

基于用户行为的启发式本体搜索机制

李江华^{1,2*}, 郑剑¹

(1. 江西理工大学 信息工程学院, 江西 赣州 341000; 2. 北京科技大学 国家材料服役安全科学中心, 北京 100083)

(* 通信作者电子邮箱 ljh@mail.jxust.cn)

摘要: 为了能够以较高的准确率搜索到用户所需要的领域本体, 在分析本体搜索需求和研究用户搜索行为的基础上, 提出了一种基于用户行为的启发式本体搜索机制, 利用不同用户由于领域认知不同, 输入的具有领域共性的搜索关键词不同, 实现用户搜索关键词的启发式扩展和搜索匹配度的提高。实验表明, 使用该方法执行本体搜索具有较高的准确率和召回率。

关键词: 本体; 本体搜索; 关键词; 关键词扩展; 关键词搜索

中图分类号: TP18 **文献标志码:** A

Heuristic ontology search mechanism based on user behavior

LI Jiang-hua^{1,2*}, ZHENG Jian¹

(1. School of Information Engineering, Jiangxi University of Science and Technology, Ganzhou Jiangxi 341000, China;

2. National Center for Materials Service Safety, University of Science and Technology Beijing, Beijing 100083, China)

Abstract: In order to search domain ontologies needed by users in higher precision, a heuristic ontology search mechanism was proposed on the basis of analyzing demands for ontology search and studying user search behavior, which took full advantages of the different search keywords belonging to same domain input by different users for their different domain knowledge to realize heuristic extension for users' search keywords and improvement for search matching. The experimental results show that the proposed approach could help users to search and get the relevant ontologies at a higher precision and recall.

Key words: ontology; ontology search; keyword; keyword extension; keyword search

0 引言

互联网已经成为人们获得信息、享受服务的重要渠道。但是, 由于互联网是面向人类设计的, 其信息描述方式不能被机器自动处理和理解。随着语义 Web^[1] (Semantic Web) 的诞生, 它利用对数据的语义描述, 使计算机能够理解 Web 信息, 实现计算机之间的智能交互, 从而使智能 Web 成为可能。语义 Web 以资源描述框架为基础, 以本体为核心, 利用本体对领域内资源在语义层次上的描述, 使领域内的资源从内容级别提升到语义级别, 使得资源的管理和利用更加智能和高效。随着语义 Web 的发展和知识共享需求的驱动, 人们对本体的认识不断加深, 本体已经被证明对描述领域知识很有价值并且迅速成为语义 Web 的支柱^[2]。

随着本体重要性的提高, 使用本体的应用也越来越多, 然而, 本体的构建耗时耗力且成本昂贵, 另一方面, 互联网上存在的本体越来越多, 利用本体重用的优点, 人们可以根据具体应用的需要对本体进行扩展和精简, 从而达到高效共享领域知识的目的。因此, 采用合适的搜索机制, 高效准确地从本体库或通过本体搜索引擎搜索到用户所需要的本体是一个亟待解决的问题。

1 相关研究

当前获取本体的方法主要有两种: 本体库和本体搜索引擎。目前比较常用的本体库有 DAML library^[3]、Ontolingua^[4]、

the Protégé OWL library^[5] 等, 本体库主要提供本体的构建、维护、复用和应用等, 缺乏从互联网上自动收集本体的功能, 因此库中本体的数量很少, 而且不提供搜索功能, 或提供的搜索功能非常有限, 用户难以快速准确找到所需要的本体。为了能够自动搜集互联网上分布的大量本体, 一些组织开发了专用的本体搜索引擎, 采用不同搜索机制的用户接口, 供用户查找所需要的本体。

Swoogle^[6] 是当前最重要的本体搜索引擎之一, 提供本体和元数据的搜索及排序服务。用户接口类似于 Google, 用户输入关键词进行搜索。这种接口搜索机制很难准确表达用户的搜索需求, 因此为了改善这种表达的不足, 提高搜索的准确性, Swoogle 在高级搜索中提供了一系列的搜索约束, 如: 1) URL (Unified Resource Location) 约束; 2) 内容约束, 包括类、属性及三元组; 3) 语言和编码格式约束等。类似的本体搜索工具还有 AKTiveRank^[7], 它通过调用 Swoogle 的 Web APIs 进行搜索。

OntoKhoj^[8] 提出了一种面向上下文的查询接口, 利用 WordNet, 首先对用户输入的关键词进行语义消歧, 从 WordNet 中获取关键词的近义词和上位词集合。在搜索时遵循以下原则: 1) 如果存在匹配结果, 则按用户输入的关键词搜索; 2) 如果没有匹配结果, 使用近义词搜索; 3) 如果仍然没有匹配, 则使用上位词进行搜索, 直到有搜索结果。

OntoFetcher^[9] 通过查询 WordNet 获取用户输入关键词的近义词、上位词和下位词集合, 提供接口供用户选择查询词之

收稿日期: 2012-04-06; 修回日期: 2012-05-30。 基金项目: 国家“十一五”科技基础平台项目 (2005DKA32800)。

作者简介: 李江华 (1976-), 男, 河南新野人, 副教授, 博士研究生, 主要研究方向: 语义 Web、数据工程; 郑剑 (1977-), 男, 湖北黄石人, 副教授, 博士, 主要研究方向: 语义 Web、软件工程。

间的关系,包括定义域、值域和属性关系,对查询进行扩展。

OntoQA^[10]使用 WordNet 中的近义词、上位词和下位词对用户输入关键词进行查询扩展,调用 Swoogle 的 Web APIs 进行搜索。类似的工作还有 Jones^[11-14],与 OntoQA 的不同之处在于, Jones 调用 Google Web APIs 进行本体搜索。OntoSearch^[15]则根据用户输入的关键词调用 Google Web APIs 搜索本体,不对用户输入进行扩展。

综上所述,不管 Swoogle、Google 或其他搜索方式,共性都是基于关键词的匹配方式,不同点在于有没有对用户输入的关键词进行扩展。不管哪种搜索方式,如果搜索执行前不对查询关键词进行扩展,结果都已经被证明是很低效的,搜索结果中往往包含大量无关的文档,这是因为用户在搜索时输入的关键词很少,不足以表达用户的搜索需求。而 Swoogle 通过增加约束、OntoFetcher 通过增加关键词间定义域、值域和属性关系的行为,要求用户具有丰富的领域知识,且并不能准确表达用户的搜索需求,也不适用于普通用户。基于 WordNet 的搜索扩展,通过引入近义词,上位词和下位词集合,增加了关键词之间的关系和语义,在一定程度上增强了用户搜索的需求表达。但也存在着以下问题:1)搜索工具严重依赖于知识库,WordNet 是一个有限词典的知识库,当用户输入的关键词不属于知识库时,导致无法扩展;2)脱离了具体应用的扩展具有很强的盲目性,一方面,扩展的层次不宜确定,另一方面,不管以何种方式扩展都将导致搜索范围和搜索结果的极度扩大,这将会引起搜索准确率的下降,增加用户甄别的工作量。鉴于以上分析,本文在研究本体搜索需求和用户搜索行为的基础上,提出了一种无需知识库的相关关键词启发式扩展搜索机制(Heuristic Ontology Search Mechanism, HOSM),该机制可以克服上述方法的缺点,能够根据不同用户的搜索行为进行针对性的扩展,极大地提高本体搜索的准确率。

2 基于用户行为的启发式本体搜索机制

2.1 本体搜索需求和用户行为分析

本体主要包括类、类的属性和类间关系,类及其之间的关系通常被认为是一种图式结构,类表示顶点,关系表示边。在进行搜索时,除了考虑类和属性的匹配外,用户还希望满足搜索要求的领域本体定义了尽可能丰富的类间关系。理想情况下,在进行本体搜索时,用户接口提供一种机制,描述一个本体的子图来表达用户的搜索语义。但由于类与关系相互依存,且关系复杂,在进行搜索时,用户输入的关键词不一定都匹配本体中的类,即使匹配,这些类在本体中也不一定存在直接关系,而且也缺乏描述这种关系的方法,因此很难提供这样的接口去描述用户的搜索语义,这就决定了目前的本体搜索引擎都只能采用基于关键词的搜索方式。

本体搜索与普通文档搜索的重要区别在于,在进行普通文档搜索时不需要考虑匹配词之间的关系,只要存在匹配就满足搜索要求,而本体是一种领域性很强的文档,由于领域交叉性和词汇多义性的存在,增加了本体搜索的难度,为了提高搜索时领域的分辨精度,需要用于搜索的关键词集合能够尽可能覆盖某一领域的语义。通常情况下,用户搜索时输入的关键词都比较少,因此在搜索执行前需要采用某种机制,对用户输入的关键词进行扩展,使其能够表达用户的领域搜索需求。

同时,由于本体中类和属性的命名都采用领域专有术语,而软件工程师和普通用户通常缺乏相关的领域知识。因此在搜索时,用户输入的关键词不一定是类或属性的领域专用术语,可能是其同义词或近义词,这就决定了在进行本体搜索时,为了提高搜索的准确性,需要扩展后的搜索关键词集合和待考查本体有尽可能多的类和属性相匹配,这就要求用于扩展的关键词一方面和用户输入的关键词具有某种语义关联,另一方面应该尽可能是领域专有术语。

从用户搜索的角度看,即使执行同一应用领域的本体搜索,不同用户输入的搜索关键词不可能完全相同。这是因为,针对同一领域,不同用户由于领域背景知识不同,关注面和兴趣不同,用户在搜索时只能输入自己熟知的关键词,而不同用户的熟知关键词不可能完全相同;另一方面,由于忽略或遗忘等原因,用户通常不能输入足够多的关键词,使之能够表达清楚自己的搜索需求。而针对同一领域搜索的多个不同用户输入的不同关键词,具有领域共性,弥补了单个用户输入搜索关键词少的不足,组合覆盖能力可以增加领域的分辨率,增强用户搜索需求的表达能力。由此,为了提高搜索结果的准确性,可以利用不同用户输入的不同关键词实现搜索关键词的启发式扩展,提高领域分辨率,以及和领域本体的匹配度。为了便于说明,先作以下定义。

定义1 给定关键词 kw (用户输入), 关键词集合 $kwset$, 如果 $kw \in kwset$, 对于任意给定的 $kw_i, kw_i \in kwset$ 并且 $kw_i \neq kw$, 那么称 kw_i 是 kw 的相关关键词, $kwset$ 是 kw 的相关关键词集合。

定义2 给定关键词集合 $kwset$, 且 $1 \leq |kwset| \leq n$, $|kwset|$ 表示集合中的关键词数, 如果使用 $kwset$ 的任意非空子集作为关键词进行搜索, 总可以得到本体文档 O_i , 称 $kwset$ 是 O_i 的副索引。

定义3 使用给定的关键词集合 $kwset$ 作为本体搜索的关键词集合, $1 \leq |kwset| \leq n$, $|kwset|$ 表示集合中的关键词数, O 表示最终的搜索结果, 即本体文档集合, 对于 $O_i \in O$, $kw_1, kw_2, \dots, kw_m (m \leq n) \in kwset$, 在搜索过程中如果是因为 $kw_1, kw_2, \dots, kw_m \in O_i$ 而得到 O_i , 则称 kw_1, kw_2, \dots, kw_m 是 O_i 的种子。

2.2 搜索机制描述

基于用户行为的启发式本体搜索机制分为两个阶段, 第一阶段是预处理阶段, 为数据库中存储的每一个本体增加一个空字段, 即创建副索引, 初始状态为空, 第二阶段是搜索处理阶段。算法的核心思想是, 如果用户输入的关键词匹配于某个本体中的类或属性, 就把这个关键词记录到该本体的副索引中, 经过一定量的搜索后, 本体的副索引中记录了用户搜索该本体时常用的关键词, 其后用户搜索时, 系统先搜索副索引, 获取相关关键词集合, 即输入关键词如果匹配于本体副索引中某个词, 则将该副索引中的其他关键词返回给用户, 供用户选择扩展, 再执行本体搜索, 最后, 把与搜索结果集中本体匹配, 但不在副索引中的新词, 即本体的种子加入到本体的副索引中, 算法处理过程如图1所示, 算法描述如下。

1) 用户输入自己熟知的或感兴趣的关键词集合。

2) HOSM 查询数据库本体的副索引, 通过用户接口返回用户输入关键词的相关关键词集合, 供用户扩展, 如果副索引为空, 则直接转向4)。

3)用户从返回的相关关键词中选择感兴趣的或可能属于潜在领域的由于认知被忽略或遗忘的关键词,形成扩展后的搜索关键词集合。

4)HOSM 使用扩展后的搜索关键词集合执行本体搜索;当搜索关键词集合中的元素多于一个时,本文档应匹配一半集合中的元素作为满足搜索的条件。

5)对于结果集中的每个文档,如果扩展搜索关键词集合中的某个关键词是其种子,并且该种子不在文档的副索引中,则将种子关键词写入到该文档的副索引中。

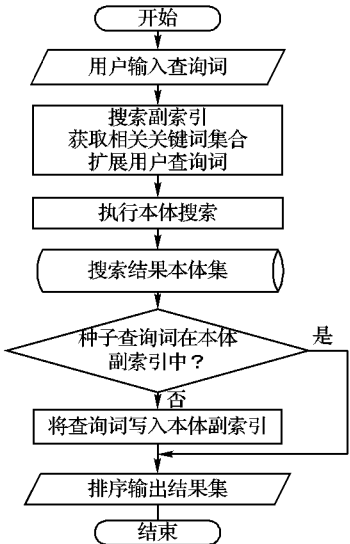


图 1 基于用户行为的启发式本体搜索算法流程

从上述算法步骤中可知,基于用户行为的启发式本体搜索机制具有三个方面的优点,第一,利用不同用户对同一领域本体搜索时输入的不同关键词进行扩展,增加了扩展的针对性,避免了扩展的盲目性,摆脱了使用知识库的依赖性。第二,由于副索引中的词都是匹配本体的领域术语,扩展的结果

必将增加与本体的匹配度,提高搜索的准确性。第三,这个过程是一个循环地自我学习地过程,随着搜索用户数的增加,库中本体的副索引将不断地被相关关键词填充,用户输入关键词的扩展针对性也越来越强,搜索的准确率必然会得到很大提高。副索引在搜索过程中的作用是为用户提供自助扩展的备选关键词,本体的搜索仍然是在本体的元数据索引中进行。

3 实验及结果分析

为了获得实验所需的数据集,开发了 HOSM 原型系统,调用 Swooge Web APIs 分别使用表 1 中第一行所示的单词作为关键词进行搜索和下载,除去无效的和重复的 URL,共得到 RDF 和 OWL 本体 602 个,分析元数据并建立索引,存储在 MySQL 数据库中。相关的实验环境包括:双核 Intel T6570 CPU 2.1 GHz,2 GB 内存,160 GB 硬盘,Window XP 操作系统,apache-tomcat 5.5.27,jdk1.6.0_21,Jena。

为了验证基于用户行为的启发式本体搜索机制的性能,进行了两项实验:1)挑选了 20 名学生志愿者,让他们依次使用我们提供的用户搜索接口,根据他们自己的认知分别输入关键词搜索关于 cancer 的本体,用以检验 HOSM 的准确率和召回率,如式(1)、(2)所示。其中:PC 表示准确率,TP 表示搜索结果集中相关文档的数量,FP 表示搜索结果集中不相关文档的数量,RC 表示召回率,FN 表示数据集中未返回的相关文档的数量。2)使用 WordNet 对搜索关键词进行扩展后搜索,比较其准确率和召回率。

$$PC = TP / (TP + FP) \quad (1)$$

$$RC = TP / (TP + FN) \quad (2)$$

表 1 第 2 行综合了 20 名用户搜索时使用的关键词,第 3 行给出了使用 WordNet 扩展后的搜索关键词集合。表 2 列出了使用 WordNet 扩展搜索和 HOSM 扩展搜索的准确率与召回率,其中 Hxx 表示第 xx 个用户的搜索,WN 表示使用 WordNet 的扩展搜索。

表 1 实验中使用的搜索关键词

说明	关键词集合
调用 Swooge Web APIs 本体搜索关键词	medicine, treatment, health, doctor, care, blood, disease, pain, hospital, surgery, department, death, chemotherapy, operation, metastasis, lung cancer, lymph, breast, stomach, cancer, culture, education, sport, hygiene
20 名志愿者输入的不同关键词	medicine, cancer, chemotherapy, liver cancer, lymph, lung cancer, death, stomach cancer, breast cancer, metastasis, patient, doctor, operation, malignant tumor, leukemia, cervical cancer, benign tumor, tumor, cancer cell, hospital, treatment, disease, carcinoma
使用 WordNet 扩展后的搜索关键词 ^[11]	cancer, cell, tumor, patient, document, carcinoma, lymphoma, disease, access, treatment, skin, liver, leukemia, risk, breast, genetic, tobacco, thymoma, malignant, gene, clinical, neoplasm, pancreatic, Tissue, therapy, lesion, blood, study, thyroid, smoking, polyp, human, health, exposure, studies, ovarian, information, research, drug, related, associated, neoplastic, oral, bone, chemotherapy, body, oncology, growth, medical, lung

图 2 是第 20 名志愿者输入关键词进行搜索时的 HOSM 用户搜索接口,从图中可以看出,当用户输入关键词时,其相关关键词被系统返回给用户,让用户可以有针对性进行选择扩展。图 3 展示了搜索结果集中本体数量随搜索用户数量的变化情况,图 4 展示了搜索结果准确率与召回率随搜索用户数量的变化情况。

从图 2 中可以看出,随着搜索用户数量的增加,搜索结果集中文档的数量逐渐减少,表明随着用户扩展关键词和本体匹配关键词的增加,无关本体在逐渐减少;从图 3 中可以看

出,当搜索的用户数达到一定量时,HOSM 具有较高且相对稳定的准确率和召回率,并且随着搜索用户数量的增加,准确率逐渐提高且趋于稳定,而召回率略有下降并趋于稳定,这是因为随着搜索关键词的增加,系统要求搜索结果匹配的关键词数量在增加,可能会导致部分匹配数量较低的本体被过滤掉。

从表 2 中看到,WN 的准确率较低而召回率达到了 1,这是因为,WordNet 扩展后的关键词集合包含了较多的与用户搜索有关的关键词,同时也包含了大量领域无关的词汇,导致搜索范围扩大,结果集包含了大量的无关文档。

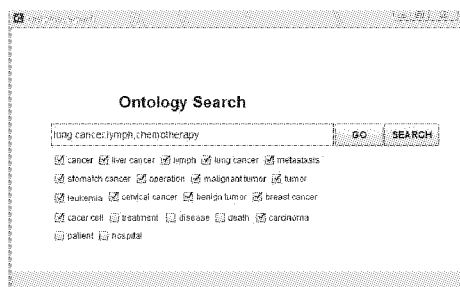


图2 HOSM 用户搜索接口

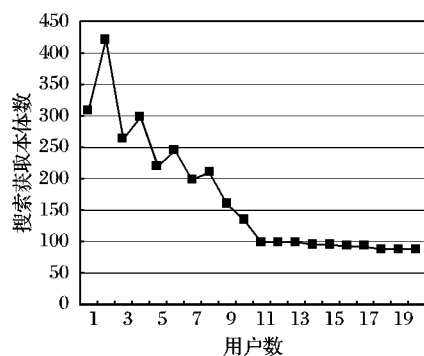


图3 HOSM 搜索结果与用户变化曲线

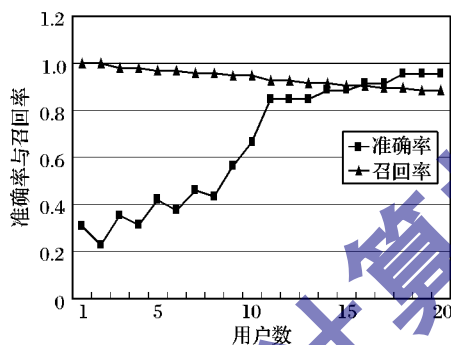


图4 HOSM 准确率与召回率与用户变化曲线

表2 两种扩展搜索的准确率与召回率

方法	准确率	召回率
H11	0.8485	0.9263
H12	0.8484	0.9263
H13	0.8485	0.9158
H14	0.8842	0.9158
H15	0.3755	0.9053
H16	0.9130	0.9053
H17	0.9130	0.8947
H18	0.9545	0.8947
H19	0.9545	0.8842
H20	0.9545	0.8842
WN	0.3755	1.0000

4 结语

本体搜索机制对于本体发现和重用都具有重要的意义。现有的本体搜索都是基于关键词的搜索,利用 WordNet 进行扩展,这种扩展具有盲目性,造成了搜索范围的扩大,在提高召回率的同时,降低了准确率。本文在研究本体搜索需求及用户搜索行为的基础上,提出了一种基于用户行为的启发式本体搜索机制,使用不同用户输入的具有领域共性的关键词,启发式地扩展搜索。领域共性使得扩展具有很强的针对性,避免了扩展的盲目性,同时由于扩展词都是领域术语,提高了

搜索时的匹配度。实验证明,随着搜索用户数的增加,HOSM 取得了较高的且较为稳定的准确率和召回率,在本体搜索中是切实有用的。

参考文献:

- [1] BERNERS-LEE, HENDLER J, LASSILA O. The semantic Web [J]. Scientific American, 2001, 284(5): 34-43.
- [2] ZHANG Y, VASCONCELOS W, SLEEMAN D. Ontosearch: An ontology search engine[C] // Proceedings of 24th SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence. Berlin: Springer-Verlag, 2004: 145-156.
- [3] DAML Ontology Library [EB/OL]. [2012-02-01]. <http://www.daml.org/ontologies>.
- [4] Ontolingua. Ontolingua ontology library[EB/OL]. [2012-03-25]. <http://Ontolingua.stanford.edu>.
- [5] Protégé. protege ontology library [EB/OL]. [2012-04-05]. http://protegewiki.stanford.edu/wiki/Protege_Ontology_Library.
- [6] DING L, PAN R, FININ T. Finding and ranking knowledge on the semantic Web [C] // ISWC'05: Proceedings of the 4th International Conference on the Semantic Web. Berlin: Springer-Verlag, 2005: 156-170.
- [7] ALANI H, BREWSTER C. Ontology ranking based on the analysis of concept structures [C] // Proceedings of the 3rd International Conference on Knowledge Capture. New York: ACM, 2005: 51-58.
- [8] PATEL C, SUPEKAR K, LEE Y, et al. OntoKhoj: A semantic Web portal for ontology searching, ranking, and classification[C] // Proceedings of 5th ACM International Workshop on Web Information and Data Management. New York: ACM, 2003: 58-61.
- [9] SHAH S A H, KHALID A, QADIR M A. OntoFecher: An approach for query generation to gather ontologies and ranking them by ensuring user's context [C] // Proceedings of the 4th International Conference on Emerging Technologies. Islamabad: [s. n.], 2008: 247-252.
- [10] TARTIR S, ARPINAR I B, MOORE M. OntoQA: Metric-based ontology quality analysis[C] // Proceedings of IEEE Workshop on Knowledge Acquisition from Distributed, Autonomous, Semantically Heterogeneous Data and Knowledge Sources. Houston: IEEE, 2005: 45-53.
- [11] JONES M, ALANI H. Content-based ontology ranking [C] // Proceedings of the 9th International Protégé Conference. Stanford, USA: Stanford University, 2006: 96-99.
- [12] PAN J Z, THOMAS E, SLEEMAN D. Ontosearch2: Searching and querying Web ontologies [C] // Proceedings of the IADIS International Conference. San Sebastian, Spain: [s. n.], 2006: 42-49.
- [13] AQUIN M, SABOU M, DZBOR M, et al. Watson: A gateway for the semantic Web [C] // Poster Session of the 4th European Semantic Web Conference. Innsbruck: ESWC, 2007: 3-7.
- [14] CHENG G, GE W Y, QU Y Z. Falcons: Searching and browsing entities on the semantic Web[C] // Proceedings of the 17th International Conference on World Wide Web. New York: ACM, 2008: 1101-1102.
- [15] BUITELAAR P, EIGNER T, DECLERCK T. OntoSelect: A dynamic ontology library with support for ontology selection[C] // Proceedings of the Demo Session at the International Semantic Web Conference. New York: ACM, 2004: 54-57.