中文网络百科开放分类层次结构树及其聚类算法研究*

贾 真, 尹红风, 李天瑞
(西南交通大学 信息科学与技术学院, 成都 610031)

摘 要: 为利用开放分类进行百科条目的分类和检索,提出了基于词共现和语义分析的开放分类聚类算法以及开放分类层次结构树构建方法;为了进一步提高层次结构树的聚合度,提出了基于相似度和相关度计算的层次结构树聚类算法。以互动百科开放分类为实验数据集,实验结果表明,所构建的开放分类层次结构树的准确率较高,利用开放分类层次结构树有效提高了百科条目检索的效率。

关键词: 开放分类; 聚类; 共现; 语义分析; 层次结构树
中图分类号: TP391;TP301.6 **文献标志码**: A **文章编号**: 1001-3695(2013)06-1660-04
doi:10.3969/j.issn.1001-3695.2013.06.014

Research on Chinese online encyclopedia open category hierarchy tree and clustering algorithms

JIA Zhen, YIN Hong-feng, LI Tian-rui
(*School of Information Science & Technology, Southwest Jiaotong University, Chengdu* 610031, *China*)

Abstract: To use open category to categorize and retrieve the encyclopedia entries, this paper proposed the open category clustering algorithm based on word co-occurrence and semantic analysis, and the method of generating open category hierarchy trees. It proposed a hierarchy tree clustering algorithm based on similarity and correlation computing for increasing the quality of hierarchy trees. Experimental data set was downloaded from Hudong online encyclopedia. The experimental results show that the proposed algorithms achieve high precision and the generated hierarchy trees effectively improve the efficiency of encyclopedia entries retrieval.

Key words: open category; clustering; co-occurrence; semantic analysis; hierarchy tree

"开放分类"是中文网络百科(互动百科、百度百科等)的特定词汇。它是指比传统的目录式分类更灵活、更具自主性的分类方式。用户可以直接根据条目内容来添加相关的一个或多个开放分类,而不必将条目强制归入某个已经设定好的分类。与开放分类类似的协同标注、大众分类法多是让用户自由使用标签对文章、图片等进行描述,并利用这些用户标签进行信息资源的分类、组织和检索。然而,开放分类是用户按照自己的理解和习惯自由添加的,用词没有统一标准,并且开放分类是平行的,没有构成层次结构,对信息的浏览和检索带来诸多不便。对开放分类进行聚类研究,对分类体系的自动构建、百科知识条目的分类、检索和推荐等方面有着广泛的应用前景。开放分类的聚类与传统的文本聚类存在很大差别,因为开放分类一般都是单个的词或短语,包含信息量少,难以自动抽取其特征[1]。开放分类聚类的相关工作主要有文献[2~7]等。文献[2]根据标签的共现次数,对每个标签建立相关标签无向图,再采用图聚类算法进行标签聚类,得到语义相关的标签。文献[3]提出基于余弦向量相似度计算标签相似度,对网站 Del.icio.us 中的约六万个标签进行聚类,将相关标签连接成无向图,再采用相关算法将无向图转换为层次结构。文献[4]采用凝聚式层次聚类算法对标签进行聚类研究,利用相关标签的权重,计算标签之间的相关度,从而实现标签的聚类。Sanderson 等人[5]提出利用词语共现分析代替标准聚类算法获得词语之间的语义包含关系或层次关系,再生成词语的概念层次结构。Schmitz[6]改进了文献[5]的共现概率模型,将图片网站(www.flickr.com)中约二十万个标签进行聚类并归纳出分面本体,将分面本体与标签系统集成,提供了更加灵活的搜索和浏览接口。Ponzetto 等人[7]利用语义分析从维基百科分类标签中找出具有 Is-A 关系的分类标签,再构建分类层次关系。

本文将共现分析与语义分析两种方法相结合,提出开放分类包含关系对生成方法,并利用开放分类之间的包含关系构建出开放分类层次结构树。使用该分类树能够有效提高对百科条目进行检索的效率。针对开放分类层次结构树个数过多的问题,本文又提出了层次结构树的相似度和相关度计算方法,将满足最小相似度和最小相关度阈值的层次结构树进行合并,进一步提高了开放分类层次结构树的聚合度。

1 开放分类包含关系生成方法

开放分类包含关系生成方法分为三个步骤:a)基于共现

收稿日期: 2012-09-24; **修回日期**: 2012-11-08 **基金项目**: 国家自然科学基金委员会主任基金资助项目(61152001);中国科学院自动化所复杂系统管理与控制重点实验室开放课题(20110102);中央高校基本科研业务费专项基金资助项目(SWJTU11ZT08)

作者简介: 贾真(1975-),女,河南开封人,讲师,博士研究生,主要研究方向为信息抽取、内容安全(zjia@home.swjtu.edu.cn);尹红风(1964-),男,教授,博士,主要研究方向为大数据、搜索引擎; 李天瑞(1969-),男,教授,博导,博士,主要研究方向为智能信息处理、数据挖掘、云计算.

模型生成开放分类包含关系对;b)基于语义分析生成开放分类包含关系对;c)将两种方法生成的包含关系对合并。

1.1　基于共现模型的开放分类包含关系生成方法

本文基于文献[6]的共现模型计算开放分类的共现概率,其计算方法如下:

$$若\ P(X|Y)\geqslant t\ 并且\ P(Y|X)<t,则\ X\ 包含\ Y$$
$$E_x\geqslant E_{\min},E_y\geqslant E_{\min}$$

其中:X、Y 分别代表两个开放分类;E_x、E_y 代表 X 和 Y 标记百科条目的次数;t 为共现阈值,取值范围为(0,1],满足共现阈值 t 的两个开放分类构成包含关系;$E_{\min}$为开放分类最小出现次数,用于过滤出现次数较少的开放分类;(X,Y)为包含关系对,又称为父子对。

基于共现模型的开放分类聚类算法 OCCA(open category clustering algorithm)如下。

输入:百科条目开放分类集合 $C=\{(c_{11},c_{12},\cdots),(c_{21},c_{22},\cdots),\cdots,(c_{n1},c_{n2},\cdots,c_{nk},\cdots)\}$($c_{nk}$表示第 n 个百科条目的第 k 个开放分类),共现阈值 t,开放分类最小出现次数 $E_{\min}$。

输出:开放分类集合 $C=\{c_i\}$,父子对集合 $P=\{(c_i,c_j)\}$。

a)将每个条目的开放分类 c_{nk} 存入 HASH_{tag} 表中,对 HASH_{tag} 中每个开放分类的出现次数进行统计,记为 $\text{Supp}(c_i)$。

b)将条目中同时出现的两个开放分类(c_{nq}, c_{nk})放入 $\text{HASH}_{\text{co-occur}}$ 表中,对表中开放分类的共现次数进行统计,记为 $\text{Supp}(c_i,c_j)$。

c)若 $\text{HASH}_{\text{co-occur}}$ 表中共现的两个开放分类 c_i 和 c_j 满足 $\text{Supp}(c_i)\geqslant E_{\min}$,且 $\text{Supp}(c_j)\geqslant E_{\min}$,则计算 $P(c_i|c_j)$ 和 $P(c_j|c_i)$。

d)若 $P(c_i|c_j)\geqslant t$,且 $P(c_j|c_i)<t$,则 c_i 包含 c_j,(c_i,c_j)为父子对,将(c_i,c_j)存入 $\text{HASH}_{\text{subsumption}}$ 中。

e)遍历 HASH_{tag} 输出开放分类集合 C。

f)遍历 $\text{HASH}_{\text{subsumption}}$ 输出父子对集合 P。

1.2　基于语义分析的开放分类包含关系生成方法

Is-A 关系指一个概念包含在另一个概念的外延之中,或指概括性的事物类别与该类事物的个体实例之间的关系。例如:“苹果”与“水果”之间具有 Is-A 关系,“水果”与“植物”之间具有 Is-A 关系。汉语词汇在字面上构成具有重心后移的特点。重心后移是指在汉语词汇中,语素越靠后,在表达主题概念中所起的作用越大[8],主题中心词往往在词的后半部分。如果某些词的后半部分相同,则可以认为是同一主题的词语。例如,“中国演员”和“香港演员”的后半部分都是“演员”,则可认为这两个词是同一主题,并且这两个词都是“演员”的下位词,与“演员”具有 Is-A 关系。根据汉语词汇重心后移特点,对开放分类进行字面相似性分析,可以从中找出具有 Is-A 关系的开放分类。本文将具有 Is-A 关系的开放分类组成包含关系对,如(演员,香港演员)为一个包含关系对。本文采用西南交通大学的耶宝智慧中文分词系统(http://www.yebol.com.cn/)对开放分类进行分词预处理。分词后的开放分类表示为 $c_i=\text{seg}_{i1}\text{seg}_{i2}\text{seg}_{i3}\cdots\text{seg}_{in}$($c_i\in C$,$C$ 为开放分类集合)。seg_{in}表示分词后的开放分类中第 n 个词。将开放分类集合中的每个开放分类与其他开放分类进行比较,如果一个开放分类 c_j 包含于另一个开放分类 c_i,并且 $c_j=\text{seg}_{ik}\text{seg}_{i(k+1)}\cdots\text{seg}_{in}$($2\leqslant k\leqslant n$),即 c_j 中词的个数小于 c_i 中词的个数且为 c_i 的后面部分,则 c_i、c_j 为具有 Is-A 关系的词对,即 c_j 包含 c_i,记为(c_j,c_i)。

1.3　将两种方法生成的包含关系对合并

仅利用共现概率和共现阈值生成包含关系对会造成某些包含关系的遗漏。不满足共现阈值的开放分类之间也可能具有包含关系。采用语义分析找出具有 Is-A 关系的开放分类,能够将某些不满足共现阈值的开放分类组成包含关系对,在一定程度上弥补了共现概率的缺点。为了得到更多的父子对,需将两种方法结合,即将两种方法中产生的父子对进行合并。若两种方法产生的父子对相矛盾,如共现模型得到的父子对(中国版画家,版画家)与字面相似性分析得到的父子对(版画家,中国版画家)相矛盾,则以字面相似性分析得到的父子对为准,将共现模型得到的父子对删除。

2　开放分类层次结构树构建方法

生成开放分类父子对后,利用父子关系(包含关系)构建开放分类层次结构树。例如,图 1 中有四个包含关系对:X 包含 Y、X 包含 Z、Y 包含 Z、Y 包含 L。X 包含 Y 和 Y 包含 Z 两个包含关系对连接成 X、Y、Z 的层次结构,同时去掉 X 包含 Z 的关系。如图 2 所示,X、Y、Z、L 四个开放分类构成一个层次结构树。

开放分类层次结构树生成算法 HTGA(hierarchy tree generation algorithm)如下:

输入:父子对集合 $P=\{(c_i,c_j)\}$。

输出:层次结构树集合 $T=\{t_1,t_2,\cdots,t_n\}$。

a)将 c_i 及 c_i的子存入链表 list_c_i。c_i 为链表的头。

b)所有链表组成链表集合 List。

c)for each list_$c_i\in$ List do

(a)如果 $c_j\in$ list_c_i 并且 $c_j\in$ list_c_k 并且 $c_k\in$ list_c_i,则将 c_j 从 list_c_i 中移去;

(b)如果 $c_k\in$ list_c_i,递归遍历 list_c_k,输出以 c_i 为根的树 t_{ci}。

end for

3　层次结构树聚类方法

由于生成的层次结构分类树个数较多,不便于进行百科条目的检索,本文提出了层次结构树相似度与相关度计算方法。将满足最小相似度与最小相关度阈值的层次结构树聚合在一起,减少孤立分类树的个数。

3.1　层次结构树相似度计算方法

不同的层次结构树可能具有相同的节点或子树。例如,图 3 中的两个层次结构树具有相同的节点 P 和 S。虽然两个层次结构树的根节点不是包含关系对,但如果两个层次结构树中有很多相同的节点,说明它们的根节点之间可能具有语义关系。相同节点数目越多,两个树根节点之间的语义关系越紧密[9]。基于层次结构树相同节点的情况,本文提出一种层次结构树相似度计算方法。

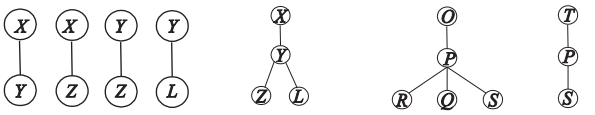

图1　包含关系对　　图2　层次结构树　　图3　具有相同节点的层次结构树

定义 1　层次结构树相似度,即 $\text{sim}(A,B)$的计算公式为

$$\text{sim}(A,B)=\frac{\sum_{i=0}^{m}\ \sum_{j=0}^{n}f(A_i,B_j)}{N_B}$$
$$N_A\geqslant N_{\min},N_B<N_{\min}\qquad(1)$$

其中:A、B 分别代表两个开放分类层次结构树,N_A 为树 A 节点数,N_B 为树 B 节点数,$N_{\min}$为节点数阈值。满足相似度阈值的两个树合并为一个树。为了降低计算复杂度,将节点数小于 $N_{\min}$的树作为 B,节点数大于等于 $N_{\min}$的树作为 A;A_i 为 A 的节点,i 为节点编号,A_0 为 A 的根节点;B_j 为 B 的节点,j 为节点编

号,B_0 为 B 的根节点。

$$当 A_i = B_j 时, f(A_i, B_j) = 1$$
$$当 A_i \neq B_j 时, f(A_i, B_j) = 0$$

满足相似度阈值的两个层次结构树根节点成为包含关系(A,B),从而使两个树合并为一个树。

3.2 层次结构树相关度计算方法

开放分类的共现情况可以反映两个开放分类是否具有语义关系。层次结构树之间节点的共现情况可以反映出这两个树之间是否相关。层次结构树上节点之间的共现是指一个树上的节点与另一个树上的节点同时标注一个百科条目。两个树共现节点的个数越多,则这两个树之间的相关性越强。本文提出层次结构树之间的相关度计算方法,将满足相关度阈值的两个层次结构树进行合并。

定义 2 层次结构树相关度,即 rel(A,B)计算公式为

$$\mathrm{rel}(A,B) = \frac{\sum_{i=0}^{m} \sum_{j=0}^{n} Co(A_i, B_j)}{\sum_{j=0}^{n} \mathrm{Num}(B_j)}$$
$$N_A \geqslant N_{\min}, N_B < N_{\min} \tag{2}$$

其中:A、B 分别代表两个开放分类层次结构树,A_i 为 A 的节点,B_j 为 B 的节点,$Co(A_i,B_j)$为 A_i 与 B_j 共现的次数,$\sum_{i=0}^{m}\sum_{j=0}^{n} Co(A_i,B_j)$为两个树上共现节点的共现次数之和,$\mathrm{Num}(B_j)$为 B_j 出现的次数,$\sum_{j=0}^{n}\mathrm{Num}(B_j)$为树 B 上所有节点出现次数之和,N_A 为树 A 节点数,N_B 为树 B 节点数,$N_{\min}$为节点数阈值。满足相关度阈值的两个层次结构树根节点成为包含关系(A, B),从而使两个树合并为一个树。

3.3 基于相似度和相关度计算的层次结构树聚类算法

层次结构树聚类算法 HTCA(hierarchy tree clustering algorithms)描述如下:

输入:层次结构树集合 $T = \{t_1, t_2, \cdots, t_n\}$,相似度阈值 min_sim,相关度阈值 min_rel。

输出:聚类后的层次结构树集合 $T' = \{t_1', t_2', \cdots, t_m'\}$

a)计算 T 中树的个数 N_t。

b)计算 T 中每个树的节点个数 N_{node}。

c)如果 T 中每个树的节点个数 $N_{\mathrm{node}} \geqslant N_{\min}$或每个树的节点个数 $N_{\mathrm{node}} < N_{\min}$,则结束。

d)若 T 中有树 t,且 t 节点数 $N_{\mathrm{node}} < N_{\min}$, $T_B \leftarrow t$。

e)若 T 中有树 t,且 t 节点数 $N_{\mathrm{node}} \geqslant N_{\min}$, $T_A \leftarrow t$。

f)$M = T_B$ 中树的个数。

g)for each $t_i \in T_B$

for each $t_j \in T_A$

(a)如果 $\mathrm{sim}(t_i, t_j) \geqslant$ min_sim,则将 t_i 和 t_j 合并为 t_j', $T_A \leftarrow t_j'$,将 t_i 和 t_j 从 T_B 和 T_A 中移除。

(b)如果 $\mathrm{rel}(t_i, t_j) \geqslant$ min_rel,则将 t_i 和 t_j 合并为 t_i', $T_A \leftarrow t_i'$,将 t_i 和 t_j 从 T_B 和 T_A 中移除。

end for

end for

h)for each $t_i \in T_B$ and $t_j \in T_A$

$T' \leftarrow t_i, T' \leftarrow t_j$

end for

i)聚类收敛条件判断:$M' = T_B$ 中树的个数;if $M' < M$,则 $M = M'$,转到步骤 f);else $T' = T_A + T_B$。

结束。

4 实验与结果分析

4.1 实验数据集

实验数据取自互动百科。实验数据包含 116 814 个开放分类、783 343 个互动百科条目。为了计算共现概率,本文选取的百科条目均被两个及两个以上的开放分类标注。出现次数最多的开放分类有人物、文化、科学、生活、艺术、地理、文学等,出现次数均超过四万次。开放分类非常稀疏,有 113 566 个开放分类出现次数少于 100 次。部分开放分类出现次数统计如图 4 所示。开放分类的稀疏性不便于检索百科条目,因此,有必要将语义相关的开放分类进行聚类并构建层次结构树,便于用户对百科条目进行分类、检索和浏览。

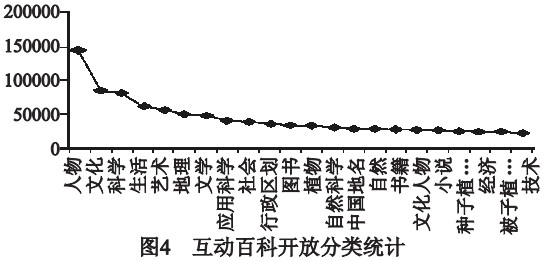

图4 互动百科开放分类统计

4.2 开放分类数据预处理

开放分类是用户按照自己的理解和习惯自由添加的,用词没有统一的标准,导致了大量同义词的出现,如书籍和图书、药物和药品、唐朝和唐代等。因此在聚类分析前,需要将同义词进行替换。本文采用哈工大信息检索研究中心语言技术平台中的《同义词词林扩展版》[10]找出开放分类中的同义词。在具有同义关系的两个开放分类中,用词频高的开放分类替换词频低的开放分类。

在开放分类实验数据集中,具有同义关系的开放分类共有约 4 200 个。例如,电脑和计算机、农村和乡村、年份和年代、餐饮和饮食、法规和法律、气候和气象、卡通和动画、番薯和甘薯等。用高频词替换低频词,被替换的开放分类约有 2 700 个。

4.3 父子对生成实验

基于词共现模型生成的父子对个数随 t 和 $E_{\min}$的变化而不同。图 5 显示当 $E_{\min}$和 t 分别取不同值时生成的父子对个数的变化情况。从图中可看出,给定 $E_{\min}$时,t 越小,父子对越多;给定 t 时,$E_{\min}$越小,父子对越多,并且随着 t 逐渐增大时父子对个数逐渐减少,随着 $E_{\min}$逐渐增大时父子对个数逐渐减少。

在生成的父子对中,有一些是错误的父子对。本文把相反的包含关系、相关但不正确和不相关都视为错误的父子对。例如,在以“人物”为父、以“复旦大学药学院”“文化人物”“高尔夫球”“白羊座名人”等为子组成的父子对中,复旦大学药学院和高尔夫球与人物不相关。

以“河流”为父、以“各国地理”“水系”“中国河流”“中国名水”“著名河流”为子组成的父子对中,“各国地理”与“河流”相关但不正确,“河流”和“水系”的父子关系相反。由于在 $E_{\min}$和 t 取不同值时,生成的父子对个数不同,本文分别对在不同的 $E_{\min}$和 t 下生成的父子对进行随机抽样,计算平均错误率,分析平均错误率随 $E_{\min}$和 t 的变化情况。例如,当 $E_{\min}$分别等于 5、10、20、40 时,$t = 0.6$ 时的平均错误率为 8%,$t = 0.7$ 时的平均错误率为 6.43%。图 6 显示了当 $E_{\min}$分别等于 5、10、20 和 40 时,平均错误率随 t 的变化情况。图 7 显示了当 t 分别等于 0.6、0.7、0.8 和 0.9 时,平均错误率随 $E_{\min}$的变化情况。从图 6 可看出,当 t 在 0.75 和 0.85 之间时,平均错误率较低。从图 7 可看出,当 $E_{\min}$在 5 ~ 8 之间时,平均错误率较低。最高的平均错误率不超过 8%左右,说明词共现方法的准确率较高。

根据汉语重心后移特点使某些不满足共现阈值的开放分类

组成父子对。例如,以“人物”为父,增加了以“各职业人物”“体育领域人物”“红楼梦人物”“文学领域人物”“古代人物”“军事人物”等为子的父子对。根据汉语重心后移特点得到的父子对可以用于修正某些基于共现模型生成的“相反包含关系”的父子对。例如,父子对(空中加油机,加油机)具有相反的包含关系,通过语义分析调整为(加油机,空中加油机)。利用语义分析,新增加的父子对以及改变包含关系的父子对共有约 6 000 个。因此,结合语义分析和词共现分析,能够得到更多的父子对,并能够弥补基于词共现模型方法的不足。

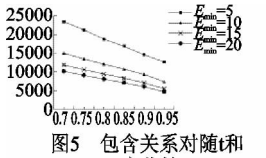

图5　包含关系对随t和E_{min}变化情况

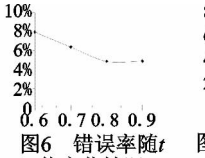

图6　错误率随t值变化情况

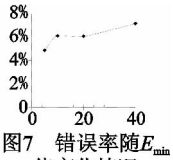

图7　错误率随E_{min}值变化情况

4.4　层次结构树生成与聚类实验

4.4.1　层次结构树生成实验

对父子对进行连接和剪枝,生成开放分类层次结构树和层次结构森林。当 E_{min} 和 t 取不同值时,生成的父子对个数不同,因此层次结构树个数也随之不同。表 1 列出在不同 E_{min}、t 下层次结构树个数统计。从表 1 可看出,当 E_{min} 不变、t 逐渐增大时,层次结构树个数先增加后减少。其原因是当 t 增大时,父子对个数会减少,导致某些父子对不能连接起来,因而层次结构树个数会有所增加;然而,随着父子对个数的不断减少,层次结构树个数最终仍会减少。

表 1　层次结构树统计

t	$E_{min}=5$	$E_{min}=10$	$E_{min}=15$	$E_{min}=20$
0.7	1 356	933	792	697
0.75	1 421	959	824	735
0.8	1 370	971	840	752
0.85	1 360	958	844	765
0.9	1 303	920	804	741

图 8 是以“人物”为根的层次结构树部分内容。图 9 是以“文化”为根的层次结构树部分内容。

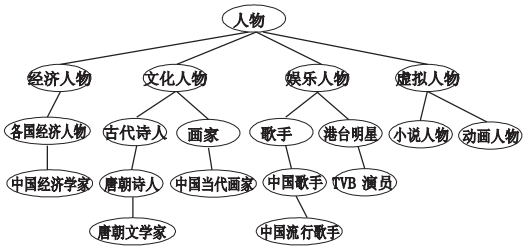

图8　人物层次结构树

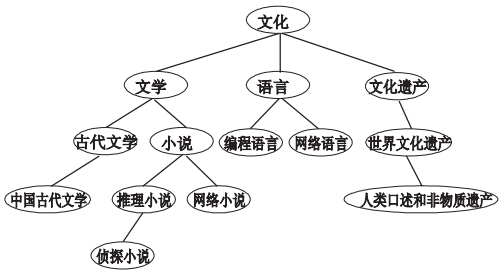

图9　文化层次结构树

4.4.2　层次结构树聚类实验

层次结构树个数较多,不便于百科条目的检索。例如表 1 中,当 $E_{min}=10$、$t=0.85$ 时,层次结构树个数为 958。本文采用 HTCA 算法将层次结构树进行聚类。层次结构树聚类将很多孤立的层次结构树合并。如“书法家”“艺术家”“学者”“中国人”“网络红人”“模特”“物理学家”等层次结构树并入了“人物”层次结构树;“河流”“海洋”“岛屿”“湿地”“地震”“西藏”“山脉”等层次结构树并入了“地理”层次结构树;“昆虫”“鸟纲”“宠物狗”“鸠鸽科”“猫”“软体动物门”“龟科”等层次结构树并入“动物”层次结构树。聚类后的层次结构树个数统计如表 2 所示。从表 2 看出,聚类后的层次结构树个数明显减少。

表 2　聚类后的层次结构树统计

$sim(A,B)$	N_{min}												
$rel(A,B)$	10	20	30	40	50	60	70	80	90	100	500	900	1000
0.7	250	234	211	204	207	201	185	184	267	338	836	836	958
0.75	276	253	225	216	234	251	216	213	294	405	846	846	958
0.8	326	304	288	275	302	316	268	332	396	504	855	855	958
0.85	354	332	318	302	330	338	295	356	417	515	858	858	958
0.9	400	376	373	364	391	398	385	414	472	541	860	860	958

聚合后的层次结构树个数随着相似度、相关度阈值以及 N_{min} 的变化而变化。相似度和相关度阈值越高,层次结构树的个数越多。聚合后的层次结构树个数与 N_{min} 之间的变化情况是当 N_{min} 从 10 逐渐增大至 40 时,层次结构树的个数逐渐减少,说明随着 N_{min} 的增大,有更多的层次结构树并入其他的层次结构树;当 $N_{min}=50$ 时,层次结构树个数却有所增加,说明节点数少于 40 的层次结构树与节点数在 40 ~ 50 之间的层次结构树进行了聚合;当 $N_{min}=70$ 时,层次结构树个数有所减少,说明节点数在 50 ~ 70 之间的层次结构树与节点数大于 70 的层次结构树进行了聚合;随着 N_{min} 的不断增大,层次结构树的个数也不断在增加,当 N_{min} 增加到某个值,层次结构树的个数变化很小,如 $N_{min}=500,900$,说明节点数在此区间的层次结构树个数非常少;当 N_{min} 超过树最大节点个数时,如超过 1 000,层次结构树没有聚合,个数不变,仍为 958。

对层次结构树聚类结果进行观察,发现某些没有关系的层次结构树被聚合了起来。如“吉祥物”“漫画”“赛事”等层次结构树并入了“人物”层次结构树。此外,某些层次结构树合并后造成层次关系的错误,如“生物”层次结构树并入了“种子植物门”层次结构树,“传统节日”层次结构树并入了“春节”层次结构树。本文对聚合后的层次结构树进行了准确率分析与统计。从表 3 可看出,准确率随着相似度与相关度阈值的增加而增加,随着最少节点数的增加而降低,平均准确率均在 80% 以上。其中,当 $sim(A,B)=0.9$、$rel(A,B)=0.9$、$N_{min}=10$ 时,准确率达到 93.4%。

表 3　层次结构树聚类准确率分析

$sim(A,B)$	N_{min}										平均准
$rel(A,B)$	10	20	30	40	50	60	70	80	90	100	确率/%
0.7%	89.1%	88.7%	88.5%	88.3%	87.3%	87.2%	85.1%	72.3%	69.4%	64.2%	82
0.75%	90.4%	89.7%	88.8%	88.6%	87.8%	87.3%	86.7%	78.8%	76.5%	73%	84.8
0.8%	90.9%	89.8%	89.1%	88.7%	87.8%	87.6%	87.5%	86.6%	85.5%	84.5%	87.8
0.85%	91.4%	90.3%	89.7%	89%	88.2%	88.1%	88.1%	87.1%	85.8%	85.1%	88.3
0.9%	93.4%	92.5%	92.2%	90.4%	90.4%	90.3%	90.2%	89.7%	88.1%	87%	90.4

4.5　开放分类层次结构树检索百科条目

出现次数最多的开放分类依次为人物、文化、科学、生活、艺术、地理(图 4)。在聚类前利用开放分类或聚类后利用层次结构树检索百科条目个数对比如表 4 所示。从表 4 中可看出,聚类后利用“人物”层次结构树检索到的条目数比聚类前增加了约 2 倍;利用“地理”层次结构树检索到的条目数比聚类前增加了约 4 倍。聚类后检索条目的数量明显增加,使条目检索效率得到提高。

(下转第 1674 页)

五级与六级等级之间。

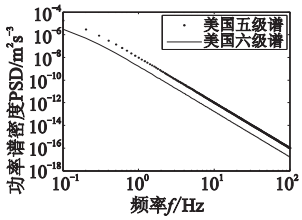

图3 仿真用美国AAR轨道谱

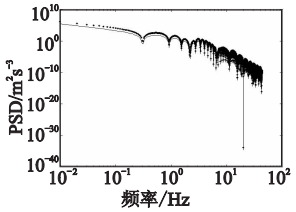

图4 不同级别轨道谱作用下车体垂向加速度功率谱密度

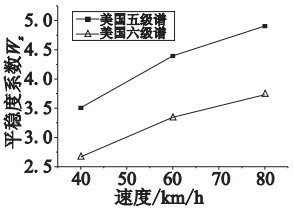

图5 不同轨道谱和车速下平稳性指标变化

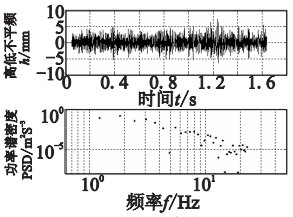

图6 实测地铁高低不平顺及其功率谱

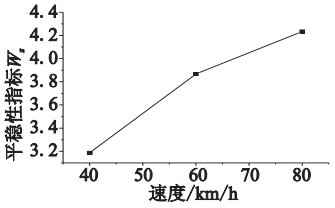

图7 实测轨道谱作用下平稳性指标

6 结束语

本文分析了目前轨道评价系统的不足,针对地铁线路的特点提出了基于车辆响应仿真计算的轨道不平顺状态评价方法,引入随机振动虚拟激励法计算了标准美国 AAR 轨道谱和地铁轨检车实测轨道谱作用下的平稳性指标,详细介绍了计算方法和过程,结果对比分析可知实测路段的整体平顺状态介于美国 AAR 五级与六级之间。该评价方法对于轨道的养护计划制定具有一定的参考价值。

参考文献:

[1] 刘金朝,刘秀波. 轨道质量状态评价方法[J]. 铁路技术创新,2012(1):38.

[2] BERGGREN E G, LI M X D, SPÄNNAR J. A new approach to the analysis and presentation of vertical track geometry quality and rail roughness[J]. Wear,2008,265(9-10):1488-1496.

[3] LI M X D, BERGGREN E G, BERG M. Assessment of vertical track geometry quality based on simulations of dynamic track-vehicle interaction [J]. Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit,2009,223(2):131-139.

[4] LI M X D, BERGGREN E G, BERG M, *et al*. Assessing track geometry quality based on wavelength spectra and track-vehicle dynamic interaction[J]. Vehicle System Dynamics,2008,46(1):261-276.

[5] LUBER B, HAIGERMOSER A, GRABNER G. Track geometry evaluation method based on vehicle response prediction[J]. Vehicle System Dynamics,2010,48(1):157-173.

[6] LUBER B. Railway track quality assessment method based on vehicle system identification[J]. Elektrotechnik & Informationstechnik, 2009,126(5):180-185.

[7] 陈宪麦,杨凤春,吴旺青,等. 秦沈客运专线轨道谱评判方法的研究[J]. 铁道学报,2006,28(4):84-88.

[8] 林家浩,张亚辉. 随机振动的虚拟激励法[M]. 北京:科学出版社,2004.

[9] 王振来. 轨道车辆平稳性频域建模与仿真[D]. 长春:吉林大学,2007.

[10] 单涛涛,楼梦麟,贾宝印. 关于上海地铁一号线轨道高低不平顺问题的探讨[J]. 振动与冲击,2012,31(12):53-58.

(上接第 1663 页)

表 4 聚类前后检索条目数量对比

开放分类/层次结构树	聚类前检索条目	聚类后检索条目	开放分类/层次结构树	聚类前检索条目	聚类后检索条目
人物	143 075	414 850	生活	61 078	177 597
文化	85 548	291 211	艺术	56 058	188 929
科学	80 008	374 309	地理	48 823	231 664

5 结束语

本文提出的聚类算法有效实现了对互动百科开放分类的聚类和重新组织。生成的层次结构树准确率较高,提高了对百科条目进行分类、检索的效率。接下来,笔者将对下列内容作进一步研究。

a)由于总是小树(节点数目少)并入大树(节点数目多),造成某些层次关系的错误。如上述“生物”和“种子植物门”层次结构树之间、“传统节日”和“春节”层次结构树之间的层次关系错误。今后的一个主要工作是研究更加灵活的聚类算法,进一步提高聚类准确率。

b)当前仅能对层次结构树进行聚类,下一步研究层次结构树的子树之间的聚类算法,实现层次结构树子树的聚合,使层次结构树的层次关系在语义上更加准确。

c)利用知网 HowNet 等外部语义资源对聚类结果进一步优化。例如,提取开放分类词汇的语义概念,根据语义概念将语义相关的开放分类进行聚合。

致谢 在此,向对本文工作给予支持和建议的同行,尤其是西南交通大学计算机科学与技术系杨燕教授、戴齐副教授,以及云计算与智能技术实验室的老师和同学们表示感谢。

参考文献:

[1] ANDRES S, OSCAR C, HARITH A, *et al*. Review of the state of the art: discovering and associating semantics to tags in folksonomies[J]. The Knowledge Engineering Review,2012,27(1):57-85.

[2] BEGELMAN G, KELLER P, SMADJA F. Automated tag clustering: Improving search and exploration in the tag space[C]//Proc of the 15th International World Wide Web Conference. 2006.

[3] HEYMANN P, HECTOR G M. Collaborative creation of communal hierarchical taxonomies in social tagging systems[R]. [S.l.]: Stanford InfoLab,2006.

[4] CAO Gao-hui, JIAO Yu-ying, CHENG Quan. Research on tag cluster based on hierarchical agglomerative clustering algorithm[J]. New Technology of Library and Information Service,2008,163(4):23-27.

[5] SANDERSON M, CROFT B. Deriving concept hierarchies from text [C]//Proc of the 22nd ACM Conference of the Special Interest Group in Information Retrieval. 1999: 206-213.

[6] SCHMITZ P. Inducing ontology from flickr tags[C]//Proc of Collaborative Web Tagging Workshop at WWW2006. 2006.

[7] PONZETTO S P, STRUBE M. Taxonomy induction based on a collaboratively built knowledge repository[J]. Artificial Intelligence, 2011,175(9-10):1737-1756.

[8] 宋明亮. 汉语词汇字面相似度性原理与后控制词表动态维护研究[J]. 情报学报,1996,15(4):261-271.

[9] SUN Ji-gui, LIU Jie, ZHAO Lian-yu. Clustering algorithms research [J]. Journal of Software,2008,19(1):48-61.

[10] CHE Wan-xiang, LI Zheng-hua, LIU Ting. LTP: a Chinese language technology platform[C]//Proc of COLING. 2010: 13-16.