

云环境下的语义本体构建及其在语义检索中的应用

鲍文燕, 沈岑诚, 刘 博

(扬州大学 信息工程学院, 江苏 扬州 225127)

摘 要: 随着云技术的广泛应用, 云环境下的语义检索也逐渐成为重要的研究内容, 而语义本体构建是首先需要解决的问题. 本文首先提出了一种基于相似度计算的数据部署策略, 综合考虑了概念之间的结构相似度和语义相似度, 将对象数据最优化的部署到不同的虚拟机分别构建语义本体, 并给出了相应的语义检索算法.

关键词: 语义本体; 数据部署; 相似度; 语义检索

中图分类号: TP18

文献标识码: A

文章编号: 1000-7180(2012)11-0066-04

Semantic Ontology Construction and its Application of Semantic Retrieval in Cloud Environment

BAO Wen-yan, SHEN Cen-cheng, LIU Bo

(College of Information Engineering, Yangzhou University, Yangzhou 225127, China)

Abstract: With the wide application of cloud technology, semantic retrieval in cloud environment becomes a important research content, but semantic ontology must be constructed. This paper firstly put forward data deployment strategy based on the similarity, object data will be deployed to different virtual machine according to the structure similarity and semantic similarity of the concept, and then semantic ontology will be constructed respectively and the corresponding semantic retrieval algorithm is proposed.

Key words: semantic ontology; data deployment; similarity; semantic retrieval

1 引言

云计算技术^[1]的出现, 让许多在大数据集下进行数据挖掘与运算的研究拥有了合适的平台. 相对于常规的分布计算, 云计算在理论上拥有无限的空间. 然而, 云环境的分布特性同时也对数据分布提出了较高的要求. 一个合理的数据部署策略可以最大化地减少处理器之间的数据交流, 而最大化单个处理器内部的数据联系.

语义本体是进行语义检索的基础. 概念格可以作为语义本体的基本结构, 其语义本体构建包含了构格, 建树, 映射生成本体三个基本步骤^[2]. 只要将形式背景中的对象进行合理划分, 就可将这三个基

本步骤拆分到云环境下的各个处理机或虚拟机上独立执行. 随着云技术的广泛应用, 云环境下的语义检索也逐渐成为研究内容^[3-4], 但其中如何充分利用云环境的特点降低语义本体的构建复杂度是语义检索效率的关键^[5].

本文提出一种基于相似度计算的数据部署策略, 综合考虑了概念之间的结构相似度和语义相似度, 利用关联度矩阵的办法比较概念间的相似度之后, 对对象数据进行最优化的部署以便于语义本体在云环境下的构建.

2 云环境下的语义本体构建方法

2.1 云环境下的数据部署策略

定义 1 概念 C 的对象集合 $Obj(C)$ 存在对象

收稿日期: 2012-05-12; 修回日期: 2012-06-16

基金项目: 国家自然科学基金(61070133, 61003180); 江苏省自然科学基金(BK2010311); 江苏省教育厅自然科学基金(11KJD510011)

i , 则概念 C 为对象 i 的相关概念, 对象 i 的相关概念集合 T_i 为: $T_i = \{C | i \in \text{Obj}(C)\}$.

定义 2 设 T_i, T_j 分别为对象 i 和对象 j 的相关概念集合, 则对象 i 和对象 j 的关联度 R_{ij} 为对 T_i 和 T_j 中同时出现的概念个数, 即: $R_{ij} = \text{count}(T_i \cap T_j)$.

关联度计算完毕之后, 使用关联度矩阵和矩阵的相关运算对对象进行数据分布部署的规划.

定义 3^[7] M 为对象间的关联度矩阵, 假设总共有 N 个对象, 则 M 为一个 $N \times N$ 的矩阵.

$$M = \begin{bmatrix} R_{11} & R_{12} & \cdots & R_{1N} \\ R_{21} & R_{22} & \cdots & R_{2N} \\ \cdots & \cdots & \cdots & \cdots \\ R_{N1} & R_{N2} & \cdots & R_{NN} \end{bmatrix}$$

对矩阵 M 进行基于关联度的聚类, 使得关联度高的对象尽可能地处于一个子矩阵中. 对关联度的聚类有三步.

$$(1) \text{ 令 } Q = \sum_{i=N, j=N} M_{ij} \times (M_{ij-1} + M_{ij+1});$$

对矩阵进行任意的行列置换, 每次置换之后, 获得当前矩阵的 Q 值;

(2) 选取最大的 Q 值, 将那种行列置换的结果作为关联度聚类的结果;

(3) 将聚类到一起的对象放入一个处理机中进行处理. 可以根据不同的情况来决定对象是否需要重复放置.

具体的操作过程如下:

(1) 对全集形式背景进行属性约简和对象约简, 然后将模糊属性进行 01 化表示之后, 使用渐增构格的方法, 得到一个全局概念格 L_0 .

(2) 设定一个对象支持度阈值 δs , 并且确保 δs 小于等于单个处理机中能够存储的最大对象个数. 对于任何 $C \in L_0$, 如果 C 的对象个数小于等于 δs , 则将 C 存入概念集合 C_0 .

(3) 对于每一个对象 O_i , 计算它的相关概念集 T_i . 然后计算每两个对象之间的关联度 R_{ij} .

(4) 生成对关联度矩阵 M 进行聚类, 并得到对象初步部署策略.

(5) 综合考虑检索要求和相关对象间关联度的大小, 得到对象最终部署策略.

2.2 云环境下的本体构建方法

按照关联度部署完成, 每个处理机上都会一个特定的对象集合, 加进相应的属性集合之后, 每个处理机上都有一个独立的子形式背景. 在每一个处理

机上构建独立的子概念格, 然后转化为子概念树, 并通过映射生成为子语义本体. 为每一个子语义本体设定一个语义标签, 通常这个标签由该子语义本体中根节点的属性集合来表示.

子语义本体完全生成之后, 通过计算子语义本体根节点之间的相似性^[6], 将相似性满足阈值的子本体根节点直接相连, 并在云环境内的虚拟机中生成新的虚拟根节点. 然后再计算这些虚拟根节点之间的相似度. 重复过程直到根节点只剩一个. 此过程中生成的所有节点都称为虚拟节点.

首先使用根节点之间的对象相似度来作为节点之间的相似性^[7]. 但是由于在对象分布过程中, 已经尽可能地将关联度小的对象分布到不同处理机上, 不同处理机之间的对象相似度(即使存在)也是非常小的. 如果在对象分布过程中, 因为检索要求等原因, 不采用对象重复分布的话, 处理机之间的对象相似度为零. 因此在子语义本体根节点相似度计算过程中, 如果对象相似度不满足阈值, 就只能考虑两个节点之间的属性相似度.

定义 4 设 $\text{Obj}(C_i)$ 为概念 C_i 的对象集合, 则概念节点 C_i 和 C_j 之间的对象相似度 $\text{SIM}_1(C_i, C_j)$ 为:

$$\text{SIM}_1(C_i, C_j) = \frac{|\text{Obj}(C_i) \cap \text{Obj}(C_j)|}{|\text{Obj}(C_i) \cup \text{Obj}(C_j)|}$$

定义 5 设 $\text{Atr}(C_i)$ 为概念 C_i 的属性集合, 概念节点 C_i 和 C_j 之间的属性相似度 $\text{SIM}_2(C_i, C_j)$ 为:

$$\text{SIM}_2(C_i, C_j) = \frac{|\text{Atr}(C_i) \cap \text{Atr}(C_j)|}{|\text{Atr}(C_i) \cup \text{Atr}(C_j)|}$$

定义 6 若 $V(C_i, C_j \cdots C_k)$ 为概念节点 $C_i, C_j \cdots C_k$ 所生成的虚拟节点, 则有:

$$V(C_i, C_j \cdots C_k) = ((\text{Obj}(C_i) \cup \text{Obj}(C_j) \cup \cdots \cup \text{Obj}(C_k)), (\text{Atr}(C_i) \cap \text{Atr}(C_j) \cap \cdots \cap \text{Atr}(C_k)))$$

构建虚拟节点过程中先判断两个节点间的 SIM_1 , 如果 SIM_1 不满足阈值则计算它们的 SIM_2 . 如果有一个节点与其他所有节点的 $\text{SIM}_1, \text{SIM}_2$ 都不满足阈值, 则先不连接, 等到之后将它与新生成的虚拟节点再做比较. 注意, 随着层次的升高, $\text{SIM}_1, \text{SIM}_2$ 的判断阈值都应该根据情况适当降低.

构建过程如图 1 所示, 虚线圈即为虚拟节点.

3 云环境下的语义检索算法

传统的语义检索流程, 以 ORS 系统(ORS-Retrieval)^[8] 中的检索模块为例, 过程相对简单, 只需要

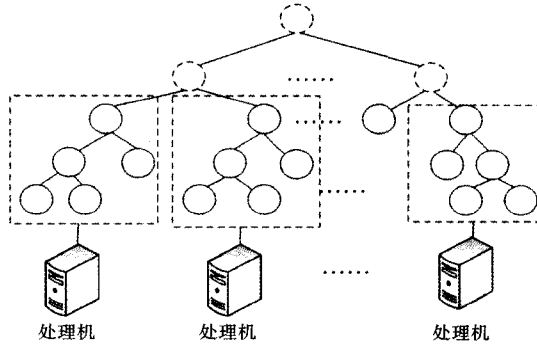


图1 云环境下本体构建示意图

先过滤检索词集,然后寻找匹配概念并返回。但是当本体较大而检索词集又比较精确时,此类检索的时耗就会相对较高。而在云环境下的语义本体构建完成之后,就可以开始进行云环境下的语义检索。具体步骤如下:

(1)首先读取用户输入的一个或多个检索词,去除连词、助词等没有语义含义的词语,得到初始检索词集 Q_0 。

(2)将 Q_0 和全局形式背景的最终属性集合 A_0 进行交运算,得到检索词集 Q 。

(3)将 Q 和各个子语义本体的标签 T 进行重合度计算,计算 $|Q \cap T|/|T|$ 。如果重合度满足初次匹配阈值,再将 Q 和该本体下的各个子节点进行匹配比较,如果和某个节点的属性集满足返回匹配阈值,则返回该节点及其所有子节点的对象集合。

(4)如果 Q 和任意一个语义本体的标签重合度都不满足初次匹配阈值,则向上寻找虚拟节点 S ,计算最大 $|Q \cap S|/|Q|$,如果满足匹配阈值,则输出该节点下面所有子语义本体内的节点对象。

算法1为云环境下的语义检索(C-Retrieval)算法的伪码。

算法1 C-Retrieval(Q_0 , Ont, δ_1 , δ_2 , δ_3)

输入:检索词集 Q_0 ,语义本体 Ont,定位阈值 δ_1 ,精确阈值 δ_2 ,模糊阈值 δ_3 ,子本体标签 T ,虚拟节点 V 。

输出:查询结果 R

- 1 $i=0; V_s = \{V_i | V_i \in \text{Ont}\}; N = \text{count}(V_s);$
- 2 $Q_0 = Q_0 - P;$ /将检索词集中无意义的连词、助词约简掉
- 3 $Q = Q_0 \cap A_0;$ /提取出语义本体中存在的语义特征
- 4 Par do/在各个子本体上并行操作
- 5 if $|T_i \cap Q|/|T_i| \geq \delta_1$ /如果第 i 个子本体的标签与 Q 的重合度满足定位阈值

- 6 if $|S_j \cap Q|/|S_j| \geq \delta_2$ /如果该本体下第 j 个节点的属性与 Q 满足检索阈值
- 7 $R_i = R_i \cup \text{obj}(S_j)$ /将第 j 个节点的对象加入该子本体的子结果集
- 8 Par end
- 9 $R = R \cup R_i$; return(R) /返回检索结果
- 10 else/如果不与任何子本体的标签重合
- 11 For $i=1$ to N
- 12 If $|V_i \cap Q|/|Q| \geq \delta_3$ /如果和第 i 个虚拟节点满足模糊阈值
- 13 return $R = \text{Obj}(\text{Sub}(V_i))$ /返回该虚拟节点所有子节点的对象
- 14 else return("empty")
- 15 End For
- 16 else return("empty")

4 实验及分析

随机使用了15个检索关键词集,采用传统检索方法(ORS-Retrieval)和本文提出的云环境下的检索方法(C-Retrieval)分别进行查询,云环境采用虚拟机软件搭建的Hadoop集群,对检索的时间消耗和准确率进行了比较,如图2和图3所示。

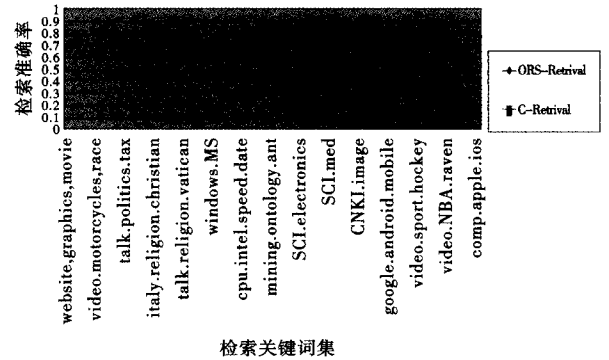


图2 检索准确率比较

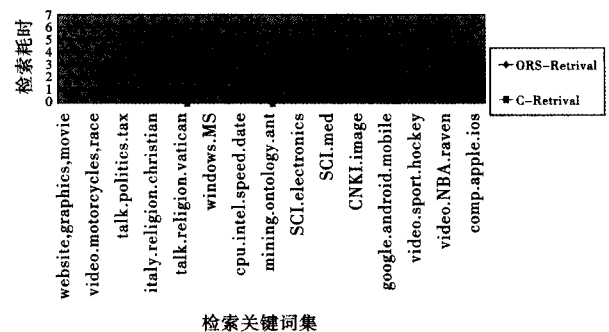


图3 检索耗时比较

可以看出,在大部分情况下,云环境下语义检索的准确率与传统的语义检索不相上下,当检索词集专业性较强时,云环境下的语义检索准确率更高些.同时,在大多数情况下,云环境下语义检索的时间消耗远远优于传统方法.但是在少数特殊情况下,本章的检索方法不仅耗时巨大,而且准确率极低,这是因为检索过程中无法在某个处理机上找到匹配节点,而不得不寻找虚拟节点并返回虚拟节点的所有子节点对象,导致准确率和效率都大大降低.

5 结束语

云环境的应用在近几年越来越受到重视.利用云环境的分布特性,可以大大降低很多运算的复杂度.本章考虑了云环境下的语义本体构建问题,利用计算对象间相似度的方法,结合矩阵的转置,对对象进行了合理的数据分布,不仅很好地利用了云环境可无限扩展的特性增加了领域本体的扩展性,还利用了云环境内可虚拟硬件的方式对构建本体进行帮助.同时针对云环境下的新型语义本体模型,提出了一种更符合云环境下语义检索的流程.

参考文献:

- [1] Michael Armbrust, Armando Fox, Rean Griffith, et al. Above the clouds; a berkeley view of cloud computing [R]. UC Berkeley Reliable Adaptive Distributed Systems Laboratory, 2009; 1-23.
- [2] Liya Fan, Tianyuan Xiao. FCA-Mapping: a method for ontology mapping[C]// Proc. of The 4th European Semantic Web Conference. Innsbruck, Austria, Springer-Verlag, 2007; 192-206.
- [3] Jaeyong Kang, Kwang Mong Sim. Towards agents and ontology for cloud service discovery[C]// 2011 In-

ternational Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC). India, Kanpur, 2011; 483-490.

- [4] Jaeyong Kang, Kwang Mong Sim. Ontology and search engine for cloud computing system[C]// 2011 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC). India, Kanpur, 2011; 276-281.
- [5] Hong Zhou, Hongji Yang, Hugill A. An ontology-based approach to reengineering enterprise software for cloud computing[C]// Computer Software and Applications Conference (COMPSAC), 2010 IEEE 34th Annual. Seoul, Korea, 2010; 383-388.
- [6] Dong Yuan, Yun Yang, Xiao Liu, et al. A data placement strategy in scientific cloud workflows[J]. Future Generation Computer Systems, 2010(26); 1200-1214.
- [7] Yang Liu, Zhiqing Shao. The similarity calculation of concepts from different ontologies based on cosine[C]// 2010 International Conference on Information Management, Innovation Management and Industrial Engineering (ICIII). Hongkong, 2010; 130-134.
- [8] Myunggwon Hwang, Hyunjang Kong, Pankoo Kim. The design of the ontology retrieval system on the web [C]// Advanced Communication Technology, ICACT 2006. The 8th International Conference. Korea, 2006; 1815-1818.

作者简介:

鲍文燕 女, (1964—), 工程师. 研究方向为信息处理、计算机应用技术等.

沈岑诚 男, (1986—), 硕士研究生. 研究方向为概念格、数据挖掘、信息检索等.

刘博 男, (1987—), 硕士研究生. 研究方向为数据挖掘, 云计算等.