通过对文献进行语义标注以及计算概念与文献的相关度,因此能够检索出更多相关的文献。根据以上检索结果总数和相关文献数量,利用查全率和查准率计算方法,并把基于本体的检索系统和原检索系统的查全率和查准率绘制成 P-R 曲线。从图 5 中可以看出,领域本体的加入在一定程度上提高了检索的效率。

5　结束语

本文分析了文献检索智能化的发展趋势,对国内外相关研究进行了系统介绍,提出了顺应趋势发展的基于领域本体的文献检索模型。该模型通过对文献进行语义标注,强化了领域概念与文献间的相关度,并利用领域本体对概念及概念间关系的描述,对用户检索

式进行扩展。通过实验分析,基于领域本体的文献检索系统在一定程度上提高了检索效率,但由于领域本体构建的规模较小,涉及到的领域概念较少,所以优势并不明显。因此,完善电力系统领域本体将是本研究下一步进行的研究工作。

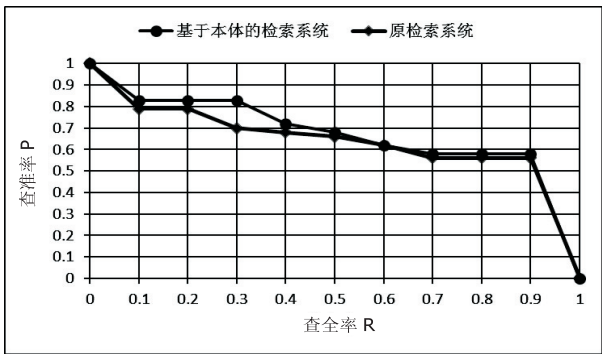

图 5 两个检索系统的 P-R 曲线

参 考 文 献

[1] 王 峰,汪华方. 数字图书馆信息检索技术的智能化发展趋势[J]. 现代情报,2008,28(11):93-95, 99

[2] 石 静,肖航宇,熊前兴. 基于 SWRL 规则与本体相似度的语义检索模型研究[J]. 计算机应用与软件, 2010,27(7):236-238,273

[3] Mauro Dragoni, Célia da Costa Pereira, Andrea G. B. Tettamanzi. A Conceptual Representation of Documents and Queries for Information Retrieval Systems by Using Light Ontologies[J]. Expert System with Applications,2012,39(12):10376-10388

[4] Castells P,Fernandez M, Vallet D. An Adaptation of the Vector-Space Model for Ontology-Based Information Retrieval [J]. Knowledge and Data Engineering, 2007, 19(2):261-272

[5] Zhai J,Liang Y, Jiang J, Yu Y. Fuzzy Ontology Models Based on Fuzzy Linguistic Variable for Knowledge Management and Information Retrieval[M]. Intelligent information processing Ⅵ, 2008

[6] Calegari S, Sanchez E. A Fuzzy Ontology-Approach to Improve Semantic Information Retrieval [C]. Proceedings of the Third ISWC Workshop on Uncertainty Reasoning for the Semantic Web -URSW, 2007

[7] Li Z,Ramani K. Ontology-Based Design Information Extraction and Retrieval [J]. AI EDAM, 2007, 21(2):137-154

[8] Díaz-Galiano M, García-Cumbreras MA, Martín-Valdivia M T,et al. Advances in Multilingual and Multimodal Information Retrieval [M]. Berlin: Springer-Verlag Berlin Heidelberg, 2007:601-606

[9] Stojanovic N. An Approach for the Efficient Retrieval in Ontology-Enhanced Information Portals[J]. Lecture Notes in Computer Science,2004,3336:414-424

[10] 王存刚,王 斌,姚文琳,等. 基于 Ontology 的 Web 信息检索系统研究[J]. 计算机工程与设计,2008,29(24):6316-6317

[11] 郭承霞,王爱继,陈庆海. 基于领域本体的智能信息检索模型研究[J]. 计算机科学,2009,36(4A):101-103

[12] 熊忠阳,李春玲,张玉芳. 一种基于领域本体的混合信息检索模型[J]. 计算机工程,2008,34(21):68-70

[13] 丁政建,李 飞. 基于本体的信息检索技术的研究[J]. 科学技术与工程,2008(13):3660 -3663

[14] 李 飞,赵世霞. 基于本体的语义信息检索技术的研究[J]. 信息与电脑,2010(6):106-107

[15] 张 娜,李宝敏. 语义检索及其关键技术研究[J]. 计算机技术与发展,2006,16(11)

[16] 贾松浩,杨 彩,张海玉,等. 基于本体的 Web 改进的信息检索方法[J]. 微计算机信息,2009(33):212-213,156

[17] 张 明. 基于本体的智能信息检索研究[D]. 保定:河北大学,2007

[18] 陈 欣,李晓菲. 基于领域本体的专业文献信息检索研究[J]. 现代图书情报技术,2009(7):59-64

(责编:刘影梅)