双聚类算法在本体构建中的应用

朱利达,徐少华,韩立新,曾晓勤
(河海大学 计算机与信息学院,江苏 南京 211100)

摘 要:随着互联网上数据的增长,如何更有效地利用数据成为了一个亟待人们解决的问题。为此语义网被提出,使得机器可以帮助人们处理这些数据。语义网的核心是本体,因此语义网的发展和人们对互联网上信息的本体构建相关。如何快速准确地构建本体是语义网发展的关键。构建语义本体本身又是一件繁琐的工作,当今的本体学习技术利用机器学习和数据挖掘的技术来实现本体的自动或半自动构建。文中将双聚类算法的思想引入到本体构建当中,同时提出了 FIU-CTWC 双聚类算法,解决了一维聚类不能聚出多重语义的问题。

关键词:双聚类;本体自动构建;语义网
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2013)03-0027-04
doi:10.3969/j.issn.1673-629X.2013.03.007

Application of Biclustering Algorithm in Construction of Ontology

ZHU Li-da,XU Shao-hua,HAN Li-xin,ZENG Xiao-qin
(College of Computer and Information,Hohai University,Nanjing 211100,China)

Abstract:As the amount of data in the Internet become more and more large,how to use the data more effectively turn into a problem which need people to solve. At this background,the semantic web was provided. Using semantic web computer can help people to handle the data. The development of semantic web largely depends on the number of ontology in the Internet. How to construct ontology quickly and precisely is the key of the semantic web's development. Ontology learning aims to construct the ontology automatically or semi-automatically with the help of techniques like machine learning and data mining. It uses a refine biclustering algorithms FIU-CTWC to construct the ontology in order to overcome the shortage of 1D clustering algorithms.

Key words:biclustering algorithms;ontology auto construction;semantic web

0 引 言

万维网的兴起带来了人们利用信息的一场革命。从 20 世纪 80 年代后期万维网被提出以来,万维网上数据的增长速度是惊人的,目前万维网已经达到了很大的规模。受制于当前万维网的局限性,人们并不能更有效地利用数据。因为当前绝大部分的内容都是面向人的,也就是说人们不能充分利用计算机的能力来处理如此海量的数据。1998 年万维网联盟的蒂姆·伯纳斯-李提出了语义网的概念[1]。语义网的核心是通过给万维网上的资源添加能够被计算机理解的元数据,使得当前的万维网变得更加智能,信息更加整合。本体是语义网中最重要的概念。Nechs[2] 给出本体定义是"给出构成相关领域词汇的基本术语和关系,以及利用这些术语和关系构成的规定这些词汇外延的规则的定义"。一个更加流行的本体定义由 T. R. Gruber[3] 提出,"一个本体是一个概念体系的显示的形式化规范。本体提供了对给定领域的一种共识,这种共识消除了资源语义上的差别。这种共识使得计算机的理解成为可能"。语义网的普及依赖于本体的构建,但手工方式构建本体需要大量的人力和时间,因此迫切地需要有方法能够自动或是半自动地构建本体,国内也已经展开了相关的研究工作[4]。

本体结构定义为一个 5 元组[5,6] $O=\{C,R,H^c,Re/,A^o\}$。C 是概念的集合;R 是概念间关系的集合;H^c 是概念间的分类关系;Rel 是概念间的非分类关系;A^o 表示公理。本体学习由易到难包括概念学习,关系学习和公理学习。当前人们对概念学习和关系学习进行着广泛而深入的研究。本体学习算法可以采用基于统计的方法、基于关联规则的算法和基于聚类[6] 的方法。文中将双聚类的方法引入本体学习中,提出了 FIU-CTWC 双聚类算法。该方法克服了传统单聚类方法在

收稿日期:2012-06-11;**修回日期**:2012-09-13
基金项目:国家自然科学基金资助项目(60971088);江苏省高校"青蓝工程"中青年学术带头人培养对象资助项目
作者简介:朱利达(1986-),男,硕士研究生,主研领域为数据挖掘、信息检索;韩立新,教授,博士生导师,主研领域为信息检索、模式识别、数据挖掘。

本体学习中不能有效聚出多重语义的不足。

1 双聚类算法简介

数据矩阵 $\boldsymbol{M}_{m\times n}$ 是 m 行 n 列的矩阵,其中 $\boldsymbol{X}=\{X_1,X_2,\cdots,X_m\}$ 和 $\boldsymbol{Y}=\{Y_1,Y_2,\cdots,Y_m\}$ 表示矩阵 $\boldsymbol{M}$ 的行和列向量的集合,那么 $\boldsymbol{M}$ 可以表示成($\boldsymbol{X}$,$\boldsymbol{Y}$),m_{ij}表示矩阵中的第 i 行 j 列的数值。如果 $\boldsymbol{I}\subset\boldsymbol{X}$,$\boldsymbol{J}\subset\boldsymbol{Y}$,分别表示矩阵 $\boldsymbol{M}$ 的行和列向量子集,那么 $\boldsymbol{M}_{\boldsymbol{IJ}=(\boldsymbol{I},\boldsymbol{J})}$ 表示矩阵 $\boldsymbol{M}$ 的子矩阵,其中它包含了 $\boldsymbol{I}$ 中所有的行和 $\boldsymbol{J}$ 中所有的列。一个双聚类(bicluster)就是 $\boldsymbol{M}$ 中这样的一个子矩阵 $\boldsymbol{M}_{\boldsymbol{IJ}=(\boldsymbol{I},\boldsymbol{J})}$,其中 $\boldsymbol{I}\subset\boldsymbol{X}$,$\boldsymbol{J}\subset\boldsymbol{Y}$,并且在这个子矩阵中的每一行或列都表现出一定程度的相似性[7]。

双聚类根据聚集方式和程度的不同,最终形成的双聚类类型也是不同的,主要有四种双聚类[8]:

类型 1:双聚类的所有的数值都是相等的;

类型 2:双聚类同一行或者同一列上的数值是相等的;

类型 3:双聚类的同行或者列上数值的变化趋势近似相同;

类型 4:双聚类行或者列上的信息变化趋势一致。

双聚类的结果的结构,也就是双聚类算法最终在搜索到的子矩阵在原来总矩阵中的相对位置关系,也有多种类型。根据具体的应用背景,应该选择适宜的结构。

双聚类算法中主要的候选集合选择策略有以下五类:

类型 1:迭代合并行聚类和列聚类的结果;

类型 2:分割和攻破策略;

类型 3:贪心代搜索子空间的方法;

类型 4:穷举法双聚类;

类型 5:分布参数识别法。

迭代合并行聚类和列聚类的结果的方法最为直接,它是将传统聚类的方法用在行向量集合和列向量集合上面,然后通过某种迭代过程来合并这些聚类以得到结果。

耦合双向聚类算法[9,10](coupled two-way clustering, CTWC)采用迭代合并行和列的聚类方法来得到双聚类。该算法对行的集合和列的集合分别进行层次聚类(步骤 3,4),并迭代这个过程来寻找符合条件的稳定子集,直到不再有符合条件的新的稳定聚类出现则终止。此算法的描述如下:

TWOWAY($\boldsymbol{U}$, $\boldsymbol{V}$, ALG)

$\boldsymbol{U}$:矩阵行向量

$\boldsymbol{V}$:矩阵列向量

ALG:一维聚类算法, 输入一个矩阵,输出稳定的行或者列聚类

开始:$\boldsymbol{u}=\{\boldsymbol{U}\}$,$\boldsymbol{v}=\{\boldsymbol{V}\}$

步骤 1:当列向量集合 $\boldsymbol{u}$ 和行向量集合 $\boldsymbol{v}$ 不为空;

步骤 2:对每个集合 $\boldsymbol{u}$ 和集合 $\boldsymbol{v}$ 生成的矩阵;

步骤 3:用 ALG 进行聚类,生成稳定的行聚类 $\boldsymbol{u}$;

步骤 4:用 ALG 进行聚类,生成稳定的列聚类 $\boldsymbol{v}$;

步骤 5:保存中间结果;

步骤 6:用 $\boldsymbol{u}$ 和 $\boldsymbol{v}$ 迭代进行上述步骤。

普通聚类方法的选择要具体问题具体分析,在 Getz, Levine 以及 Domany 提出这个算法的时候,他们是用于对基因的聚类,采用的普通聚类是层次聚类的方法[11]。在寻找符合条件的稳定子集这步,可以利用启发式的搜索方法来避免聚类产生所有可能的行列合并。

2 FIU-CTWC——CTWC 算法在语义本体领域的改进

2.1 聚类候选集合选取

由词和词的语义所构成的语义矩阵有一个非常明显的特点,那就是无论行向量还是列向量都是比较稀疏而且分散的。比如说“苹果”这个词,具有多重的语义特性,可以是一个品牌,也可以属于一种水果,那么解释“苹果”这个词的行向量参加普通的聚类必定得不到良好的结果。可能会得到和水果不太相似,和品牌也不太相似。如图 1, W_1 和 W_n 共同由 P_1 中的词解释,W_n 和 W_2 又共同由 P_2 中的词解释,W_n 有和 W_1 相近的意思也有和 W_2 相近的意思,但是在整个矩阵上运用一维聚类,可能 W_n 和 W_1 及 W_2 的相似度都不高。文中在考虑了本体构建特点的情况下,改进了 CTWC 算法的候选集选取使得此算法更加适用于本体构建。

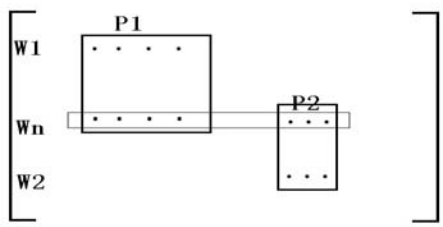

图 1 词和解释矩阵

改进候选集合选取算法的主要思想是将多重语义的词在预处理阶段就分离到不同的候选集中去,即在运用一维聚类之前将整个空间划分成不同的超平面,如图 2,在不同的超平面中再进行一维聚类 。一个词在不同的超平面中出现,在不同的超平面中语义不同。这样在不同候选集上进行的聚类,就能聚类出词的多重语义。定义如下的语义矩阵 $W\times P$,此矩阵是一个 0-1 矩阵,行由一个词的解释所组成。定义这个矩阵的行向量为 $W_i=(p_1,p_2,\cdots,p_n)$,行向量的含义为词 W_i 由 $p_1,p_2,\cdots,p_n$ 解释。定义此矩阵的列向量为 $p_i=$

$(w_1, w_2, \cdots, w_n)$,列向量的含义为词 p_i 能解释 $w_1, w_2, \cdots, w_n$ 这些词。可以利用列向量求出 w_i 和 w_j 的解释共同包含词 p。也可以求得共同被 n 个词解释过的词的集合。这样就可以得到很多个小矩阵。

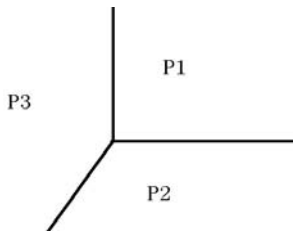

图 2　不同词的子空间

候选集选取算法如下:

子程序

输入: properties 的个数为 $2n$ 的候选集合

输出: properties 的个数为 $2n+1$, $2n+2, \cdots, 2^{n+1}$ 的候选集合

步骤 1: $\text{outputpro}_k = \text{inputpro}_i \cup \text{inputpro}_j$

对 Input 集合中的任意两个元素 input_i 和 input_j 的 properties 求并集,得到一个新的 properties 集合,这个集合中 properties 的元素个数可能是 $2n+1$ 到 2^{n+1}。

步骤 2: $\text{outputword}_k = \text{inputword}_{i \cap \text{inputword}_j}$

对第 1 步中两个元素 input_i 和 input_j 的 word 求交集,得到一个新的 word 集合。

步骤 3:将 1,2 两步中得到的 properties 和 word 集合组成一个新的元素,将这个元素放置到对应 properties 个数的集合中。

步骤 4:重复 1,2,3 步骤直到 Input 集合中任意两个元素都完成了上述步骤。

主程序 MPD:

步骤 1:

for i from 0 to 3:

取出 properties 个数为 $2i$ 的集合,调用上述子程序求得 properties 个数为 $2i+1$ 到 2^{i+1} 的集合。

步骤 2:

for j from 16 to 1:

for k from $j-1$ to 1

查看 properties 个数为 j 的集合中每个元素中 word 子集,如果 properties 个数为 k 的集合中有与之相同的 word 子集的元素,那么就删除 k 中的该元素,也就是说使得同样的 word 子集的 properties 最大化。

通过上述的方法,求得候选集合 properties 个数从 1 到 16 的集合。假设这样的集合 word 个数是 I, properties 个数是 J,那么 $I \times J$ 的子矩阵就是候选矩阵。

2.2　聚　类

改进 CTWC 算法,根据特点命名为 First-Intersection-Union CTWC(FIU-CTWC),那么 FIU-CTWC 算法的双聚类部分:

1)候选集合,也就是双聚类子矩阵已经通过上面的方法选出,这样可以保证优质的双矩阵不会被其他信息干扰而导致其信息丢失。每个集合中的每个元素就是一个子矩阵,即 Word × Properties 的矩阵。

2)在最原始的语义本体矩阵 $\boldsymbol{M}$ 中,找到每个 Word × Properties 子矩阵,对每个这样的子矩阵进行行向量集合的聚类。这里使用的聚类与普通聚类十分相似,但是也有不同的地方。这里使用的聚类方法,作者称之为动态复用聚类(Dynamic Multiplexing Clustering, DMC)。

3)DMC 聚类算法,先作个简单的介绍,Dynamic 说明聚类结果的类别个数不定;Multiplexing 说明在所有的元素中同一个元素可能属于不同的几个聚类中。DMC 算法思想:对任意两个元素求其相似度,使用向量之间夹角公式,假设两个向量是 $\boldsymbol{A}$, $\boldsymbol{B}$: $\cos<\boldsymbol{A},\boldsymbol{B}> = \frac{\boldsymbol{A}\cdot\boldsymbol{B}}{|\boldsymbol{A}||\boldsymbol{B}|}$,当两个向量的角度小于 30°,那么可以看做两个本体相似。将所有的本体建立一个图,将相似的本体之间连接一条边找到图中的完全子图,将这些完全子图当成一类。见图 3。

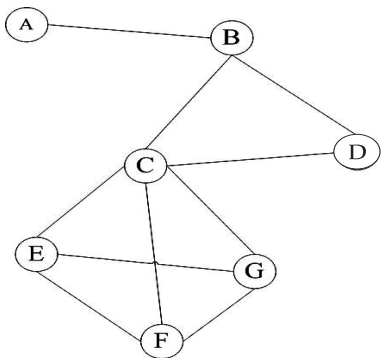

图 3　本体关系的图表示

在这个 DMC 聚类的结果图中,可以有三个聚类,也就是 3 个完全图(A, B), (B, C, D), (C, E, F, G)。这里把原本为(A, B, C, D, E, F, G)的候选集合双聚类成了 3 个集合。每个集合中的 properties 不变,word 部分被拆分。这里本体 B 和 C 都被复用了,分别都在两个不同的聚类结果中。

3　实验及结果分析

文中选用了《现代汉语词典》作为实验数据。《现代汉语词典》中的每一项由词和一段对该该词的解释组成,这个词的每种解释占一行。

比如名词苹果的解释构成如下:

苹果.

(1)落叶乔木,叶子椭圆形,花白色带有红晕。果

实圆形,味甜或略酸,是普通水果。

(2)该公司由乔布斯、盖瑞等于 1976 年 4 月 1 日创立,以咬了一口的苹果为公司 LOGO。

实验期待从中找出词之间的分类关系和层次关系。实验步骤如下:

步骤 1:对词的解释进行分词。得到 $W_i = (p_1, p_2, \cdots, p_n)$,其中 w 为词,p_i 由词 w 的解释通过分词得到。在这里分词使用最大正向匹配方法进行分词。所有的词和解释组成矩阵 $\boldsymbol{M} = W \times P$。

步骤 2:计算 p_i 对 w_i 的解释的贡献度,给第一步得到的 $\boldsymbol{M}$ 矩阵赋值。利用 TFIDF[12] 的算法思想,用一个称之为 TFIOF 的参数来表示 (w_i, p_i) 值。TFIOF 中有如下几个参数:

1. TF (j) 词频,指在 w_i 的解释中 p_i 出现的次数。

2. OF(j) 指属性 p_i 用以解释的词数。

3. IOF(j) :逆向频率,用总共的词数量除以 OF(j) ,可以算出 p_i 对所有语义本体的平均重要度,即 $\mathrm{IOF}(j) = n/\mathrm{OF}(j)$ 。

TFIOF(j) 则可以类似于 TFIDF 的求法,可以使用公式 $\mathrm{TFIOF}(j) = \mathrm{TF}(j) \times ln(n/\mathrm{OF}(j))$

步骤 3:对 $\boldsymbol{M}$ 用文中之前介绍过的 FIU-CTWC 算法进行候选集合挑选,得到一系列的矩阵 $\boldsymbol{M}_i$ 之后在这些得到的矩阵上进行聚类。

对实验结果的分析发现,对名词能得到较好的结果,能够较好地聚得同义词和层次关系,同时能够发现名词的多义。比如“苹果”一词由 FIU-CTWC 的候选集选取算法可将它的水果词义和品牌词义挑选到不同的候选集合中,在一个候选集合中,苹果和香蕉聚得到相近的词义,同时由最大权重的解释“水果”,得到“苹果”这词的上位词是“水果”,而在品牌候选集中,苹果又和其他品牌聚得相近的词义。

但是本实验也存在着很多不足。首先本实验所使用的数据不够一般化。本实验所用的《现代汉语词典》已经给出了词和它和的解释,通过对解释的分词能够得到词和属性关系。怎么在无结构的纯文本中抽取出词和它的描述属性更具有实际意义。其次用 TFIOF 值来描述属性对一个词的重要程度也存在一定问题。因为属性出现的频数和语义上的重要程度不那么明显。通常聚类的结果依赖于两个方面,距离矩阵的获取和聚类算法的选择。不准确的距离矩阵也影响预处理后一维聚类的准确性。TFIOF 方法也影响层次关系的获取,因为这种方法不是很能衡量哪个属性是词的上一层,但是在考虑名词的情况下实验能得到不错的结果。

4 结束语

人们有效利用互联网数据的愿望,催生了语义网的诞生和发展,因此也带动了本体自动构建的研究热潮。文中受到双聚类算法 CTWC 的启发,将双聚类的思想引入到了本体构建中。又根据本领域的特点,调整了 CTWC,提出了 FIU-CTWC 双聚类算法。最后实验验证了文中的想法。本体自动构建的研究仍有大量的工作要做。文中的方法是对本体语义的分类关系的发掘的一种尝试,实验中仍然有许多不足和有待改进的地方。

参考文献:

[1] Berners-Lee T, Hendler J, Lassila O. Semantic web[J]. Scientific American, 2000, 1(1): 68-88.

[2] Neches R, Fikes R E, Gruber T R. Enabling technology for knowledge sharing[J]. AI Magazine, 1991, 12(3): 36-56.

[3] Gruber T R. A translation approach to portable ontology specification[J]. Knowledge Acquisition, 1993, 5(2): 199-200.

[4] 薛中玉,李春梅. 基于文本挖掘的本体自动构建系统架构解析[J]. 计算机技术与发展,2011,21(1):100-103.

[5] 杜小勇,李 曼,王 珊. 本体研究综述[J]. 软件学报,2006,17(9):1837-1847.

[6] Maedche A. Ontology Learning for the Semantic Web[M]. Boston: Kluwer Academic Publishers, 2002.

[7] Bisson G, Nedellec C, Canamero D. Designing clustering methods for ontology building: The Mo'k work bench[C]//Proceeding of the ECAI 2000 Workshop on Ontology Learning (OL'2000). [s. l.]: [s. n.], 2000.

[8] Cheng Y, Church G M. Biclustering of expression data[C]// Proceeding of the 8th International Conference on Intelligent Systems for Molecular Biology. [s. l.]: [s. n.], 2000: 93-103.

[9] Gu Jiajun, Liu J S. Bayesian biclustering of gene expression data[J]. BMC Genomics, 2008, 9(Sup): S1-S4.

[10] Busygin S, Prokopyev O, Pardalos P M. Biclustering in data mining[J]. Computers & Operations Research, 2008, 35(9): 2964-2987.

[11] Tanay A, Sharan R, Shamir R. Biclustering Algorithms: A Survey[D]. Israel: Tel-Aviv University, 2004.

[12] Zhang Wen, Yoshida T, Tang Xijin. TFIDF, LSI and Multi-word in Information Retrieval and Text Categorization[C]// 2008 IEEE International Conference on Systems, Man and Cybernetics. [s. l.]: [s. n.], 2008: 108-113.

双聚类算法在本体构建中的应用

作者： 朱利达， 徐少华， 韩立新， 曾晓勤
作者单位： 河海大学 计算机与信息学院,江苏 南京211100
刊名： 计算机技术与发展
英文刊名： Computer Technology and Development
年，卷(期)： 2013(3)