

# 一种面向微生物领域的知识库构建方法

陈 航<sup>1,2</sup>, 张 辉<sup>1,2</sup>

(1 北京航空航天大学 国家科技资源共享服务工程技术研究中心, 北京 100191;

2 北京航空航天大学 计算机学院, 北京 100191)

**摘 要:** 从构建面向微生物领域的知识库出发, 研究了微生物领域的知识词典构建算法, 在此基础上实现了语义扩展检索功能, 设计并实现了微生物知识检索系统, 对系统的查全率和查准率两个方面进行了实验分析. 实验结果表明, 与传统的关键词匹配技术相比, 基于知识库的语义检索技术在查全率方面提高了 72%, 在查准率方面提高了 5%.

**关键词:** 微生物; 知识库; 语义信息; 语义检索; 领域知识词典; 语义扩展检索

**中图分类号:** TP391

**文献标识码:** A

**文章编号:** 1000-7180(2016)08-0005-05

## A Method for Constructing the Knowledge Base in the Field of Microbiology

CHEN Hang<sup>1,2</sup>, ZHANG Hui<sup>1,2</sup>

(1 National Engineering Research Center for S&T Resources Sharing Services, Beihang University,

Beijing 100191, China; 2 School of Computer Science, Beihang University, Beijing 100191, China)

**Abstract:** This paper starts from the construction of knowledge base in the field of microbiology, study the domain knowledge dictionary construction algorithm, and implement semantic retrieval function on this basis. We designed and implemented a knowledge retrieval system and analyzed the recall and precision of the system. The results showed that compared with the keyword-based retrieval, semantic retrieval based on the knowledge base improved the recall of 72% and 5% in precision.

**Key words:** microbe; knowledge base; semantic information; semantic retrieval; domain knowledge dictionary; semantic extension retrieval

### 1 引言

随着互联网技术的发展, 相比于传统数据库, 知识库已经成为计算机中各领域的研究热点. 人工智能和数据库两项计算机技术的有机结合, 促成了知识库系统的产生与发展. 本文根据微生物领域专家对于知识检索的需要, 研究如何构建微生物领域知识库, 并进一步实现语义检索功能.

本文将微生物领域概念及其关系的集合称为微生物领域知识词典. 为了构建微生物领域知识词典, 首先需要将网络上组织方式多样微生物信息进行必

要的整合. 为此, 需要使用网络爬虫抓取微生物相关文档. 然后, 使用随机步长模型和马尔科夫聚类等相关算法, 从文档中抽取微生物相关概念及关系. 领域知识词典构建完成后, 本文进一步提出了基于关键词的语义扩展检索方法, 实现了语义检索功能.

### 2 相关研究

为了处理信息孤岛问题, 文献[1]提出了一种基于领域本体的语义检索和推理系统, 系统提供了三个用户接口. 用户可以从检索接口中选择一个检索项进行检索, 也可以从查询接口中输入或者选择一

个查询提交到 SPARQL 处理模块中, 用户还可以在推理规则接口选择或者输入规则信息, 根据已有信息进行推理. 系统主要基于课程领域本体, 使用 SPARQL 语言和 Jena API 实现查询和推理.

文献[2]提出了领域本体驱动的模糊语义检索, 使用本体来理解用户所要查询的信息. 系统的用户接口仍然接收用户输入的关键词, 查询处理模块首先将关键词进行模糊匹配, 然后转化为 SPARQL 查询, 并返回处理结果.

文献[3]使用了对文档进行语义标注的方法实现了基于本体的语义检索. 首先, 用本体实体对文档进行标注, 即记录下对于每个本体实体, 哪些文档出现了这些实体. 这样, 在检索的时候, 将用户关键词转化为本体实体, 再进一步进行搜索, 即可得到相关文章. 这种思路对于本文语义扩展检索的实现有很大的借鉴作用.

文献[4]在使用了本体和 RDF 等技术的基础上实现了一个语义网络搜索引擎. 系统首先将用户的输入转化为本体中的类或实例, 然后对类或实例根据本体树中的上下位关系进行扩展, 最后进行搜索得出检索结果. 这篇文章中使用本体对用户输入关键词进行上下位关系扩展, 对于本文实现语义扩展检索有借鉴作用.

文献[5]利用模糊逻辑理论和语义检索技术构建了一个模糊搜索引擎. 在使用网络爬虫抓取网页时, 首先根据领域对网页进行分类. 接着构建一个模糊本体, 并计算本体中不同词项之间的相似度. 用户检索时可以选择一个合适的领域并输入带有不同权重的关键词. 之后, 对关键词使用模糊本体进行扩展, 来发现同义词或类似的词项, 最后搜索结果根据关键词的重要程度进行排序. 文章中使用模糊本体进行对关键词扩展的方法, 对于本文实现的对关键词模糊匹配具有启发作用.

文献[6]对网页信息进行预处理并进行语义标注, 使用本体对用户输入的关键词进行匹配, 并进行上下位关系、同位关系、相关属性扩展, 使用扩展后的结果对标注好的网页信息进行检索, 得到检索结果.

文献[7]利用文献[8]中构建本体的方法构建领域本体, 并进一步实现语义检索功能. 文献[8]利用基于图表示的方法实现领域本体的构建. 首先, 对每篇文档进行分词, 将分词后的词项作为顶点, 词项之间的关系作为边, 这样就将每篇文档表示成了一个图. 接着, 利用随机步长模型计算每个词项的权重,

并使用马尔科夫聚类方法对词项进行聚类, 这样每篇文档就表示成一些概念及概念之间的关系. 进一步将这些概念及其之间的关系抽取出来, 完成本体的构建. 这篇文章中构建本体的方法对本文领域词典的构建起到了很大的启发作用. 构建完成本体后, 文献[7]用语义扩展的方法实现了语义检索功能.

综上所述, 这些搜索引擎大多使用领域本体来完成对关键词的语义扩展功能. 而本文使用了领域知识词典来实现领域本体的功能, 从而省去了使用相关工具的麻烦. 在语义检索环节, 本文也没有使用 SPARQL 等复杂的查询语言, 这样可以方便用户使用、简化系统并提高检索速度. 最后, 本文借鉴了相关文献中提到的一些方法来构建领域知识词典, 并实现了语义检索功能.

### 3 微生物领域知识库的构建

微生物知识库的构建包含两个部分, 一部分是微生物相关信息的获取, 另一部分是领域知识词典的构建. 微生物相关信息获取主要是使用爬虫从相关专业网站中爬取所需数据, 然后存储到数据库中. 领域知识词典的构建, 使用了将文档表示成图的办法, 然后对图进行进一步处理, 抽取微生物相关概念及概念间的关系. 系统流程图如图 1 所示.

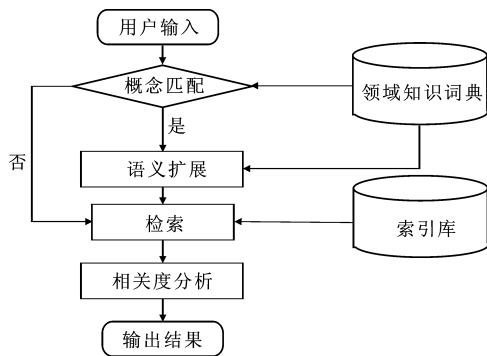


图 1 系统流程图

#### 3.1 微生物相关信息的获取

要构建微生物领域知识库, 首先需要获取微生物领域的相关信息. 系统使用 Jsoup 网页解析工具, 对微生物专业网站上的信息进行解析, 并爬取所需信息存储到数据库中. 然后使用 Solr 对数据库中的数据建立索引, 以供检索的时候使用. 微生物知识库数据架构如图 2 所示.

数据主要来源于知网、万方、百科以及 NCBI 网站中的专业数据.

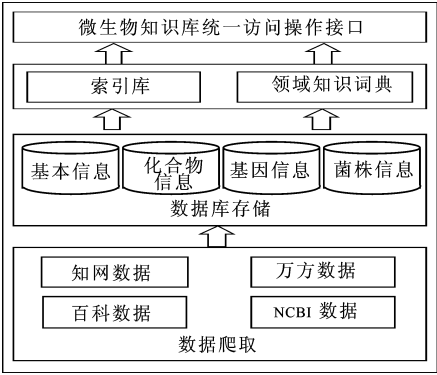


图 2 数据架构图

3.2 领域知识词典的构建

微生物领域知识词典是微生物领域概念以及概念之间关系的集合,是对用户输入的关键词进行语义扩展,实现语义检索的基础. 本文使用领域知识词典的构建主要分为以下五个步骤.

Step1 对每篇文档的摘要进行分词,对分词后得到的词项进行过滤,只留下动词和名词. 以这些词项作为顶点,若两个词项在原来的文档中相邻,则对对应的词项之间用边相连,这样每篇文档摘要就被表示成一个无向图,注意,摘要中重复出现的词,在图中只用一个顶点表示.

Step2 统计每个词项以及词项之间关系在原文档摘要中出现的次数,将得到的次数分别除以所有词项和关系出现的总次数,得到每个词项以及词项之间关系出现的频率.

Step3 注意到第二步产生的数值仅代表词项或者词项之间关系出现的次数,所以为了计算每个词项以及词项之间关系的权重,系统使用了随机步长模型,进行迭代计算每个词项的权重,并进一步得到词项之间关系的权重. 随机步长模型的迭代公式如下:

$$WS(V_i) = (1 - d) + d \times \sum_{V_j \in \text{Neighbour}(v_i)} \frac{w_{ji}}{w_i} WS(V_j) \tag{1}$$

$$W_i = \sum_{V_j \in \text{Neighbour}(v_i)} W_{ki} \tag{2}$$

式中, $WS(V_i)$  表示  $V_i$  迭代后的权重; $d$  为阻尼系数,本文中取 0.5; $W_{ji}$  表示连接点  $j$  和点  $i$  的边的权重; $W_i$  表示所有和  $i$  相邻的边的权值之和. 词项之间关系的权重计算公式为:

$$WE(e_{i,j}) = \frac{\sqrt{WS(V_i)^2 + WS(V_j)^2}}{|\sqrt{WS(V_i)}| + |\sqrt{WS(V_j)}|} \tag{3}$$

Step4 第三步完成以后,就可以将文档摘要所

表示成的无向图,表示成一个矩阵. 为了抽取出文档中出现的概念以及概念之间的关系,需要使用马尔科夫聚类算法对上述矩阵进行进一步迭代,最终得到聚类之后的结果. 这样,每篇文档摘要就被表示成了一些词类,每个词类中包含了相关联的词项.

Step5 对聚类之后的结果进行分析处理,抽取每篇文档摘要中出现的概念,以及概念之间的关系,构造出微生物领域知识词典.

下面以描述抗菌素的摘要为例,介绍领域知识词典的构建过程. 描述抗生素的摘要如下:

“抗菌素是微生物代谢过程中所产生的杀菌物质,具有抑制其他微生物的生长发育,以及阻碍其生理机能的作用. 迄今为止,世界上已发现的抗菌素有 2 000 多种,大部分属放线菌的分泌产物,其中约 50 种用于医药卫生方面,称医用抗菌素;有 10 多种用于农业,称农用抗菌素简称农抗. 如使用比较普遍的有链霉素、井冈霉素、内疗素、庆丰霉素、灰黄霉素、放线菌酮、灭瘟素等,都是当前应用较广的主要抗菌素品种.”

通过对摘要进行分词,保留名词和动词,并出现频率较高的词语. 将词语表示为结点,词语之间的相邻关系表示为边,可将摘要表示为图 3 所示.

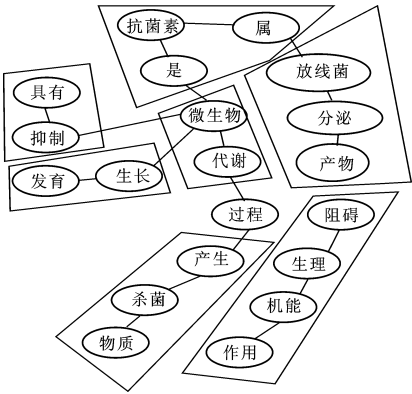


图 3 基于图的文档表示

通过计算词语及词语间关系的权重,可将图 3 所示的无向图转换为邻接矩阵. 通过使用随机步长模型和马尔科夫聚类算法对矩阵进行迭代,可以得到将词语进行聚类之后的结果. 图 3 使用方框显示了经过马尔科夫聚类之后的结果. 可见,摘要所对应的图经过矩阵的迭代赋予权值,并经过聚类之后,不同的词项被归位了不同的类别. 通过图 3 可以归纳出的概念有抗菌素、微生物、代谢、放线菌、分泌产物、产生杀菌物质、阻碍生理机能、抑制、生长发育等,并可以得到这些概念之间的关系. 通过这样的方

法对每篇文档的摘要进行同样的分析和计算,就可以得到大量的概念以及概念之间的关系,进而可以构建出领域知识词典.

#### 4 语义检索的实现

语义检索的实现主要是利用已经构建好的领域知识词典对用户输入的关键词进行语义扩展检索并检索的过程. 为了加快检索速度,并实现模糊匹配,可以事先对存储在数据库中的领域知识词典建立索引. 语义检索具体实现步骤如下:

Step1 对于用户输入的关键词,首先使用领域知识词典中的概念对关键词进行模糊匹配,即对用户输入的关键词到词典索引中进行检索,若无匹配,则直接进行关键词检索. 若有匹配,则得到匹配后的概念;

Step2 根据第一步匹配得到的概念在词典事先定义好的关系中进行语义扩展,主要是同义关系和相关关系的扩展,即获取到和该概念相近的所有概念,得到扩展后的概念集;

Step3 对于第二步中得到的概念集,可以根据事先记录好的每个词在所有文档中出现的频率分配不同的权值;

Step4 使用第三步中产生的带有不同权值的概念,在索引库中进行搜索,最后得到搜索结果列表.

语义检索的实现流程如图 4 所示.

$$\text{查准率} = (\text{检索出的相关信息量} / \text{检索出的信息总量}) \times 100\%$$

前者是衡量检索系统和检索者检出相关信息的能力,后者是衡量检索系统和检索者拒绝非相关信息的能力. 两者合起来,即表示检索效率.

为了验证检索效果,使用了基于关键词匹配和基于知识库的语义扩展检索两种方法,随机选取了专业领域的五个关键词,利用已经爬取好的 2500 多条万方数据,做了五组对比试验. 实验结果如表 1、2 所示. 表 1 显示了基于关键词匹配和基于语义扩展方法的查准率情况.

表 1 检索结果查准率

| 关键词  | 基于关键词匹配的查准率 | 基于语义扩展的查准率 |
|------|-------------|------------|
| 青霉素  | 0.445 9     | 0.548 9    |
| 放线菌酮 | 0.991 9     | 1          |
| 麦迪霉素 | 0.896 1     | 0.894 7    |
| 链霉素  | 0.345 1     | 0.472 7    |
| 土霉素  | 0.688 9     | 0.622 6    |

表 2 显示了基于关键词匹配和基于语义扩展方法的查全率情况.

表 2 检索结果查全率

| 关键词  | 基于关键词匹配的查全率 | 基于语义扩展的查全率 |
|------|-------------|------------|
| 青霉素  | 0.514 1     | 0.781 7    |
| 放线菌酮 | 0.276 1     | 0.414 1    |
| 麦迪霉素 | 0.133 3     | 0.270 6    |
| 链霉素  | 0.52        | 0.98       |
| 土霉素  | 0.22        | 0.413 3    |

从两个表中可以直观地看到,使用基于知识库的语义扩展检索,大大提高了检索结果的查全率和查准率.

从实验结果可以看出,基于知识库的语义扩展检索方法明显优于基于关键词匹配的技术. 表明了我们使用的方法对于传统的检索方法进行了一定的改进,并且得到了满意的效果. 另外,由于对关键词扩展时加入了和关键词相关的语义信息,因此,从检索结果中可以看出文档的相关性更大,文档中所包含的用户关心的信息更加丰富,因此,基于知识库的语义扩展检索更能够满足用户的需要.

#### 5 实验结果

查全率和查准率是评估搜索引擎搜索结果好坏的两个重要指标,本文使用这两个指标来评价检索系统的检索效果. 查全率和查准率的定义如下:

$$\text{查全率} = (\text{检索出的相关信息量} / \text{系统中的相关信息总量}) \times 100\%$$

#### 6 结束语

本文主要研究了微生物领域知识库的构建及在此基础上如何实现语义检索功能. 本文采用了基于将文本表示成图,并进一步对图中的词项进行聚类的方法完成了概念的抽取,并构建出了微生物领域

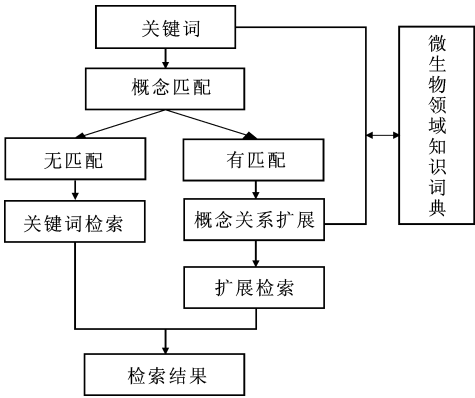


图 4 语义检索流程图

知识词典. 语义检索的实现主要是利用词典中的同义关系和相关关系对用户输入的关键词进行语义扩展,进一步得到检索结果.

本文研究在抽取概念时,没有考虑各词项之间的复杂关系. 并且抽取出来的概念仍然不够精确,有些概念与关键词之间的关系不够密切. 对于领域知识词典的构建方法还需进一步研究. 另外,知识库的构建不是一蹴而就的,需要经过反复的更新迭代才能达到令人满意的效果. 因此,下一步工作还需要对微生物知识库进行持续的扩充.

## 参考文献:

- [1] Song Wanli, Ni Lixian. Semantic query and reasoning system based on domain ontology[C]// International Symposium on Computers & Informatics. France, Atlantis Press, 2015.
- [2] Remi S, Varghese S C. Domain ontology driven fuzzy semantic information retrieval[C]// International Conference on Information and Communication Technologies. Holland, Procedia Computer Science, 2015.
- [3] Miriam Fernández, Iván Cantador, Vanesa López, et al. Semantically enhanced information retrieval: an ontology-based approach[J]. Web Semantics: Science, Services and Agents on the World Wide Web, 2011, 9

(4):434-452.

- [4] Debajyoti Mukhopadhyay, Aritra Banik, Sreemoyee Mukherjee, et al. A domain specific ontology based semantic web search engine[C]// International Workshop on MSPT Proceedings. Jeonju, 2007.
- [5] Lai Lienfu, Wu Chaochin, Lin Peiying, et al. Developing a fuzzy search engine based on fuzzy ontology and semantic search [C] // International Conference on Fuzzy Systems. Taiwan, IEEE, 2011.
- [6] Tao Tengyang, Zhao Ming. An ontology-based information retrieveval model for vegetables E-commerce[J]. Journal of Integrative Agriculture, 2012, 5 (11): 800-807.
- [7] Xutang Zhang, Xin Hou, Xiaofeng Chen, et al. Ontology-based semantic retrieval for engineering domain knowledge[J]. Neurocomputing, 2013, 116 (20): 382-391.
- [8] Xin Hou, Ong S K, Nee A Y C, et al. GRAONTO: a graph-based approach for automatic construction of domain ontology[J]. Expert Systems with Applications, 2011, 38(9): 11958-11975.

## 作者简介:

陈 航 男, (1991-), 硕士研究生. 研究方向为知识库、知识图谱. E-mail: chenhang@buaa.edu.cn.

(上接第4页)

- [3] Song H. Protocol-oblivious forwarding: Unleash the power of SDN through a future-proof forwarding plane [C] // Proceedings of the second ACM SIGCOMM workshop on Hot topics in software defined networking. China, Hongkong, ACM, 2013: 127-132.
- [4] Bosshart P, Daly D, Gibb G, et al. P4: Programming protocol-independent packet processors [J]. ACM SIGCOMM Computer Communication Review, 2014, 44(3): 87-95.
- [5] Rossi D, Rossini G. On sizing CCN content stores by exploiting topological information [C] // INFOCOM Workshops. USA, Florida, Orlando, 2012: 280-285.
- [6] Wang Y, Li Z, Tyson G, et al. Optimal cache allocation for content-centric networking [C] // Network Protocols (ICNP), 2013 21st IEEE International Conference on. Germany, Göttingen, IEEE, 2013: 1-10.
- [7] Dong L, Zhang Y, Raychaudhuri D. Enhance content broadcast efficiency in routers with integrated caching [C] // Computers and Communications (ISCC), 2011 IEEE Symposium on. Greece, Kerkyra, IEEE, 2011:

320-322.

- [8] 刘外喜, 余顺争, 蔡君, 等. ICN 中的一种协作缓存机制 [J]. Journal of Software, 2013, 24 (8): 1947-1962.
- [9] Rossi D, Rossini G. Caching performance of content centric networks under multi-path routing (and more) [R]. Relatório técnico, Telecom ParisTech, 2011.
- [10] Psaras I, Chai W K, Pavlou G. Probabilistic in-network caching for information-centric networks [C] // Proceedings of the second edition of the ICN workshop on Information-centric networking. Finland, Helsinki, ACM, 2012: 55-60.

## 作者简介:

刘 静 女, (1991-), 硕士研究生. 研究方向为未来网络. E-mail: ljsmile@ustc.edu.cn.

王 雷 男, (1972-), 博士, 副教授. 研究方向为未来网络、云计算环境中的内容存储分发以及分布式系统优化.

高翔宇 男, (1991-), 硕士研究生. 研究方向为未来网络、SDN.

王 嵩 男, (1975-), 讲师. 研究方向为未来网络.