

基于差异化信息融合的语义信息检索模型

王 哲^{1,2}

(1 河南中医学院 信息技术学院, 河南 郑州 450046; 2 东华大学 信息科学与技术学院, 上海 201620)

摘 要: 研究语义信息检索方法, 提高检索的效率. 差异化信息由于特征差异过大, 在检索过程中存在较大排异现象, 传统的语义信息检索模型针对大差异信息检索过程中, 以多轮次检测为主, 效率很低. 为此, 提出一种基于决策树算法的语义信息检索方法. 根据多层次解析融合相关理论, 计算窗口函数, 并且根据窗口函数进行不同层次数据的融合, 得到差异信息融合结果. 根据上述结果, 建立决策树, 实现语义信息的检索. 实验结果表明, 利用改进算法进行语义信息检索, 能够提高检索的效率.

关键词: 信息融合; 信息检索; 多层次解析; 决策树

中图分类号: TP311

文献标识码: A

文章编号: 1000-7180(2014)08-0159-04

The Semantic Information Retrieval Model Based on Differential Information Fusion

WANG Zhe^{1,2}

(1 College of Information Technology, Henan University of Traditional Chinese Medicine, Zhengzhou 450046 China;

2 College of Information Science and Technology, Donghua University, Shanghai 201620 China

Abstract: The research of semantic information retrieval method, improve the efficiency of retrieval. Differentiated information due to the characteristics of the difference is too big, the larger rejection in the retrieval process, the traditional semantic information retrieval model in view of the large differences in the process of information retrieval, is given priority to with testing rounds, efficiency is very low. For this, put forward a kind of semantic information retrieval method based on decision tree algorithm. Calculate according to the related theory, multi-level analytic fusion window function, and based on window function for different levels of data fusion, the differences in information fusion results are obtained. Based on the above results, the establishment of a decision tree, realization of semantic information retrieval. Experimental results show that the improved algorithm of semantic information retrieval, and can improve the efficiency of retrieval.

Key words: information fusion; information retrieval; multi-level analytic; decision tree

1 引言

如何在海量的计算机信息中, 快速检索目标语义信息, 已经成为计算机领域需要研究的核心问题, 发挥着越来越重要的作用^[1-2]. 检索效率, 也已经成为衡量语义信息检索方法的一个重要指标. 因此, 语义信息检索方法, 已经成为计算机数据处理领域需要研究的热点问题, 受到了极为广泛的关注. 现阶段, 主要的语义信息检索方法包括基于均值聚类算

法的语义信息检索方法、基于马尔科夫算法的语义信息检索方法和基于聚类式云计算算法的语义信息检索方法. 其中, 基于聚类式云计算算法的语义检索方法是当前最常用的语义信息检索方法^[3-4]. 由于聚类式云计算算法能够进行复杂数据挖掘, 因此成为很多专家学者研究的热门课题, 发展前景十分广阔.

使用计算机进行语义信息检索的过程中, 语义信息中存在大量复杂无序的信息, 这些信息存在较

大差异^[5]. 如果使用传统的聚类式云计算算法进行语义信息挖掘处理, 将会导致算法出现收敛速度慢、计算复杂度高、计算数据量大, 容易陷入无穷对比等弊端, 严重降低了计算机数据处理的效率. 为了避免传统方法的缺陷, 提出一种基于决策树算法的语义信息检索方法^[6]. 利用多层次解析融合理论能够得到差异信息的窗口函数, 利用得到的窗口函数进行不同层次差异信息数据的融合, 能够得到差异信息融合结果, 利用获得的结构能够建立语义信息检索决策树, 从而实现了语义信息的检索. 实验结果表明, 利用改进算法进行语义信息检索, 能够提高语义检索的效率, 相对传统算法具有较强的优越性.

2 传统语义信息检索原理

2.1 计算动态语义信息关联性

根据关联规则相关理论, 能够有效地对海量的语义信息动态数据的关联性进行计算, 其详细过程如下所述:

设置语义信息动态数据构成的集合能够用 Z 进行描述, 全部语义信息动态数据的数量能够用 q 进行描述, 上述语义信息动态数据属性的数量能够用 p 进行描述^[7-9]. 全部模糊语义信息动态数据属性构成的集合能够用 $D = \{d_1, d_2, \dots, d_r\}$ 进行描述. 利用下述公式能够计算出不同属性语义信息动态数据的支持度:

$$H(D) = \sum_{l=1}^q \prod \omega(d_k, z_l) \quad (1)$$

利用下述公式能够计算语义信息动态数据的关联度:

$$y_{kl} = \frac{\exp(-\frac{\beta}{\kappa})}{\sum_{k=1}^p \exp(-\frac{\beta}{\kappa})} \quad (2)$$

2.2 建立语义信息检索的马尔科夫模型

在马尔科夫模型中, 语义信息数据的层次能够用 $N = \langle T, B_T, S_T^b, Q_T^b, W \rangle$ 进行描述, 其中, T 是语义信息检索系统的语义状态参数数据集, B_T 是语义信息检索系统中的语义状态参数数据集到语义信息数据集的映射, Q_T^b 是语义信息状态参数到 T 的映射, S_T^b 是检索决策的概率, W 是检索时间参数.

利用下述公式能够计算出检索决策的概率:

$$S_T^b = F\{s_{u+1} + \eta s_{u+2} + \eta^2 s_{u+3} + \dots + \eta^{r-1} s_{u+p} \mid t_u = t, t_{u+p} = t^s\} \quad (3)$$

对差异语义信息相关的数据进行非结构化处理, 能够得到不同检索决策的状态参数:

$$W_i(t) = F_i(\sum_{i=1}^p \eta(i) S_i \mid t_0 = t) \quad (4)$$

式中, $\eta(i) = \eta^{y_1+y_2+\dots+y_p}$, 用来描述检索到第 i 种语义信息的决策执行力.

对上述公式进行计算能够得到最优语义信息决策状态参数:

$$W^+(t) = \sum_b \tilde{\omega}(t, b) \sum_t q_t^b (S_t^b + \eta W(t^b)) \quad (5)$$

结合上述方法, 利用下述公式, 能够建立语义信息检索的马尔科夫模型:

$$R^+(t, b) = \sum_t Q_t^b (S_t^b + \eta \max(\sum_t q_t^b (S_t^b + \eta W(t^b)))) \quad (6)$$

根据上述阐述的方法可知, 利用语义信息数据重构方法, 能够对差异语义信息相关的参数进行初始化处理, 对差异语义信息检索决策提供了准确的数据. 建立的语义信息检索的马尔科夫模型, 能够根据模型的输出结果对差异语义信息进行检索.

3 基于决策树算法的语义信息检索方法

利用传统语义信息检索方法进行复杂语义数据挖掘, 难以避免由于海量数据过于复杂导致算法的收敛速度较慢的弊端, 从而降低了计算机数据处理的效率. 为了传统算法的弊端, 提出一种基于决策树算法的语义信息检索方法.

3.1 实现语义信息检索过程中差异语义信息融合

根据数据特征多层次解析融合的相关算法, 能够对差异语义信息进行融合, 实现语义信息检索.

在语义信息检索的过程中包含大量的差异语义信息, 上述语义信息的能量存在较大的差异性. 针对不同的语义信息数据机型加权能量计算, 能够得到下述结果:

$$\begin{aligned} PN(k, l) &= \left| \frac{\partial^2 K(k, l)}{\partial k^2} \right| + \left| \frac{\partial^2 K(k, l)}{\partial l^2} \right| = \\ &|2K(k, l) - K(k - step, l) - K(k + step, l)| + \\ &|2K(k, l) - K(k, l - step) - K(k, l + step)| \\ UON(k, l) &= \sum_k \sum_l NP(k, l) \end{aligned} \quad (7)$$

根据差异信息检索窗口函数, 能够对差异信息进行融合处理, 利用下述公式能够描述处理过程:

$$NP(k, l) = \sum_r \sum_s Y(r, s) [PN(k + r, l + s)]^2 \quad (8)$$

利用下述公式能够对差异语义信息数据特征进行多层结构, 计算结果如下所述:

$$\psi_c = \frac{PN^c(k, l)}{PN^c(k, l) + PN^d(k, l)}$$

$$\psi_D = \frac{PN^D(k, l)}{PN^C(k, l) + PN^D(k, l)} \quad (9)$$

设置语义信息检索系统中,子检测系统能够用 C 和 D 表示,则能够得到差异语义融合公式如下:

$$K_H(k, l) = \begin{cases} K_C(k, l) & NP_C(k, l) - NP_D(k, l) > v \\ K_C(k, l) \times \psi_C + K_D(k, l) \times \psi_D & |PN_C(k, l) - PN_D(k, l)| \leq v \\ K_D(k, l) & PN_C(k, l) - PN_D(k, l) < v \end{cases} \quad (10)$$

根据上述阐述的方法,能够提取语义信息检索过程中差异语义信息特征,并针对上述特征进行语义融合,实现了语义信息检索过程中的高效性。

3.2 建立语义信息检索决策树

设置全部动态语义信息数据构成的集合能够用 $Z = \{(z_k, a_k) \mid k = 1, 2, \dots, total\}$ 进行描述,其中 $z_k = (z_{k1}, z_{k2}, \dots, z_{kf})$ 表示上述集合中的全部动态语义信息数据.上述动态语义信息数据的属性构成的集合能够用进行描述 $\{C_1, C_2, \dots, C_f\}$. 利用下述公式能够计算动态信息数据的期望熵:

$$K(q_1, q_2, \dots, q_p) = - \sum_{l=1}^p \frac{q_l}{total} \log_2 \left(\frac{q_l}{total} \right) \quad (11)$$

设置全部语义信息的动态数据属性能够用 $C_h (h = 1, 2, \dots, f)$ 进行描述,其中包含 s 个不同的属性值,利用下述公式能够对上述语义信息的动态属性数据进行分界处理:

$$G(C_h) = \sum_{u=1}^s \frac{q_{1u} + \dots + q_{pu}}{total} K(q_{1u}, q_{2u}, \dots, q_{pu}) \quad (12)$$

根据下述公式能够计算语义信息动态数据属性的信息增益与信息增益比:

$$\begin{aligned} E(C_h) &= K(q_1, q_2, \dots, q_h) - G(C_h) \\ u(C_h) &= - \sum_{u=1}^s \frac{q_u}{total} \times nd \left(\frac{q_u}{total} \right) \\ I_{-t}(C_h) &= \frac{I(C_h)}{u(C_h)} \end{aligned} \quad (13)$$

利用下述公式能够建立语义信息动态数据优化参数决策树:

$$C_k = MIN + \frac{MAX - MIN}{Q} \times k \quad (14)$$

式中, $k = 1, 2, \dots, Q$.

根据上述阐述的方法,获得的语义信息动态数据优化参数决策树能够用图 1 进行描述.

将全部语义信息增益比为极大值的动态语义信息数据作为决策树的节点,能够实现语义信息检索

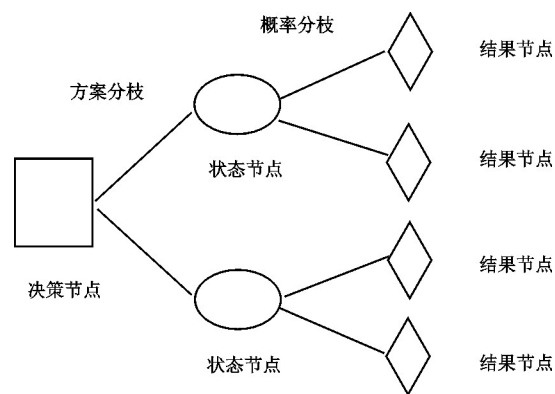


图 1 决策树模型

决策树的构建。

4 实验结果及分析

完了验证改进算法的有效性,需要进行一次实验.利用 Visual C++6.0 构建实验环境.云数据挖掘过程需要的时间是衡量不同语义信息检索方法的一个重要指标.

设置数据库中的语义信息样本数量是 1 000 个,全部云数据语义信息种类是 15 种.

将上述语义信息样本数据以及属性进行整理分析,能够得到表 1.

表 1 语义信息样本数据表

序号	属性	数据量
1	酒店	19
2	航空	86
3	港口	78
4	世界	63
5	工业	91
6	经济	25
7	政治	38
8	改革	49
9	文化	66
10	开放	81
11	政府	77
12	安保	54
13	农业	39
14	交通	83
15	国家	76

为了验证改进算法的优越性,需要进行一次实验.分别利用传统算法和改进算法对上述语义信息数据进行数据挖掘处理,获得实验结果能够用图 2 进行描述.

根据图 2 实验结果可知,在语义信息样本复杂度较低时,不同算法进行数据挖掘的时间基本一致.

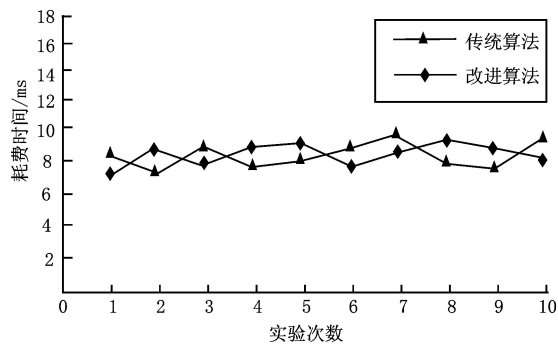


图2 样本复杂度较低时检索结果

将全部语义信息样本数据作为实验对象,设置需要在这些语义信息中获取与改革相关的数据.分别利用改进算法与传统算法进行语义信息挖掘处理,则信息检索耗费的时间能用图3进行描述.

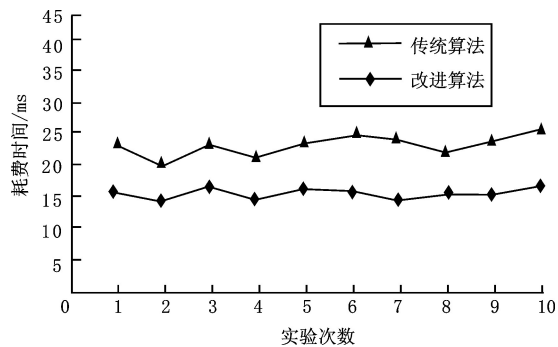


图3 样本复杂度较高时检索结果

根据图3实验结果可知,在语义信息样本复杂度较高时,利用改进算法进行数据挖掘耗费的时间在15 ms左右,利用传统算法进行数据挖掘耗费的时间在24 ms左右,充分体现出改进算法在样本复杂度较高时的优越性.

对上述实验过程中的数据进行分析整理,能够得到实验结果能够用表2和表3进行描述.

表2 语义信息样本复杂度较低时实验结果

实验次数	传统算法耗费时间/ms	改进算法耗费时间/ms
1	8	7
2	7	8
3	8	7
4	7	8
5	7	8
6	8	6
7	8	7
8	8	8
9	8	8
10	7	7

表3 语义信息样本复杂度较高时实验结果

实验次数	传统算法耗费时间/ms	改进算法耗费时间/ms
1	24	15
2	20	14
3	25	15
4	21	14
5	24	15
6	25	15
7	24	14
8	22	15
9	23	15
10	25	15

根据上述实验结果可知,利用改进算法进行语义信息检索,能够避免传统算法由于数据量过于复杂导致的算法收敛速度变慢的缺陷,从而提高了计算机处理云数据的挖掘速度.

5 结束语

针对传统语义信息检索方法进行数据量过于庞大复杂的语义信息检索时容易出现收敛速度变慢的弊端,提出一种基于决策树算法的语义信息检索方法.利用信息数据多层次解析融合理论,计算窗差异信息的窗口函数,利用差异信息的窗口函数进行不同层次数据的融合,能够得到差异信息数据融合结果,利用获得的结构能够建立语义信息检索决策树,从而实现语义信息的检索.实验结果表明,利用改进算法进行语义信息检索,能够避免传统算法的弊端,提高了计算机的检索效率.

参考文献:

- [1] 徐章艳, 张师超. 挖掘关联规则中一种优化的 Apriori 算法[J]. 计算机工程, 2003, 29(19): 83-84.
- [2] 李德仁, 王树良, 史文中. 论空间数据挖掘和知识发现[J]. 武汉大学学报: 信息科学版, 2001, 26(6): 491-499.
- [3] 瞿晓静. 试论基于语义网的智能代理[J]. 情报杂志, 2004(01): 35-36.
- [4] 张海霞, 吴江. 基于语义网的知识管理系统框架设计[J]. 计算机技术与发展, 2006(4): 46-48.
- [5] Welly C, Guarino N. Supporting ontological analysis of taxonomic relationships[J]. Data and Knowledge Engineer, 2007, 39(1): 51-74.
- [6] 白同强, 刘磊. 语义网的研究与展望[J]. 吉林大学学报: 信息科学版, 2005(7): 121-125.
- [7] Tim Berners-Lee, James Hendler, Ora Lassila. The semantic web[J]. Scientific American, 2001(5): 29-37.

(下转第 166 页)

能变化比较大. 由于实验应用程序满足对 9% 的对数据库的用户请求, 我们假设 d 的值为 0.9. 略微修改 d 的值会得到成正比的结果.

如我们预计的那样, 队列 2 的等待时间小于队列 1. 队列 2 中的用户流量降低也降低了其平均等待时间. 因此, 等待时间随 d 减少. W 是 1 和 2 队列中的等待时间的总和. $W_{\max}$, $W_{\min}$ 将决定同一时刻客户/连接数量.

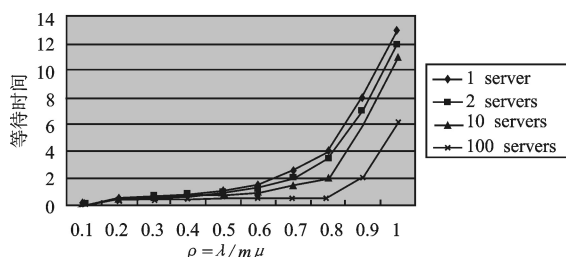


图 4 队列 1 上的等待时间

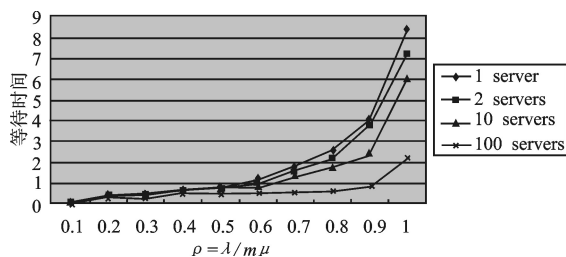


图 5 队列 2 上的等待时间

5 结束语

本文提出基于 QoS 保证机制的应用于网络教育的云模型. 在该模型中使用了两个 $M/M/m$ 队列系统相结合的方法. 第一个队列提供计算服务, 第二个队列提供访问数据库服务器的服务. 在平均等待时间方面, 该模型提供良好的 QoS, 当增加虚拟机的数量时, 第一与第二队列之间的等待时间趋于稳定. 从管理, 技术, 安全性和合法性的角度看, 该模

型可以提高网络教育服务的水平. 未来的工作中, 将进一步考虑资源性能动态预测、调度安全等问题.

参考文献:

- [1] Khazaei H, Misic J, Misic V. Performance analysis of cloud computing centers using $M/G/m/m+r$ queuing systems[J]. IEEE Trans Parallel Distributed Syst, 2012(23):5-11.
- [2] Mao M, Li J, Humphrey M. Cloud auto-scaling with deadline and budget constraints[C]// Grid Computing (GRID), 2010 11th IEEE/ACM International Conference. [s. l.]:IEEE, 2010:41 - 48.
- [3] Barham P, Dragovic B, Fraser K, et al; Xen and the art of virtualization[J]. SIGOPS Oper Syst Rev, 2003, 37(5):164 - 177.
- [4] VMware Sta; Virtualization overview. White paper [EB/OL]. (2012-08-25) [2013-04-20] <http://www.vmware.com/pdf/virtualization.pdf>.
- [5] Mell P, Grance T. The NIST definition of cloud computing. Gaithersburg[J]. NIST Special Publication, 2011(20):140-145.
- [6] The CLOUDS Lab. CloudSim: a novel framework for modeling and simulation of cloud computing infrastructures and services[EB/OL]. [2013-04-20] (2011-09-02). <http://www.gridbus.org/cloudsim/>.

作者简介:

赵震男, (1977-), 博士研究生, 讲师. 研究方向为数据库与云计算.

张龙昌男, (1978-), 博士, 副教授. 研究方向为云计算与食品安全信息化.

鄂旭男, (1971-), 博士, 教授. 研究方向为数据挖掘与物联网.

(上接第 162 页)

(1):18-25

- [8] Shamsfard M, Barforoush A A. Learning ontologies from natural language texts[J]. Int'l Journal Human Computer Studies, 2004, 60 (1):17-63.
- [9] Ilario Benetti, Domenico Beneventano, Sonia Bergamaschi. An Information Integration Framework for E-Commerce[J]. IEEE Intelligent Systems, 2002, 17

作者简介:

王哲女, (1981-), 博士生研究生, 讲师. 研究方向为语义信息检索、数据库与数据挖掘.