

# 基于差异化融合的语义信息检索模型仿真

王 哲<sup>1,2</sup>, 徐燕文<sup>2</sup>

(1 东华大学 信息科学与技术学院, 上海 201620;

2 河南中医学院 信息技术学院, 河南 郑州 450046)

**摘 要:** 提出一种基于差异化融合的语义信息检索算法, 并实现模型设计和仿真. 构建语义本体模型, 提取关联维差异化特征, 进行特征融合, 采用自顶向下和分治的方法构建差异化语义信息特征的融合模型, 求得融合后的差异化语义信息特征信息增益. 在语义信息检索本体模型特征空间中, 计算语义信息特征, 将不同文本分量元素进行初始化倾向性分类处理, 基于集中度、分散度和频度构建特征空间, 进行信息抽取, 对分层差异化特征进行融合处理, 实现语义信息检索算法改进.

**关键词:** 语义; 信息检索; 数据库; 特征提取

**中图分类号:** TP391

**文献标识码:** A

**文章编号:** 1000-7180(2015)01-0146-04

## Simulation Model of Semantic Information Retrieval Based on Differentiation Fusion

WANG Zhe<sup>1,2</sup>, XU Yan-wen<sup>2</sup>

(1 College of Information Science and Technology, Donghua University, Shanghai 201620 China;

2 College of Information Technology, Henan University of Traditional  
Chinese Medicine, Zhengzhou 450046 China)

**Abstract:** A semantic information retrieval algorithm is proposed based on differentiation fusion, the model design and simulation are realized. The semantic ontology model is constructed, the correlation dimension difference characteristic is extracted, fusion model difference semantic information features take top-down method, the difference of semantic information is obtained, semantic information retrieval ontology model is obtained in the feature space, the semantic feature of information is computed, different text component elements of classification processing is fused, the stratified differentiation characteristics of fusion, achieve improved semantic information retrieval algorithm.

**Key words:** semantic; information retrieval; database; feature extraction

### 1 引言

传统的语义信息检索模型设计中, 采用算法主要有 K-Means 法、FCM 法、遗传算法、最大似然法、Fisher 判别法、神经网络等<sup>[1]</sup>, 其中, 文献[2]中在语义异构环境下, 设计本体离散度增强算法, 通过构建跨越系统间的知识共享和集成本体, 实现知识协同, 提高语义离散度, 进而实现语义信息检索, 但算法计算复杂, 模型构建困难; 文献[3]中提出一种基

于关联维特征提取的网络异常数据的语义分析算法, 取得一定的特征识别性能, 提高了对异常数据的拦截能力, 但该算法无法有效识别差异化网络数据的语义特征, 语义信息检索效果不好; 文献[4]中采用多策略相似度计算和贝叶斯网络模型为本体集成系统给出本体映射方法 MSBN, 通过语义信息特征相似度提取, 实现语义信息检索, 算法采用反复迭代和概率推理的机制来建立本体之间的语义映射关系, 导致重复迭代和约束信息滞后, 检索精度

不高<sup>[5-7]</sup>.

针对上述算法出现的问题,特别是在文献[3]中提出的算法研究基础上,提出一种基于差异化融合的语义信息检索算法,首先构建语义本体模型,在模型中研究语义信息检索的自动识别和提取问题,对语义信息进行信号模型构建,提取关联维差异化特征,进行特征融合,实现语义检索算法改进设计,仿真实验进行性能验证,展示了本文算法的可行性和优越性.

## 2 语义信息检索本体模型构建与特征分析

### 2.1 语义信息检索本体模型构建

首先给出语义本体模型的形式化定义以及语义映射关系的定义和分类,语义本体  $O = \langle C, I, H^C, R, A \rangle$  与  $O = \langle C, I, H^C, R, A \rangle$  之间的语义映射关系可以通过一个映射函数  $M$  表示,在语义信息检索本体模型特征空间中<sup>[8-10]</sup>,计算语义信息特征  $t$  在  $c_i$  类内的频度:

$$TF(t, c_i) = \frac{P(t | c_i)}{\sum_{j \rightarrow n} P(t_j | c_i)} \quad (1)$$

式中,  $t$  为当前词汇的语义关联度,  $c_i$  为第  $i$  个类,  $t_j$  为词库中第  $j$  个词汇,  $n$  为词库中海量兴趣点词汇总数.

假设  $O = \langle N, E, F \rangle$  表示一个以三元组形式定义的本体模型,其中  $O$  中所包含的元素节点集合  $N$  (包括概念、实例和属性),得到多维矩阵向量的文本特征矢量为

$$Q_N R_{N-1} = J_{x_{N-1}}^{(1)} Q_{N-1} \quad (2)$$

式中,  $Q_{N-1}$  表示同一类文本中出现的词频数,  $R_0$  表示词汇表,  $Q_N$  表示局部特征映射矩阵,以此为基础,基于集中度、分散度和频度构建特征空间,进行信息抽取.将不同文本分量元素进行初始化倾向性分类处理,由指数矩阵确定元素属性类别概率,在所所有的词条中,有一个与知识类同名的主词条,得到语义信息个体与上层  $GT$  的纵向聚类关系:

$$JL_1 = \sum_{i=1}^n \sum_{j=1}^m \frac{\sqrt{(A_{iseq} - B_{jseq})^2 + 1}}{|A_{iq} - B_{jq}| + 1} \quad (3)$$

式中,  $n$  为本层节点个数,  $m$  为上层节点个数,  $A_{iseq}$  表示本层  $GT$  的位置序号,  $B_{jseq}$  表示上层  $GT$  的位置序号,  $A_{iq}$  为本层  $GT$  的聚类系数,由此建立本体之间语义映射过程,如图 1 所示.

在图 1 所示的语义本体中,概念既是实例的分类标准又是知识存储的基本单元,概念之间的分类

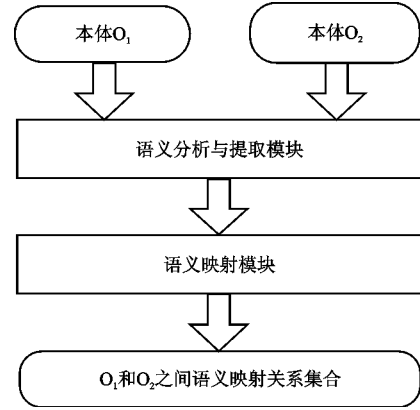


图 1 本体之间语义映射

关系集合表示行为关系和概念与实例之间的隶属关系,数据库数据集  $X = \{x_1, x_2, \dots, x_n\}$ ,  $n$  是数据集  $X$  的数目,  $X$  中的每个元素都是一个  $P$  维矢量,  $X$  含有  $c$  个类别,第  $i$  个类的中心为  $v_i = \{v_{i1}, v_{i2}, \dots, v_{ip}\}$ ,计算有向图  $G_1, G_2$  语义关系高斯边缘化程度  $S_r$ ,最后综合  $S_c$  和  $S_r$  组成统一的语义高斯边缘化的矩形窗函数  $h(t)$ ,窗函数的宽度为  $T = (2d + 1)T_s, F_s = 1/T_s$ .由此实现了语义信息检索本体模型构建,为进行语义信息检索模型设计奠定基础.

### 2.2 语义信息的差异化特征提取及分析

在上述语义本体模型中,对语义信息进行信号模型构建,提取关联维差异化特征,进行特征融合,在文本数据择取过程中,将不同文本分量元素进行初始化倾向性分类处理,由指数矩阵确定元素属性类别概率,从而确定多个导向性聚类中心,得到数据库集合的特征概率函数密度表达式为

$$p(y | \alpha, \theta) = \sum_{k=1}^K \alpha_k p_k(y | \mu_k, \sum_k) \quad (4)$$

得到语义本体结构中特征空间互信息区域文档词频向量模型:

$$\Delta d = \frac{1}{2} \times t \times u \times \sum_{i=1}^t \sum_{j=1}^u \| \bar{d}_i - M \hat{d}_i^j \|^2 \quad (5)$$

构建语义概念树,研究对象和属性集合之间的二元关系.语义主题数的三元组的形式  $K = (O, A, R)$ ,其中  $O$  是对象的集合,  $A$  是属性的集合,  $R$  是  $O$  和  $A$  之间的一个二元关系,设类别总数为  $m$ ,当前设置反例数量的类别为  $c_x$ ,每个词汇的位置顺序能反映该词汇对于文本所属类别的贡献程度,进行语义信息的差异化特征提取算法设计,语义差异化特征向量空间模型的相似度文本特征向量:

$$x_t = MSF(t) \sum_{k \in K_t} \beta_{t,k} \quad (6)$$

任意两个词语  $X, Y$ , 定义三种相似度系数公式如下:

$$WebJaccard(X, Y) = \frac{P(X \cap Y)}{P(X) + P(Y) - P(X \cap Y)} \quad (7)$$

$$WebOverlap(X, Y) = \frac{P(X \cap Y)}{\min(P(X), P(Y))} \quad (8)$$

$$WebDice(X, Y) = \frac{2P(X \cap Y)}{P(X) + P(Y)} \quad (9)$$

式中,  $P(X), P(Y)$  是分别搜索  $X, Y$  得到的综合权重的词语相似度,  $P(X \cap Y)$  是联合搜索  $X, Y$  所得到的词语的特征权值. 修正每个向量  $v_i$ , 选择包含前  $m$  个最大特征值的词语作为  $v_i$  的表示, 即  $v_i = ((w_1, t_1), (w_2, t_2), \dots, (w_m, t_m))$ ; 计算出语义本体概念树集合的中心向量  $C(Y)$ , 则  $X, Y$  的相似度计算公式为

$$Sim(X, Y) = Cos(X, Y) = \frac{C(X) \cdot C(Y)}{|C(X)| \cdot |C(Y)|} \quad (10)$$

各语义信息的差异化特征  $w_1, w_2$  之间的语义距离  $dis(w_1, w_2)$  定义为连接它们的最短路径上  $n$  条边的权值之和, 即语义信息的差异化特征空间距离为

$$dis(w_1, w_2) = \sum_{i=1}^n weight(e_i) \quad (11)$$

语义差异化特征挖掘的离散度概率函数密度表达式为

$$p(y | \alpha, \theta) = \sum_{k=1}^K \alpha_k p_k(y | \mu_k, \sum_k) \quad (12)$$

式中,  $\delta x_i = x_j - x_i, x_j$  是  $x_i$  近邻的一组指数分离向量, 综上, 实现了语义信息的差异化特征提取, 为下一步进行差异化特征融合和语义信息检索模型构建提供信息源基础.

### 3 差异化特征融合及语义信息检索算法改进实现

语义信息检索模型的构成包扩了 7 层模型, 分别可以概括为: UNICODE 和 URI 层、XML + 命名空间层、RDF + RDFS 层、本体层、逻辑层、证明层和信任层. 各个层次产生了差异化特征, 在上述特征提取的基础上, 本文针对传统的语义信息检索算法采用基于关联维特征提取的语义分析算法, 无法有效识别差异化网络数据的语义特征, 语义信息的差异化融合效果不好的问题, 提出一种基于差异化融合

的语义信息检索算法, 并实现模型设计和仿真.

采用有向图表示语义网, 其逻辑比较算法步骤为: 首先寻找两个有向图之间的交集. 假设  $G_c$  是两个语义网转换而成的有向图  $G_1$  和  $G_2$  的交集, 那么说明  $G_c$  中的语义结点同时存在于  $G_1$  和  $G_2$ , 并且  $G_c$  中的语义关系弧也同时存在于  $G_1$  和  $G_2$  中. 计算有向图  $G_1, G_2$  语义结点相似度  $S_c$ ; 其次计算有向图  $G_1, G_2$  语义关系相似度  $S_r$ , 综合  $S_c$  和  $S_r$  组成统一的语义高斯边缘化的矩形窗函数  $h(t)$ , 窗函数的宽度为  $T = (2d + 1)T_s, F_s = 1/T_s$ , 测试数据待测试窗口的特征向量表示为

$$E_{i,j} = \langle e_1, e_2, \dots, e_m \rangle \quad (13)$$

差异化语义信息特征的融合模型构建采用了自顶向下和分治的方法, 其构造方式如下, 令  $A = \{a_1, a_2, \dots, a_n\}$  为差异化语义信息特征训练集的属性集,  $B = \{b_1, b_2, \dots, b_m\}$  为差异化语义信息特征信息增益的类别集,  $a_i$  的语义属性值为  $\{c_1, c_2, \dots, c_k\}$ , 得到融合后的差异化语义信息特征信息增益表达式为:

$$Info(B) = - \sum_{i=1}^m p_i \times \log_2 p_i \quad (14)$$

$$Info_A(B) = \sum_{j=1}^v \frac{|B_j|}{|B|} \times Info(B_j) \quad (15)$$

$$Gain(A) = Info(B) - Info_A(B) \quad (16)$$

式中,  $B_j$  表示在训练集中含有  $a_x$  属性中的  $c_v$  值的元素集合. 考虑语义特征信号分解中的差异性, 进行差异化融合, 语义离散 Dopplerlet 变换:

$$DDT_{s(n)} = \sum_{n=0}^{N-1} s(n) d_{t_0, \Delta t, t, v, c}^*(n) \quad (17)$$

语义信息检索的相似度与概念深度之间存在正比的关系, 可以表示为

$$Sim(w_1, w_2) \propto \frac{dep(w_1) + dep(w_2)}{|dep(w_1) - dep(w_2)|} \quad (18)$$

式中, 符号  $\propto$  表示正比关系,  $dep(w_1), dep(w_2)$  为主题词  $w_1, w_2$  的概念深度, 在本体层次上进行数据挖掘, 通过算术平均值来计算结果集  $R(X)$  的中心向量  $C(X)$ , 计算式为

$$C(X) = \frac{1}{n} \sum_{i=1}^n \frac{v_i}{|v_i|} \quad (19)$$

对差异化融合特征进行窗函数调制, 经尺度伸缩, 克服中心频率变化对算法性能的影响, 设计了一个综合权重的词语相似度计算公式:

$$VariantSim(X, Y) = \alpha \cdot Sim(X, Y) + \beta \cdot WebJaccard(X, Y) + \gamma \cdot WebOverlap(X, Y) + \delta \cdot WebDice(X, Y) \quad (20)$$

式中,  $\alpha + \beta + \gamma + \delta = 1$ ,  $\alpha, \beta, \gamma, \delta$  是调节参数, 计算出词  $Y$  的结果集中心向量  $C(Y)$ , 输入语义主题树特征词  $X$  到语义信息检索的本体特征空间, 实现了本文提出基于差异化融合的语义信息检索。

#### 4 仿真分析及结论

为了测试本文算法在实现语义信息检索中的性能, 进行仿真实验。在语义信息检索中, 采用  $20\,000 \times 20$  表示数据行数为  $20 \times 1\,000 = 20\,000$ , 列数为 20, 用户参数  $L = 500$ , 用户参数  $L = 500$ , 运行时间单位为毫秒。语义本体特征的权值向量  $W = \{\omega_i, \omega_l, \omega_u\}^T$ , 根据现实情况确定各自在综合信任中的权重, 实验中, 分别取值为  $\{0.2, 0.2, 0.6\}$ ,  $\{0.3, 0.4, 0.3\}$  和  $\{0.4, 0.4, 0.2\}$ , 语义特征提取计算过程中, 全局迭代次数为 500 次, 变量的维数  $n$  为 30, 算法独立运行 30 次, 进行循环迭代, 得到语义特征不同尺度下的语义信息差异化特征提取结果如图 2 所示。

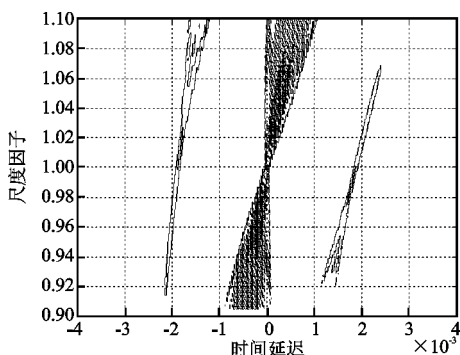


图 2 语义信息差异化特征提取结果

以图 2 结果为基础, 进行差异化特征融合和语义信息检索模型构建, 得到差异化特征融合结果如图 3 所示。

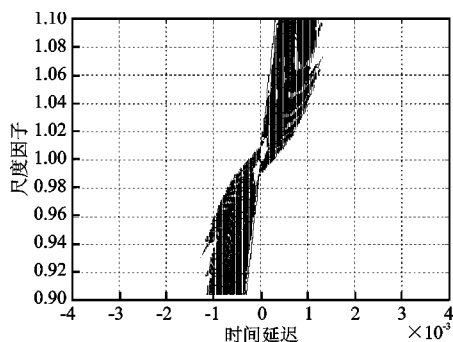


图 3 语义信息差异化融合结果

从图 2 和图 3 结果可见, 原始的关联维差异化特征具有非线性和随机散布性, 主题分布广, 更新频率大, 无法有效实现信息检索。采用本文算法, 进行差异

化特征融合, 进行窗函数调制, 经尺度伸缩, 克服中心频率变化对算法性能的影响, 适应在主体分布广和语义更新频率大的情况下的语义特征提取。在智能信息检索中, 采用查全率和查准率进行检索质量指标的测评。查准率是检索到的关联信息数目与检索到的所有信息数目之间的比值。查全率表示检索到的相关信息数与数据库中检索需求相关的所有信息数之间的比值。得到不同算法下的语义信息检索查全率 (Precision) 和查准率 (Recall) 结果见表 1。

表 1 查全率和查准率测试指标表

|   | #102 | #103 | #104 | #201 | #202 | #203 | #204 | #205 | #206 |
|---|------|------|------|------|------|------|------|------|------|
| P | 1.00 | 1.00 | 1.00 | 0.91 | 0.87 | 0.99 | 0.98 | 0.95 | 0.97 |
| R | 1.00 | 1.00 | 1.00 | 0.90 | 0.86 | 0.97 | 0.96 | 0.94 | 0.96 |
| F | 1.00 | 1.00 | 1.00 | 0.91 | 0.86 | 0.98 | 0.96 | 0.94 | 0.96 |

分析上述结果可见, 本文方法得到的语义信息文本数据召回率  $R$  为 100%, 统计所有结果发现, 89.5% 以上的同义词的相似度值大于 0.1, 随着判定阈值的增大, 查准率不断下降, 当阈值为 0.9 时, 查准率为 98.7%, 采用差异化融合设计, 有效提高了语义信息检索的查全率和查准率, 采用本文算法, 在语义本体模型信息不完整以及语义信息表示不一致的情况下, 较好地完成了语义映射任务。

仿真实验表明, 采用本文算法进行语义信息检索, 通过差异化特征融合, 适应在主体分布广和语义更新频率大的情况下的语义特征提取, 有效提高了语义信息检索的查全率和查准率, 展示了较好的应用价值。

#### 参考文献:

- [1] 张玲. POI 的分类标准研究[J]. 测绘通报, 2012(10): 82-84.
- [2] 朱剑锋, 陈珉. 语义异构环境下基于本体的供应链知识协同研究[J]. 武汉大学学报: 信息科学版, 2014, 39(1): 123-126.
- [3] 欧石燕, 胡珊, 张帅. 本体与关联数据驱动的图书馆信息资源语义整合方法及其测评[J]. 图书情报工作, 2013, 58(2): 5-13.
- [4] 邵忻, 徐倩漪. 网络伪装不良信息检测方法的研究与仿真[J]. 计算机仿真, 2012, 29(2): 135-138.
- [5] Xu Zhixiang, Chen Minmin, Q Weinberger K Q. An alternative text representation to TF-IDF and Bag-of-Words[C]//The 21nd ACM International Conference on Information and Knowledge Management (CIKM 2012), Maui, HI, USA: ACM, 2012: 58-63.

(下转第 155 页)

- 精度 TDC 芯片设计[J]. 微电子学与计算机, 2010, 27(5): 57-65.
- [5] 安琪. 粒子物理实验中的精密时间间隔测量[J]. 核技术, 2006, 29(6): 454-458.
- [6] 黄钰. 定位系统中时间间隔测量电路的设计与实现[D]. 北京:清华大学, 2013.
- [7] Jozef Kalisz, Ryszard Szplet, Jerzy pasierbinski, et al. Field-programmable-gate-array-based time-to-digital converter with 200-ps resolution[J]. IEEE Transactions On Instrumentation and Measurement, 1997, 46(1): 71-75.
- [8] Szplet R, Kalisz J. Interpolating time counter with 100-ps resolution on a single FPGA Device[J]. IEEE Trans Instrum and Meas, 2000, 49(4): 879-883.
- [9] Song Jian, An Qi, Liu Shubin, A high-resolution time-to-digital converter implemented in field-programmable-gate-arrays [J]. IEEE Transactions on nuclear-science, 2006, 53(1): 236-241.
- [10] Jinyuan Wu, Zonghan shi. The 10-ps wave union TDC: Improving FPGA TDC resolution beyond Its cell delay [C]// Nuclear Science Symposium Conference Record. Batavia: IEEE, 2008: 3440-3446.
- [11] Bayer E, Zipf P, Traxler M. A multichannel high-resolution ( $< 5$  ps RMS between two channels) time-to-digital converter (TDC) implemented in a field programmable gateArray (FPGA) [C]// IEEE Nuclear Science Symposium Conference Record. [s. l.]: IEEE, 2011: 876-879.
- [12] Eugen Bayer, Traxler M. A high-resolution ( $< 10$  ps RMS) 48-channel time-to-digital converter (TDC) implemented in a field programmable gate array (FPGA) [J]. IEEE Transactions on nuclear science, 2011, 58(4): 1547-1552.

#### 作者简介:

张 锋 男, (1985-), 硕士. 研究方向为集成电路与系统设计研究. E-mail: xdfjn@163.com

胡和益 男, (1988-), 硕士. 研究方向为集成电路与系统设计研究.

李永明 男, (1945-), 教授. 研究方向为集成电路与系统设计研究.

#### (上接第 149 页)

- [6] 王昌晶, 薛锦云, 左正康. SRLtoRadl 生成系统及其范畴论语义[J]. 电子学报, 2014, 42(1): 137-142.
- [7] 邹亮, 徐德智, 郭维. 基于参考点的大规模本体扩散映射算法 [J]. 小型微型计算机系统, 2013, 34(7): 1507-1513.
- [8] 何永强, 谷春英. 基于子任务区域分片下的分布式空间查询处理与并行调度方法[J]. 科技通报, 2014, 30(1): 110-116.
- [9] Alfaro V M, Vilanovab R. Robust tuning of 2DoF five-parameter PID controllers for inverse response controlled processes[J]. Journal of Process Control, 2013, 23(4): 453-462.
- [10] Hu W M, Hu M, Maybank S. Adaboost based algorithm for network intrusion detection [J]. IEEE Transactions on Systems, Man and Cybernetic, Part B: Cybernetics, 2008, 38(2): 577-583.

#### 作者简介:

王 哲 女, (1981-), 博士研究生, 讲师. 研究方向为语义信息检索、数据库与数据挖掘. E-mail: 13938583437 @ 126.com.

徐燕文 女, (1981-), 硕士, 讲师. 研究方向为数据库与数据挖掘.