doi: 10.13413/j.cnki.jdxblxb.2018.02.24

基于抽取规则和本体映射的语义搜索算法

周诗源[1,2], 王英林[1]
(1. 上海财经大学 信息管理与工程学院, 上海 200433; 2. 嘉兴学院 计划财务处, 浙江 嘉兴 314001)

摘要: 针对目前语义搜索过程中存在效率低、用户推荐误差大等问题, 提出一种基于抽取规则和本体映射的语义搜索算法. 首先根据用户语义搜索要求抽取语义中的元素和属性, 解决数据利用率低的缺陷; 然后建立语义模型, 构建本体之间的元素及属性之间的映射, 消除用户需求和计算机之间的语义偏差; 最后将语义搜索算法应用于用户个性化推荐系统. 实验结果表明, 该语义搜索算法有效提高了搜索效率, 降低了用户个性化推荐误差.

关键词: 信息检索; 语义搜索; 本体映射; 抽取规则; 个性化推荐
中图分类号: TP393 **文献标志码**: A **文章编号**: 1671-5489(2018)02-0329-06

Semantic Search Algorithm Based on Extraction Rules and Ontology Mapping

ZHOU Shiyuan[1,2], WANG Yinglin[1]
(1. *School of Information Management and Engineering, Shanghai University of Finance and Economics, Shanghai* 200433, *China*;
2. *Office of Budget and Finance, Jiaxing University, Jiaxing* 314001, *Zhejiang Province, China*)

Abstract: In view of the problems of low efficiency and large user recommendation error in the process of semantic search, we proposed a semantic search algorithm based on extraction rules and ontology mapping. Firstly, according to user's requirements in semantic search, the semantic elements and attributes were extracted to solve the defect of low utilization rate of data. Secondly, we established a semantic model to construct a mapping between elements and attributes between ontology, and eliminated the semantic deviation between user's need and computer. Finally, the semantic search algorithm was applied to the user's personalized recommendation system. The experimental results show that the semantic search algorithm effectively improves the search efficiency and reduces the user's personalized recommendation error.

Key words: information retrieval; semantic search; ontology mapping; extraction rule; personalized recommendation

目前, 搜索引擎已成为人们获取信息的主要工具[1-5]. 搜索引擎对网络上的信息搜索实际是一种信

收稿日期: 2016-11-21.

作者简介: 周诗源(1982—), 男, 汉族, 博士研究生, 从事自然语言处理和机器学习的研究, E-mail: zhousypaper@sina.com. **通信作者**: 王英林(1962—), 男, 汉族, 博士, 教授, 博士生导师, 从事信息抽取、数据挖掘和软件工程的研究, E-mail: wang.yinglin@shufe.edu.cn.

基金项目: 国家自然科学基金(批准号: 61375053).

息的组织形式，其通过网络爬虫不断在网络上“爬行”，对网络上的信息进行周期性搜索，然后对搜索信息进行标记，建立一个供用户按关键词进行搜索的网页索引数据库[6-9]. 在实际应用中，由于网页中的信息组织形式与传统文本信息组织形式差别较大，如果采用传统文本信息的标记方式，则无法准确描述关键词间的语义关系，搜索效率极低，且搜索结果与用户真正需求相差较远，导致面对大量信息用户却找不到自己真正需要的信息[10-11]. 这主要是因为传统文本搜索算法无法真正理解用户的需求，无法实现用户内容的查询扩展、语义理解和关联缺失. 而语义网[12-14]可排除一切平台、语言的分歧，其中本体是语义网的核心，是对知识的共同理解和描述. 目前已有许多基于本体的语义搜索算法，如采用本体中定义的语汇作为关键标记文档；利用本体的路径进行用户查询的扩展，获得了较传统算法更好的搜索结果. 这些语义搜索算法缺乏统一的语义描述，计算机程序理解的语义与用户需求之间偏差较大，导致搜索时间长，无法实现智能搜索[15-17].

针对当前语义搜索算法存在的不足，本文提出一种基于抽取规则和本体映射的语义搜索算法. 首先根据用户搜索要求抽取语义中的元素和属性，然后建立计算机和用户之间的语义本体映射，并将语义搜索算法应用于用户个性化推荐系统. 实验结果表明，本文算法提高了语义的搜索效率，增强了用户个性化推荐系统的性能.

1 算法设计

1.1 语义网

语义网是以链接为主要形式的信息网，是对万维网的一种扩展，即增加了语言标记，其中语义是关键部分，主要用于描述符号及其对应对象之间的关系. 在语义网中，信息只有具备了语义，不同应用和用户之间才能实现互相操作[18].

1.2 本体的定义

在计算机科学领域，本体定义为一定领域词汇的基本定义和关系及其之间的规则，其目标是捕获相关领域的共同知识，并定义共同认可的术语，从而实现对领域知识的推理.

1.3 语义抽取规则

网页主要包括类结构、层次结构、对象、数据和基数5个元素. 网页关键词的语义由多个义原组成，因此语义抽取时，要综合考虑每个词的义原集，对于两个关键词 w_1 和 w_2，它们分别包括 n 个和 m 个义原，即 $s_{11},s_{12},\cdots,s_{1n}$ 和 $s_{21},s_{22},\cdots,s_{2n}$，则两个关键词 w_1 和 w_2 之间的相似度可描述为

$$\mathrm{Sim}(w_1,w_2)=\max_{i=1,2,\cdots,n;\ j=1,2,\cdots,m}\mathrm{Sim}(s_{1i},s_{2j}). \tag{1}$$

根据上下文之间关系建立树状层次体系描述义原，设两个义原 p_1 与 p_2 在树中的距离为 d，则其语义距离计算公式为

$$\mathrm{Sim}(p_1,p_2)=\frac{a}{d+a}, \tag{2}$$

其中 a 表示可调节参数.

对于两个关键词，若只从语义方面分析，则它们语义相似度可能很低，而它们出现在同一网页的概率非常大，因此本文还需从关键词的义原进行考虑，即考虑关键词之间的相关性. 对于一组关键词，首先得到每个关键词的义原集，并从第一个关键词不断遍历全部关键词与义原集，若找到一个关键词与其他关键词不相同，如果它们义原集有相同的部分，则提取它们的共同义原，并将两个关键词分别放在共同义原后；但若发现一个关键词在不同行中重复出现，即将其从重复的行中删除，根据领域词典判断每个义原是否与主题相关，不相关则将其对应行删除，即完成语义抽取.

1.4 本体映射

对于结构和内容相似度非常高的两个本体，它们均可用于两种类型的网页分类，但两个本体之间存在一定的语义冲突，若根据两对概念可建立本体的语义映射关系，则会得到错误的映射结果，因此即使两个本体片段的结构和内容很相似，由于语境不同也可能导致本体语义映射的差异性. 在本体映

射过程中需考虑概念语义信息，降低映射误差. 本文建立本体间语义映射关系时，综合考虑了与语义信息相关的知识，这些信息主要包括词语知识、领域知识和结构知识.

综合分析词语知识、领域知识和结构知识对本体映射的贡献，设计一种基于内容的本体映射方法，其工作流程如图 1 所示. 步骤如下：

1）输入两个本体模型；

2）提取两个本体的语义信息；

3）通过逻辑推理的方法建立本体的语义映射关系.

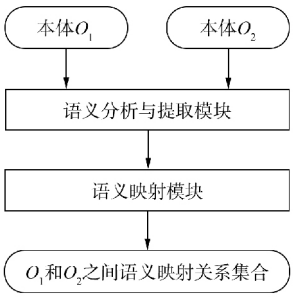

图 1　基于内容的本体映射流程

Fig.1　Ontology mapping process based on content

2　基于语义搜索算法的用户协同推荐系统

2.1　相似用户的发现

基于语义搜索算法的用户协同推荐系统的用户相似性计算方法很多. 设用户 a 和 b 对信息 i 的评分分别为 $d_{a,i}$ 和 $d_{b,i}$，用户共同评价过的信息数量为 $|I_{a,b}|$，用户 a 和 b 评价过的信息数目分别为 $|I_a|$ 和 $|I_b|$，则基于用户共同评价信息的相似性与基于用户共同评价过信息数量的相似性计算公式分别为

$$\mathrm{CPC}_{a,b}=\frac{\sum_{i=1}^{|I_{a,b}|}(d_{a,i}-\overline{d}_a)\times(d_{b,i}-\overline{d}_b)}{\sqrt{\sum_{i=1}^{|I_{a,b}|}(d_{a,i}-\overline{d}_a)^2}\times\sqrt{\sum_{i=1}^{I_{a,b}}(d_{b,i}-\overline{d}_b)^2}},\tag{3}$$

$$\mathrm{Jaccard}_{a,b}=\frac{|I_{a,b}|}{|I_a|+|I_b|-|I_{a,b}|},\tag{4}$$

其中 $\overline{d}_a$ 和 $\overline{d}_b$ 分别表示用户 a 和 b 对全部信息的平均评分[18]. 综合考虑式(3)，(4)，用户 a 和 b 的相似性计算公式为

$$\mathrm{Sim}_{a,b}=\mathrm{CPC}_{a,b}\times\mathrm{Jaccard}_{a,b}.\tag{5}$$

2.2　信任用户的确定

用户间的信任关系直接关系到推荐系统的性能优劣，根据用户邻居推荐是否成功确定用户之间的信任值. 邻居 b 向用户 a 推荐成功可描述为：a 与 b 对所推荐资源 i 评分之间的偏差小于信任阈值 ε，即

$$|d_{a,i}-d_{b,i}|\times\mathrm{Jaccard}_{a,b}<\varepsilon.\tag{6}$$

根据式(4)可初步得到用户间的信任关系，则用户 a 和 b 间的信任值计算公式为

$$\mathrm{MSD}_{a\to b}=1-\frac{\sum_{i=1}^{|I_{a,b}|}\sqrt{(d_{a,i}-\overline{d}_a)^2+(d_{b,i}-\overline{d}_b)^2}}{|I_{a,b}|\times\sum_{i=1}^{|I_{a,b}|}\left[\sqrt{(d_{a,i}-\overline{d}_a)^2}+\sqrt{(d_{b,i}-\overline{d}_b)^2}\right]}.\tag{7}$$

式(7)未考虑用户单独评价的信息，会使极少的信息评价产生极高的信任值，这与实际情况不符，因此要考虑用户的单独评价信息，用户 a 和 b 间的信任值计算公式变为：

$$\mathrm{DTrust}_{a\to b}=\begin{cases}\varepsilon\times\mathrm{Jaccard}_{a,b}, & \mathrm{MSD}_{a\to b}=0,\\ \mathrm{MSD}_{a\to b}\times\mathrm{Jaccard}_{a,b}, & \mathrm{MSD}_{a\to b}\neq 0.\end{cases}\tag{8}$$

式(8)表示用户之间的直接信任关系，但实际还存在一些间接信任关系的用户. 用户 a 信任用户 b，用户 b 信任用户 c，即可认为用户 a 在一定程度上信任用户 c，而间接信任可能存在多条信任路径，在考虑用户 a 和 c 之间的信任路径数($adj(a,c)$)基础上可得

$$\mathrm{ITrust}_{a\to c}=\frac{\sum\limits_{b\in adj(a,c)}\mathrm{DTrust}_{a\to b}\times(\mathrm{DTrust}_{b\to c}\times\beta_d)}{\sum\limits_{b\in adj(a,c)}\mathrm{DTtrust}_{a\to b}},\tag{9}$$

其中：d 表示两个用户间的距离；β_d 表示信任的衰减指数计算公式：

$$\beta_d=\frac{\mathrm{MPDist}-d+1}{\mathrm{MPDist}},\qquad d\in(2,\mathrm{MPDist}).\tag{10}$$

综上可知，用户 a 和 b 间信任值的最终计算公式为

$$\mathrm{Trust}_{a\to b}=\begin{cases}\mathrm{DTrust}_{a\to b}, & \beta_d=1,\\ \mathrm{ITrust}_{a\to b}, & 0<\beta_d<1.\end{cases}\tag{11}$$

2.3 计算用户信任权重

信任权值作为用户信任度的一个评价指标，有利于发现更多相似用户，改善用户推荐系统的性能，信任权重值计算主要考虑：用户间的信任值 $\mathrm{Trust}_{a\to b}$、用户评价声誉和语义相似性3个方面. 用户对某信息评价比例可间接描述用户在该信息推荐上的可信程度，因此用户对于某信息评价越多，表示该用户越熟悉该信息，对该信息进行推荐，该用户的信任权重值相对越大，则有

$$IR_b(IC_L)=\frac{\sharp N_b(IC_L)}{\sharp N(IC_L)},\tag{12}$$

其中：$\sharp N_b(IC_L)$表示用户评价过 IC_L 类信息的数量；$\sharp N(IC_L)$表示属于 IC_L 类信息的数量.

两个信息的语义相似性越大，对于成功向用户推荐过信息的邻居，其推荐该信息的可信度越大. 信息 i_x 和 i_y 的语义相似度与本体图中的深度($\mathrm{depth}(i_x)$,$\mathrm{depth}(i_y)$)和最近共同祖先(least common ancestor，LCA)的深度相关，计算公式为

$$\mathrm{SemSim}_{\mathrm{Hie}}(i_x,i_y)=\frac{\mathrm{depth}(\mathrm{LCA}_{i_x,i_y})}{\max\{\mathrm{depth}(i_x),\mathrm{depth}(i_y)\}}.\tag{13}$$

为充分考虑语义属性信息，引入语义相似性，此时用户信任权重值的计算公式为

$$TW_{a\to b}=w_1\times\mathrm{Trust}_{a\to b}+w_2\times IR_b(IC_L)+w_3\times\frac{1}{N}\times\Big(\sum_{j=1}^{N}\mathrm{SemSim}_{\mathrm{Hie}}(i_x,i_y)\Big).\tag{14}$$

根据信任权重值 $TW_{a\to b}$、相似用户的数量及推荐系统信任阈值 ε 之间的关系发现更多的相似用户，从而提高用户个性化推荐系统的推荐结果质量.

3 仿真实验

3.1 数据集

为了分析基于抽取规则和本体映射的语义搜索算法性能，在4核 Intel 3.0 GHz CPU，8 GB 内存的计算机上，采用 Windows7 作为操作系统，选择 JAVA 语言进行编程，采用 MovieLens 数据集作为研究对象. 为使实验结果更具说服力，采用文献[19]和文献[20]的语义搜索算法进行对比分析，用平均绝对误差(mean absolute error，MAE)对实验结果进行评价，其计算公式为

$$\mathrm{MAE}=\frac{\sum\limits_{i=1}^{N}|p_i-q_i|}{N},\tag{15}$$

其中：N 表示测试样本集的规模；p_i 表示算法的评分值；q_i 表示实际评分值.

3.2 结果与分析

1) 随机选取不同规模的训练样本和测试样本，采用本文语义搜索算法与对比语义搜索算法对数据集进行求解和分析，对比结果如图2所示. 由图2可见，训练样本的数量越多，则 MAE 的值越小，表明训练样本越多，学习结果越好，但在相同训练样本的数量条件下，本文语义搜索算法的 MAE 值最低，表明本文语义搜索算法有效提高了用户需要推荐精度，较好地解决了当前语义搜索算法存在的缺陷.

2）为了分析本文语义搜索算法的通用性，采用 Book-Crossing 数据集实现预测实验，本文语义搜索算法与对比语义搜索算法的实验结果如图 3 所示. 由图 3 可见，相对文献[19]和文献[20]的语义搜索算法，本文语义搜索算法的 MAE 大幅度提升，有效提高了用户个性化推荐精度，获得了更理想的用户个性化推荐结果.

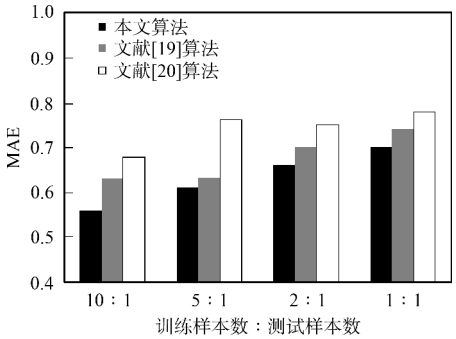

图 2　MovieLens 数据集的推荐结果对比

Fig.2　Comparisons of recommended results of MovieLens dataset

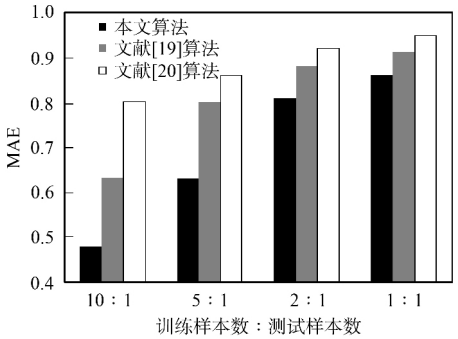

图 3　Book-Crossing 数据集的推荐结果对比

Fig.3　Comparisons of recommended results of Book-Crossing dataset

综上所述，针对传统信息检索系统没有准确描述语义，导致检索结果与用户实际需要结果相差较大的问题，为了提高用户需要检索的准确性，本文设计了一种基于抽取规则和本体映射的语义搜索算法. 首先根据抽取规则提取网页信息中的元素和属性，根据元素和属性对网页信息进行重新组织；然后建立计算机和用户语义之间的本体映射，使计算机能真正理解用户的意图，实现用户需要的智能化搜索；最后将语义搜索算法嵌入到用户个性化推荐系统. 实验结果表明，相比于传统算法，本文算法的语义搜索速度更快，减少了信息映射冗余度，取得较好的本体映射结果，能找到用户真正需要的信息，提高了用户个性化推荐系统的查询精度.

参　考　文　献

[1] Shvaiko P, Euzenat J. Ontology Matching: State of the Art and Future Challenges [J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(1): 158-176.

[2] Kumar S K, Harding J A. Ontology Mapping Using Description Logic and Bridging Axioms [J]. Computers in Industry, 2013, 64(1): 19-28.

[3] HUANG Shengjun, JIN Rong, ZHOU Zhihua. Active Learning by Querying Informative and Representative Examples [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(10): 1936-1949.

[4] 吴建. TRIZ 理论在搜索引擎创新设计中的应用研究 [J]. 重庆邮电大学学报(自然科学版), 2012, 24(6): 735-739. (WU Jian. Application of TRIZ in Search Engine's Innovative Design [J]. Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition), 2012, 24(6): 735-739.)

[5] 陈治昂，张毅，李大学. 基于 Web 智能的网络广告监测器研究与设计 [J]. 重庆邮电大学学报(自然科学版), 2009, 21(1): 115-118. (CHEN Zhi'ang, ZHANG Yi, LI Daxue. Research and Design of Web Adsmonitor Based on Web Intelligence [J]. Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition), 2009, 21(1): 115-118.)

[6] 刘显敏，李建中. 基于键规则的 XML 实体抽取方法 [J]. 计算机研究与发展, 2014, 51(1): 64-75. (LIU Xianmin, LI Jianzhong. Key-Based Method for Extracting Entities from XML Data [J]. Journal of Computer Research and Development, 2014, 51(1): 64-75.)

[7] 丁国栋，白硕，王斌. 一种基于局部共现的查询扩展研究 [J]. 中文信息学报, 2013, 20(3): 84-91. (DING Guodong, BAI Shuo, WANG Bin. Local Co-occurrence Based Query Expansion for Information Retrieval [J]. Journal of Chinese Information Processing, 2013, 20(3): 84-91.)

[8] 王璐，王国春，桂金花，等. 本体语义检索系统 [J]. 长春工业大学学报(自然科学版), 2013, 34(6): 726-730. (WANG Lu, WANG Guochun, GUI Jinhua, et al. A Semantic Retrieval System Based on Ontology [J]. Journal

of Changchun University of Technology (Natural Science Edition), 2013, 34(6): 726-730.)

[9] 李颖, 李志蜀, 邓欢. 基于 Lucene 的中文分词方法设计与实现 [J]. 四川大学学报(自然科学版), 2008, 45(5): 1095-1099. (LI Ying, LI Zhishu, DENG Huan. Design and Implementation of Chinese Words Segementation Based on Lucene [J]. Journal of Sichuan University (Natural Science Edition), 2008, 45(5): 1095-1099.)

[10] 张祥, 李星, 温韵清, 等. 语义网虚拟本体构建 [J]. 东南大学学报(自然科学版), 2015, 45(4): 652-656. (ZHANG Xiang, LI Xing, WEN Yunqing, et al. Virtual Semantic Web Ontology [J]. Journal of Southeast University (Natural Science), 2015, 45(4): 652-656.)

[11] 杨政国, 马建红. 基于领域本体科学效应知识语义检索的研究 [J]. 计算机系统应用, 2014, 23(2): 209-213. (YANG Zhengguo, MA Jianhong. Based on Scientific Effect Knowledge Domain Ontology Semantic Retrieval [J]. Computer Systems & Applications, 2014, 23(2): 209-213.)

[12] 张凌宇, 姜廷慈, 陈淑鑫. 一种基于参考本体的多本体映射方法 [J]. 四川大学学报(工程科学版), 2016, 48(5): 114-123. (ZHANG Lingyu, JIANG Tingci, CHEN Shuxin. A Method of Multiple Ontology Mapping Based on Reference Ontology [J]. Journal of Sichuan University (Engineering Science Edition), 2016, 48(5): 114-123.)

[13] 李凯, 李万龙, 郑山红, 等. 改进的多策略本体映射方法 [J]. 吉林大学学报(信息科学版), 2016, 34(4): 536-542. (LI Kai, LI Wanlong, ZHENG Shanhong, et al. Improved Multi-strategy Ontology Mapping Method [J]. Journal of Jilin University (Information Science Edition), 2016, 34(4): 536-542.)

[14] 史致远, Volker Gruhn, 朱明放. 微学习环境下基于语义的 MASHUP 架构优化 [J]. 江苏大学学报(自然科学版), 2010, 31(3): 339-342. (SHI Zhiyuan, Volker Gruhn, ZHU Mingfang. Optimization of Semantic MASHUP Architecture Applied in Microlearning Enivironment [J]. Journal of Jiangsu University (Natural Science Edition), 2010, 31(3): 339-342.)

[15] 孙煜飞, 马良荔, 吕闽晖, 等. 基于改进协同训练的本体映射方法 [J]. 系统工程与电子技术, 2017, 39(2): 459-464. (SUN Yufei, MA Liangli, LÜ Minhui, et al. Improved Co-training Based Ontology Matching Method [J]. Systems Engineering and Electronics, 2017, 39(2): 459-464.)

[16] 李华昱, 张培颖, 肖晗. 基于抽取规则和本体映射的领域 XML 语义集成 [J]. 河北科技大学学报, 2016, 37(4): 416-422. (LI Huayu, ZHANG Peiying, XIAO Han. Domain XML Semantic Integration Based on Extraction Rules and Ontology Mapping [J]. Journal of Hebei University of Science and Technology, 2016, 37(4): 416-422.)

[17] 徐德智, 易晓媛, 汤哲. 基 AHP-熵权决策的本体映射优化算法 [J]. 微电子学与计算机, 2017, 34(11): 48-52. (XU Dezhi, YI Xiaoyuan, TANG Zhe. An Ontology Mapping Optimization Algorithm Based on AHP and Entropy Weight Decision [J]. Microelectronics & Computer, 2017, 34(11): 48-52.)

[18] 韩彤, 宋文爱. 可信的第三方模糊本体映射框架及其实现 [J]. 计算机工程与设计, 2017, 38(3): 633-639. (HAN Tong, SONG Wen'ai. Trusted Third Party Fuzzy Ontology Mapping Framework and Its Implementation [J]. Computer Engineering and Design, 2017, 38(3): 633-639.)

[19] 王旭阳, 尉醒醒. 基于本体的语义检索方法 [J]. 计算机工程与设计, 2016, 37(9): 2538-2542. (WANG Xuyang, WEI Xingxing. Semantic Retrieval Method Based on Ontology [J]. Computer Engineering and Design, 2016, 37(9): 2538-2542.)

[20] 张建梁, 肖开东, 顾剑峰, 等. 基于 P2P 的结构化半分布式语义搜索算法 [J]. 计算机应用与软件, 2009, 26(4): 188-193. (ZHANG Jianliang, XIAO Kaidong, GU Jianfeng, et al. Semantic Search Algorithm Based on Structured Half-distributed P2P Network [J]. Computer Applications and Software, 2009, 26(4): 188-193.)

(责任编辑: 韩 啸)