

近十年本体应用的研究进展

宋朋

(南开大学商学院信息资源管理系, 天津 300071)

摘要: 将2005年以来本体在语义检索、信息系统、知识管理知识发现、机器翻译、医学生物学以及地理信息系统等领域的应用现状进行了总结分析; 在对本体应用领域揭示的基础上, 对今后的发展趋势进行了展望。

关键词: 本体; 语义检索; 信息系统; 知识管理; 知识发现; 机器翻译

中图分类号: TP18

DOI: 10.3772/j.issn.1673—2286.2014.10.004

1 引言

随着计算机等高新技术的不断发展, 人们可以获取的信息数量已经远远超出了人们的想象, 信息在带给人们无限便利的同时也产生了另外的问题——信息无序、知识异构。本体 (Ontology) 概念起源于哲学领域^[1]; 在图书情报领域, 最早出现于1967年S. H. Mealy所著论文中^[2]; 上世纪80年代, 计算机学科领域学者Gruber给出了广泛接受的定义“本体是共享概念模型的明确形式化规范说明”^[3]。胡亚军和刘鲁红将本体视为当前概念层面上有效的知识组织方法。本体的核心作用在于它能够对某一领域或领域内的概念以及概念之间的关系进行定义, 因其具有良好的概念层次、表达特点以及在概念间建立丰富语义联系的能力可以实现知识整合, 从而在某

种程度上解决信息和知识的无序、异构问题。

2000年12月, WWW的缔造者Tim Berners-Lee提出的语义Web为人们勾勒出了下一代互联网理念, 而本体作为语义Web实现的核心, 自然成为了学者们关注的焦点。本文将2005年以来本体在语义检索、信息系统、知识管理、知识发现、机器翻译等领域的应用情况进行总结分析并对本体应用今后的发展进行了展望。

2 2005年以来本体应用领域的确定

笔者于2014年10月10日, 在CNKI数据库以“本体应用”为主题词检索得到533篇论文, 将1975年至2014年发文趋势绘制如下 (见图1)。如图可见, 自2005年以来国内学者对本体应用领域的关注度不

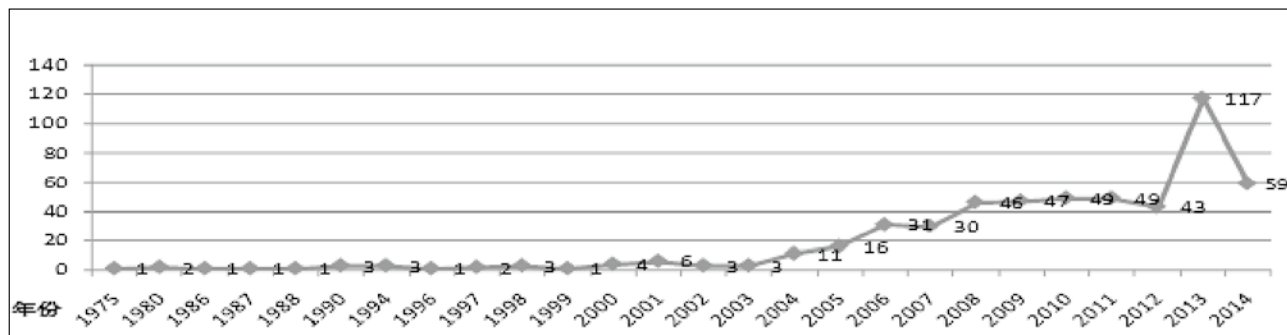


图1 CNKI数据库1975年至2014年发文趋势

断增加,并于2013年达到研究高潮。提取533篇论文的关键词,将出现频次 ≥ 5 的关键词从高到低进行排序(见表1)。然后根据高频关键词所反映的文献主题、2005年以来文献揭示的重要内容以及Google Scholar核心文献所反映的研究主题,总结得到本体在语义检索、信息系统、知识管理、知识发现、机器翻译等领域的应用热点。

3 本体的应用现状

3.1 本体在语义检索中的应用

语义检索把信息检索与人工智能、自然语言技术进行了结合。目前语义检索存在本体论和概念论两类,基于本体论的检索通过构建概念空间来实现语义检索,而基于概念论的检索则根据概念词典和关系数据库构建概念空间来实现语义检索^[4]。

TAP^[5]将语义网中的文档、概念视为结点。Hovy^[6]将半自动线性领域本体加入到大型中央本体的启发方法进行了描述。基于语言概念名称分析及概念自然语言定义,首先匹配程序使用自然语言处理技术将单词名称进行拆分;然后比较子字符串长度,发现相似概念名称;最后通过实验对相似性整合方法进行检验从而发现用户的需求满意度。Mayfield及Finin^[7]在序列和循

环中以迭代相关性反馈方法将面向本体技术与面向文本检索技术进行了整合。在层级结构和规则推理中,使用询问扩展及文档标注语义扩展方法。Pablo^[8]提出面向本体的知识库开发方法将其用于大规模文献集检索中,他将语义网的信息检索视为检索引擎,返回文档而不是返回用户查询的确切值。将语义检索与传统的关键词检索进行整合,对向量空间模型:集合权重标注及排序算法加以改进。

休斯敦大学的Ping Chen和Rakesh Verma提出使用用户询问文本摘要技术来改善医学信息检索效果。将其应用于美国国家医学图书馆开发的UMLS(一体化医学语言系统)中帮助医学信息系统理解生物医学和健康领域的概念和术语内涵及其相互关系。Aijuan Dong^[9]提出基于多本体的多媒体注解模型,通过将领域无关的多媒体本体与领域本体进行整合,实现从多个特定领域对多媒体内容进行描述,从而更好地揭示用户信息需求。

3.2 本体在信息系统中的应用

本体的本质就是领域知识的共享与复用,标准化和形式化的领域本体是信息系统之间高层互操作很好的工具^[10]。因此,本体应用的一个重要领域就是信息检索和集成,尤其在多信息源中信息表示不一致情况下,

表1 CNKI数据库1975-2014年本体应用主题高频关键词

关键词	频次	关键词	频次	关键词	频次	关键词	频次
本体	260	地理本体	11	本体进化	7	信息抽取	6
领域本体	55	知识管理	11	本体应用	7	知识工程	6
本体构建	38	本体推理	11	Jena	7	RDF	5
本体映射	28	相似度	10	本体工程	7	数据挖掘	5
语义Web	27	推理	10	语义标注	6	数据集成	5
OWL	24	本体匹配	9	关系数据库	6	Protégé	5
语义网	21	本体学习	9	本体聚合	6	本体合并	5
语义相似度	16	描述逻辑	9	概念格	6	查询扩展	5
本体论	15	Web服务	8	知识组织	6	文本分类	5
信息检索	15	本体集成	8	相似度计算	6	数据库	5
语义检索	15	WordNet	8	映射	6	形式概念分析	5
语义	13	聚丙烯装置	8	本体感觉	6	本体建模	5
知识表示	12	知识库	8	SWRL	6	向量空间模型	5
应用	11	本体法	7	可视化	6		

引入本体建立共享术语和信息表示结构,不但可以降低信息转换的难度和工作量,而且可以提供统一的模式实现信息集成。此外,本体具有良好的概念层次结构、可以实现逻辑推理,使得检索更加符合用户的需求。本体在信息系统中的应用主要包括信息组织、信息检索和异构信息系统互操作问题。目前用于信息检索的本体项目种类繁多,国际上很多著名项目都应用到了本体,如OntoAgent、Ontobroker、SKC等。

Myunggwon Hwang等^[11]为实现领域本体知识的最大化复用,利用OWL/RDF(s)对检索系统进行设计。建构系统包括网络爬虫、分类模型、排序模块及检索模块。除网络爬虫外,其他模块都采用WordNet对概念进行分析。Frederico Fonseca^[12]认为本体需要解决一般的假设问题,从而为所有领域数据进行理解和解释提供框架。另一方面,概念图示应该强调一般类目解释以及事实间关系,并在数据库领域对这一观点进行了例证。相对本体而言,概念图示应该包括特殊领域类目含义解释的规定,以及某一领域相关数据之间可能变量的结果维度。因此,概念图示使得特定领域的论据既有清晰度又有主题检测度。Frederico Fonseca认为领域的自然演化对于信息系统研发人员及用户具有重要的指导意义。

Werner Kuhn及Martin Raubal^[13]建议在语义参考咨询系统的所有部分应用简单导航程序。在本体设计规模及启示问题中,讨论了是否通过更多案例研究将该方法在重要地理空间进行实现的问题。地理信息需求生产者和使用者为数据模型投影在不同语义参考咨询系统中的转换提供了支持。在概念层级输入多元化参数实现有序推理。此外将地理信息系统置于人类活动情境下,应用术语含义与实体、关系进行关联。

3.3 本体在知识管理、知识发现中的应用

语义网技术的迅速发展为从面向文档的知识管理转换为面向知识块的知识管理提供了基础。目前网上智能“Push”服务、知识管理与事务处理集成等已经成为迫切需要解决的问题,而本体是实现这些功能的关键。本体能够对非结构化的信息进行注释以表明它们的语义,能够用于信息集成、辅助生成面向用户的视图,从而使得知识访问更加简单^[14-16]。在知识管理中,本体的主要作用表现在:为文档提供语义标注、提高检索的查

全率和准确率;为多个信息源进行信息集成提供组织架构;便于知识、数据和模型准确、有效的交换;制定信息约束系统,保证信息的正确性、一致性;创建可重用模型库^[17,18]。

Kayed^[19]提出一种新的减少抽取关系对的算法,使用文本挖掘工具对电子商务领域概念进行抽取。依据与其他概念的关系数量,对概念进行分组并提供形式化,对算法生成软件进行实验,证明减少关系数量可以降低专家的人力投入。Holsapple^[20]描述了一般性目的知识管理本体,实践者、研究者和教育者都可以利用这一本体。概念和术语通过30多个知识管理实践者和研究者采用德尔菲法获取得到,将近100个定义和公理进行了形式化表示,实现实体在知识资源上的互操作。引入分类法识别知识管理情景中,知识管理的影响因素。

随着Web的出现,数据数量、网络文件以及用户数量也在快速增长,研究者试图通过开发元数据模式来增强数据间的互操作。社会网络书签网站上可以生成大量元数据,这为研究人员提供了丰富的研究素材。Ding对UTO(上层标签本体)进行描述,通过从三个主要社会标签系统(Delicious、Flickr和YouTube)上获取元数据,对其在建模、采集、查询和分析方面的效益进行了演示^[21]。

Durbha等^[22]基于领域本体提出面向概念模型架构,首先识别出领域内的基本概念,然后根据推理层面的需求执行特殊询问。利用无监督分割算法来抽取同质区域,根据颜色、质地、形状对每个区域的原始描述符进行计算。原始描述符通过中层描述符进行定量描述。应用学习机制将非线性支持向量机(SVM)进行关联,根据语义类对关联进行分组并在询问中使用。通过界面整合执行复杂询问并将数据探索工作扩展到点对点环境。

传统文本挖掘系统采用潜解析技术,关注概念抽取及分类关系的抽取。Jiang提出CRCTOL系统,将文献中特定领域本体的丰富语义知识进行挖掘。通过全文解析技术将统计和词汇语法方法进行整合,将抽取结果与当前本体学习系统Text-To-Onto进行定性比较后发现,CRCTOL在概念、关系的抽取精度、召回率要更好,在句子及复杂结构内容中更为显著^[23]。

Chyan Yang等开发了信息技术产品规格抽取系统,将自然语言处理技术与本体概念进行融合。首先随机选取IBM和HP中的34个网页作为文档库;使用

ANNIE (基于规则方法的英文信息抽取系统)对网页文档进行分词,在JAPE(标注模式引擎)中人工建立语法规则,对信息实体进行标注;基于IT产品特征对概念及其关系进行描述,最终构建成IT产品规格本体。信息技术产品规格抽取系统在对其他品牌的IT产品规格信息抽取试验中表现良好^[24]。

3.4 本体在机器翻译中的应用

本体作为知识模型可以实现概念层次上的知识建模、概念组织从而建立丰富的概念关系。本体在知识库中的引入无疑促进了自然语言处理以及机器翻译的发展,国内的知网、北京大学计算语言研究所的“现代汉语语义词典”,国外普林斯顿大学的“WordNet”、新墨西哥州立大学的“Mikrokosmos”,都是基于本体思想而建立的。目前国内外较为知名的本体知识系统主要有以下几个: WordNet、Cyc、HowNet、FrameNet、GUM、SENSUS、Mikrokosmos。其中, HowNet和WordNet是本体开发的早期雏形, GUM、SENSUS和Mikrokosmos近年来发展得越发成熟,主要用于机器翻译领域。Cyc具备一定的推理功能,相对更加成熟完备,成为很多开发领域的概念基础。由于中文语言的特殊性,要充分利用这些已有的本体系统,目前需要解决的问题还很多。

Fu^[25]在跨语言本体映射情境环境下对机器翻译工具以及单语本体匹配技术通用方法的可靠性进行了检验。实验结果源于案例研究,单独本体映射采用中英文进行呈现。基于这些研究发现,对通用方法存在的局限进行了讨论,说明在本体标签中加入适当的翻译对于跨语言本体映射具有十分重要的作用。最后,提出识别的困难在于面向情境的跨语言本体映射(SOCOM)。Dorr^[26]对中介语言设计、建构者及中介语言系统中的语言及主要问题进行了定义,从操作系统、研究原型中对实例进行了说明,对面向中介语言的文本信息标注工作进行了讨论。Vickrey^[27]希望从情境中对词语翻译进行纠正,使用平行语料作为本体部分标签数据,提出算法解决词汇翻译问题。然后表明,词汇翻译系统可以用于机器翻译的简化并且可以有效、准确地为词汇删减候选翻译集。

许卓明^[28]提出了一种从ER模式到OWL DL本体的语义保持翻译方法。该方法在形式化表示ER模式基础上,建立ER模式和OWL DL本体之间精确的概念对应关系,通过翻译算法按照一组预定义的映射

规则实现模式翻译。理论分析表明,完全自动的机器翻译是可实现的。该方法为实现Web本体开发以及数据库和语义Web之间的语义互操作开辟了一条有效途径。

3.5 本体在地理信息、医学生物学等领域的应用

张莹^[29]从分析本体概念入手,描述了本体在地理信息科学中的含义并重点分析了本体在地理信息科学中的应用。本体思想在空间数据共享和互操作空间信息的概念模型等方面有十分重要的意义,虽然地理信息科学中的本体研究相当复杂,目前仅处于初步研究探索阶段,但是本体以其在地理信息科学中独特的优势,顺应了地理信息系统日益智能化、网络化、异构化和大众化的发展潮流,因而地理本体在地理信息科学中具有广泛的应用前景。

Deserno^[30]建议在医学CBIR研究中采用更为系统、综合的概念视角来看待概念“鸿沟”。作者定义了14个本体鸿沟,强调图像内容及特征以及系统绩效及可用性。除了这些鸿沟,作者识别出影响CBIR适用性及绩效的7个系统特征。作者所创建的架构可以用于设计前阶段,为具体医学图像技术领域及目标验证方法的比较。作者提出命名及分类规则对于医疗CBIR系统进行分析 and 评估,试图强调医疗CBIR的核心特征及所需要的功能,使用鸿沟概念作为统一概念来强调CBIR系统各方面的潜在不足,努力推动医疗CBIR的发现,通过提供更为正式、系统的方法,根据其特征、不足对CBIR系统进行改进。

Mima^[31]在生物医学领域提出了一个整合的知识挖掘系统,该系统将术语识别、术语聚类、信息检索及可视化进行了整合。系统初期的目标旨在便于从文档中进行知识整合,通过面向术语相似性计算及结构知识的自动可视化来辅助知识发现。该系统还有助于将不同类型的数据库进行整合,实现不同知识类型的同时检索。为了便于知识发现,作者还提供了基于知识相似性生成图谱的可视化方法。最后,采用GENIA语料库对系统的实用性和适用性进行了评估。

Hua Xiaohua等开发了一个 SPIE 系统(可扩展便携式信息抽取系统)。与以往使用注解语料库进行训练不同,该研究只使用少量的种子样本。SPIE系统能在大

规模生物医学文档库中抽取多个二元关系,如蛋白质间的交互作用、蛋白质与DNA的交互作用等^[32]。美国德雷塞尔大学Zhou Xiaohua等开发了一个名为MedIE(医学信息抽取)的系统,可以从大规模临床记录中抽取和挖掘患乳腺疾病的病人信息。

4 本体应用的研究趋势

随着语义网的不断发展完善,本体越发受到人们重视并在生活工作的各个领域特别是科研中扮演着越发重要的角色。如何将形形色色、各式各样的信息资源有机整合起来,并深度挖掘信息资源的价值,揭示信息资源之间的联系,以便更加有效地实现语义服务成为了本体应用的一大使命。基于上述不同领域中本体应用现状的揭示和总结,笔者提出本体应用今后发展趋势:

(1) 本体的应用已深入到各个领域,除了传统的计算机、图书情报学外,近年来本体应用在医学生物学、地理信息领域的应用也在不断普及。随着语义网的逐步实现、完善,本体应用的前景将会更加广泛。

(2) 在当今网络环境下,知识管理系统更加强调交互性。在实现知识管理系统的交互性方面,方法越发多样化,如在所包括的词汇或链接中建立整合、直接建立各种在线目录系统主题规范记录等。

(3) 在知识发现、知识挖掘方面,本体实现的技术更加复杂、多元化,挖掘的精度、准度在不断改进的同时,学者们也将目光着眼于在特定的领域情境下对不同用户信息需求的挖掘,以便满足用户个性化、情境化的信息需求。

参考文献

- [1] 冯志勇,李文杰,李晓红.本体论工程及其应用[M].清华大学出版社,2007.
- [2] YUAN G, LI H, FAN B. Discussion on the construction method of oil exploration and development domain ontology [C]// Electronics and Optoelectronics (ICEOE), 2011 International Conference on, 2011, 4: 401-403.
- [3] GRUBER T R. A translation approach to portable ontology specifications [J]. Knowledge acquisition, 1993, 5(2): 199-220.
- [4] 黄敏,赖茂生.语义检索研究综述[J].图书情报工作,2008,52(6):63-66.
- [5] GUHA R, MCCOOL R, MILLER E. Semantic search [C]// Proceedings of the 12th international conference on World Wide Web, 2003: 700-709.
- [6] HOVY E. Combining and standardizing large-scale, practical ontologies for machine translation and other uses [C]// Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC), 1998: 535-542.
- [7] BOZSAK E, EHRIG M, HANDSCHUH S, et al. KAON: towards a large scale Semantic Web[M]// E-Commerce and Web Technologies, Springer, 2002: 304-313.
- [8] CASTELLS P, FERNANDEZ M, VALLET D. An adaptation of the vector-space model for ontology-based information retrieval [J]. IEEE Transactions on Knowledge and Data Engineering, 2007, 19(2): 261-272.
- [9] DONG A, LI H. Multi-ontology based multimedia annotation for domain-specific information retrieval [C]// IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing, 2006: 158-165.
- [10] 李健康,张春辉.本体研究及其应用进展[J].图书馆论坛,2005,24(6):80-86.
- [11] HWANG M, KONG H, KIM P. The design of the ontology retrieval system on the web [C]// Advanced Communication Technology, 2006. ICACT 2006. The 8th International Conference, 2006: 1815-1818.
- [12] FONSECA F, MARTIN J. Learning the differences between ontologies and conceptual schemas through ontology-driven information systems [J]. Journal of the Association for Information Systems, 2007, 8(2): 2.
- [13] KUHN W, RAUBAL M. Implementing semantic reference systems [C]// AGILE, 2003: 63-72.
- [14] ANGELE J, SCHNURR H-P, STAAB S, et al. The Times They Are A-Changin'-The Corporate History Analyzer [C]// PAKM, 2000.
- [15] STAAB S, STUDER R, SCHNURR H-P, et al. Knowledge processes and ontologies [J]. IEEE Intelligent systems, 2001, 16(1): 26-34.
- [16] SURE Y, MAEDCHE A, STAAB S. Leveraging Corporate Skill Knowledge-From ProPer to OntoProPer [C]// PAKM, 2000.
- [17] 贺赛龙,刘柏嵩.知识管理中本体演化研究[J].情报学报,2004,23(4):469-475.
- [18] SHVAIKO P, EUZENAT J. A survey of schema-based matching approaches [J]. Journal on Data Semantics IV, Springer, 2005: 146-171.
- [19] KAYED A, HIRZALLAH N, SHALABI L A, et al. Building ontological relationships: A new approach [J]. Journal of the American Society for Information Science and Technology, 2008,

- 59(11): 1801-1809.
- [20] HOLSAPPLE C W, JOSHI K D. A formal knowledge management ontology: Conduct, activities, resources, and influences [J]. Journal of the American Society for Information Science and Technology, 2004, 55(7): 593-612.
- [21] DING Y, JACOB E K, FRIED M, et al. Upper tag ontology for integrating social tagging data [J]. Journal of the American Society for Information Science and Technology, 2010, 61(3): 505-521.
- [22] DURBHA S S, KING R L. Knowledge mining in earth observation data archives: a domain ontology perspective [C]// Geoscience and Remote Sensing Symposium, 2004. IGARSS'04. Proceedings. 2004 IEEE International, 2004.
- [23] JIANG X, TAN A-H. Mining ontological knowledge from domainspecific text documents[C]// Fifth IEEE International Conference on Data Mining, 2005: 4.
- [24] YANG C, CHEN L-C, PENG C-Y. Developing and evaluating an IT specification extraction system [J]. The Electronic Library, 2006, 24(6): 832-846.
- [25] FU B, BRENNAN R, O' SULLIVAN D. Cross-lingual ontology mapping: an investigation of the impact of machine translation [M]// The Semantic Web. Springer, 2009: 1-15.
- [26] DORR B J. Machine translation divergences: A formal description and proposed solution [J]. Computational Linguistics, 1994, 20(4): 597-633.
- [27] VICKREY D, BIEWALD L, TEYSSIER M, et al. Word-sense disambiguation for machine translation [C]// Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, 2005: 771-778.
- [28] 许卓明,董逸生,陆阳.从ER模式到OWL DL本体的语义保持的翻译[J].计算机学报,2006,29(10):1786-1796.
- [29] 张莹.地理本体的研究:研究进展与应用[J].测绘标准化,2014(2):9.
- [30] DESERNO T M, ANTANI S, LONG R. Ontology of gaps in content-based image retrieval [J]. Journal of digital imaging, 2009, 22(2): 202-215.
- [31] MIMA H, ANANIADO S, MATSUSHIMA K. Terminology-based knowledge mining for new knowledge discovery [J]. ACM Transactions on Asian Language Information Processing (TALIP), 2006, 5(1): 74-88.
- [32] HU X, LIN T Y, SONG I-Y, et al. A semi-supervised efficient learning approach to extract biological relationships from web-based biomedical digital library [J]. Web Intelligence and Agent Systems, 2006, 4(3): 327-339.

作者简介

宋朋, 男, 1985年生, 南开大学2012级情报学博士研究生, 研究方向: 领域分析、本体构建, E-mail: songpengnk@126.com。

Study on Ontology Applications in the Recent Decade

SONG Peng

(Department of Information Resources Management, Business School, Nankai University, Tianjin 300071, China)

Abstract: This paper analyzes the ontology applications in the areas of Semantic Retrieval, Information Systems, Knowledge Management & Knowledge Discovery, Machine Translation, Medical Biology and Geographic Information Systems, et al. home and abroad since the year 2005. Finally, it summarizes the situations and prospects trends in the future, based on the demonstrations of ontology applications.

Keywords: Ontology; Semantic retrieval; Information system; Knowledge management; Knowledge discovery; Machine translation

(收稿日期: 2014-10-11)