移动用户兴趣点标记语义映射方法①

崔立伟, 张晓滨

(西安工程大学 计算机科学学院, 西安 710048)

摘 要: 针对当前移动用户兴趣点标注没有统一的语义本体这一问题, 通过对移动用户兴趣点分类, 建立了 POI 概念本体层次树, 将用户标注的 POI 信息与本体树节点通过一种改进的映射方法建立映射关系, 为移动用户标注信息提供统一的规范化语义. 实验收集多名志愿者 5 周的真实标注信息来评价该方法, 实验结果表明该方法具有较高的准确率.

关键词: POI;本体层次树;映射

Semantic Mapping Method on the Mobile Users' Point of Interest

CUI Li-Wei, ZHANG Xiao-Bin

(College of Computer Science, Xi'an Polytechnic University, Xi'an 710048, China)

Abstract: In view of the problem that marked the POI (Point of interest) of the current mobile users with no unified semantic ontology,this paper categorized POI and established POI ontology hierarchy tree for mobile users, then finished the mapping between the labeled POI information of users and the node of ontology tree by an improved mapping method to provided a unified standardized semantics for mobile users to mark information. An experiment is performed to evaluate the performance of the new similarity measure by using the labeled information of users in a period of 5 weeks, the results show that the method has high accuracy.

Key words: POI; ontology hierarchy tree; mapping

随着智能手机的日渐普及, 越来越多的移动用户利用手机移动应用标注自己的 POI 信息.

POI 是重要的信息源, POI 数据的空间定位精度、属性的丰富程度及表达的清晰程度直接影响着移动地图的可用性以及用户分享 POI 信息的准确性. 张玲[1]通过研究地理信息, 总结了 POI 分类的原则和方法; 李瑞珊通过对自然语言的处理, 提出了多源 POI 融合方法并设计了数据融合系统, 构建了 POI 分类模型[2]. 但是 POI 信息中存在着大量的冗余, 特别是在用户标注 POI 信息时, 尚无统一的语义规范. 因此需要将用户 POI 信息分类映射, 寻找规范的 POI 语义, 提高数据的准确性. 因此, 本文通过建立移动用户 POI 概念本体层次树, 对用户兴趣点分类归纳, 并在此基础上提出一种移动用户兴趣点标注映射方法, 将用户标注的 POI 信息与本体树节点建立映射关系, 发现 POI 语义信息之间的相似关系, 提高 POI 标注信息的准确性.

1 移动用户POI标记语义映射

1.1 POI 分类概念本体层次树的建立

本文按照 POI 分类原则并结合移动用户的行为特点, 采用层次结构分类方式将 POI 分为一级类、二级类、三级类三类, 其中一级类的分类中, 以用户日常生活较为密切兴趣点为基础分为自然人文地理、餐饮、教育培训、购物、医疗卫生、文化休闲娱乐、日常服务 7 个大类, 在一级类基础又上划分了 34 个二级类, 115 个三级类.

对分类的 POI 信息, 采用十进制编码方式, 通过[一级类][二级类][三级类]表示. 如图 1 所示:

POI 分类具有层次性, 根据 POI 层次分类得到 POI 概念本体层次树片段如图 2 所示:

① 基金项目:教育部春晖计划(Z2009-1-71001)

收稿时间:2014-08-09;收到修改稿时间:2014-09-13

图 1　POI 分类代码

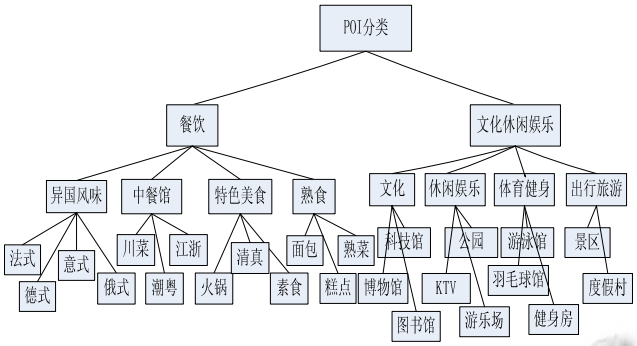

图 2　POI 概念本体层次树片段

1.2 移动用户 POI 标注信息映射

移动用户在终端上通过各种应用来标注 POI 信息, 例如百度地图、微博、人人网客户端等. 这些应用采用开放式的形式, 通过用户对自己感兴趣的 POI 标注, 将信息发布到服务器上, 与其它用户共享. 但这就将用户这个庞大的群体纳入到了数据生产者行列, 产生大量信息的同时带来了冗余信息甚至是错误信息, 因此需要有统一语义对其规范化.

1.2.1 POI 信息预处理

在建立映射关系前, 由于移动地图的标注数据在精度、重复和格式等方面存在问题, 所以必须先对信息缺失点、异形同义点以及冷地点数据进行处理, 以提高数据精度, 减少重复.

1) 信息缺失处理主要处理具有一定关联的数据, 建立一个标准格式来筛选出有效的数据, 采用“店名+位置”的方式,将其区分开来. 例如用户在开元路标注为“肯德基”的信息存在信息缺失, 处理后 POI 格式应为“肯德基(开元路店)”.

2) 对于同一个 POI 实体, 有时会有多种不同的称呼方式, 其中包括实体的标准名称、俗称等. 例如“西安工程大学”这一条信息, 标准名称为“西安工程大学”, 而同时又存在“纺院”、“西工科”等别称. 因此, 对于这类语义异构问题, 将同一地理目标的不同别名、俗称与标准名称进行合并[3].

3) 冷地点由于到访人数很少而不会成为大众所关注的 POI, 因此须将其剔除. 对于这类问题, 可以设定用户到访数量的阈值, 将低于该阈值的 POI 剔除.

1.2.2 POI 标注语义映射方法

预处理后的 POI 信息包含了 POI 的位置及名称信息, 但缺乏语义上的特征, 没有语义关联, 属于自然语言表达的字符串. 因此需要将自然语言表达的 POI 信息通过映射关系[4]来完成 POI 语义的规范化, 实现由自然语义到标准语义的映射, 即:

$$F_{Map}:(POI_{标},O_{SS})\rightarrow(POI_{预},O_{GS})$$

其中 $POI_{标}$ 表示用户标注 POI 信息, $POI_{预}$ 表示预处理 POI 信息, O_{SS} 源本体语义即自然语义, O_{GS} 表示目标本体语义即标准语义.

1) 基本方法　为了建立语义映射关系, 先在 WordNet 语料库中考虑语义距离和层次深度两方面因素, 通过设定计算结果的阈值, 得到用户标注 POI 信息相似语义集合 S.

定义 1(语义距离). 针对概念本体层次树, 连接两个概念节点的通路中最短路径所跨的边数我们称之为语义距离[5]. 两个概念的语义距离越小, 其相似度越高; 而处在离根较远的概念间的相似度要比离根近的概念间的相似度要大. 采用边的权重来计算语义距离, 即连接两个概念节点的最短路径所跨的边的权重之和. 假设概念 $S_1,S_2.....S_n$ 为概念 S_1 和 S_2 之间的最短路径, 则概念之间的语义距离为:

$$Distence(S_1,S_2)=\sum_{i=1}^{n-1}W(S_i)\qquad n\neq 1 \tag{1}$$

其中 $W(S_i)$ 表示从概念 S_i 引出的边具有相同的权重, 用概念 S_i 的权重表示, 即连接 S_1,S_2 最短路径上第 i 条边的权值. 对于权重的计算, 利用公式(2):

$$W(S)=\frac{1}{Wid(S)} \tag{2}$$

其中 $Wid(S)$ 为概念 S 的直接孩子节点数目.

定义 2(层次深度). 对于所建立的概念本体层次树, 每层子概念都是对上层父概念的不断细化, 越靠近叶子节点, 其概念的差异程度越小. 在本体层次树中, 两个概念具有相同的子概念, 则这两个概念可能是相似的; 两个概念具有相同的兄弟, 则这两个概念

可能是相似的. 因此对层次深度做归一化处理, 层次深度影响因素 N 由公式(3)决定:

$$N(S_1,S_2)=\frac{\alpha-1}{\alpha}*\frac{|Dep(S_1)-Dep(S_2)|+1}{Dep(S_1)+Dep(S_2)} \tag{3}$$

其中 $Dep(S)$ 为概念 S 深度，α 为调整参数.

综合考虑两方面因素影响，得出 POI 语义相似度计算公式如公式(4):

$$sim(S_1,S_2)=1-\sqrt{N*Distence} \tag{4}$$

在计算结果中选取高于某一阈值的部分得到 POI 相似语义集合S.

2) 改进方法 在得到相似语义集合 S 后，引入 Jaccard 相似方法，Levenshtein Distance 方法，The Jaro-Winkler Distance 匹配算法，通过上述 3 种方法以加权的方式计算得出语义集合 S 中与本体树节点具有较高相似性的语义点.

定义 3(Jaccard 相似方法[6]). 方法描述为: 两个词组对中相同词(无重复)的个数与所有词(无重复)个数的比值，即对于 S_1,S_2，两者的 Jaccard 相似度可定义为:

$$Jaccard(S_1,S_2)=\left|\frac{S_1\cap S_2}{S_1\cup S_2}\right| \tag{5}$$

定义 4(Levenshtein Distance 算法[7]). Levenshtein Distance 被定义为将词组 s 变换为词组 t 所需的删除、插入、替换操作的次数. Levenshtein Distance 算法表示为:

$$LD(S_1,S_2)=1-\frac{dist}{\max len(S_1,S_2)} \tag{6}$$

其中 $dist$ 是字符串 S_1,S_2 的莱文距离，LD 计算结果越大，相似度越大.

定义 5 (The Jaro-Winkler Distance 匹配算法[8]) The Jaro–Winkler Distance 适合于较短的字符之间计算相似度. 算法得分公式为:

$$d_w(S_1,S_2)=d_j+L*P(1-d_j) \tag{7}$$

其中 d_j 为 Jaro Distance 公式:

$$d_j(S_1,S_2)=\frac{1}{3}*\left(\frac{m}{S_1}+\frac{m}{S_2}+\frac{m-t}{m}\right) \tag{8}$$

Match Window(匹配窗口)计算公式:

$$MW=\frac{Max(|S_1|,|S_2|)}{2}-1 \tag{9}$$

其中 m 是匹配的字符数, t 是换位的数目. 公式(7)定义了一个范围因子常量 P，如果前缀部分有长度为 L 的部分字符串相同，就用常量 P 来调整前缀匹配的权值，标准默认设置值 P=0.1.

综上所述，由于 Jaccard 相似法未考虑相同词出现的位置以及顺序，因此在 Jaccard 算法的基础上引入莱文斯坦算法和 The Jaro-Winkler distance 匹配算法修正相同词的位置以及顺序因素的影响. 得到改进算法如公式(10)所示:

$$Sim(S_1,S_2)=\alpha Jaccard+\beta SLD+\gamma d_w \tag{10}$$

其中，$\alpha+\beta+\gamma=1, Jaccard\in[0,1], SLD\in[0,1], d_w\in[0,1]$.

通过改进方法，能够更准确找出集合S中不同POI点的语义相似性，从而使 WordNet 中语义相似度的计算更准确.

1.2.3 实验分析

为了验证计算方法的有效性，本文利用街旁网用户标注数据集，选取其中 10 位用户的标注信息,先对训练集数据预处理，之后利用本文所提出的方法计算处理后的信息与本体树节点相似性.

1)取阈值高于 0.6 部分得到相似语义集合 S,在 S 中取 α=0.3，β=0.3，γ=0.4，对于不同标注 POI 点利用不同方法计算该 POI 点与本体树节点映射结果如图 3 到图 6 所示.

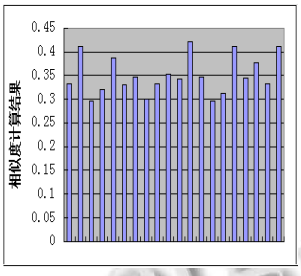

图 3 Jaccard 方法

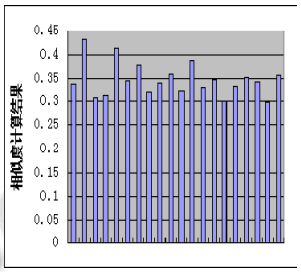

图 4 Levenshtein 方法

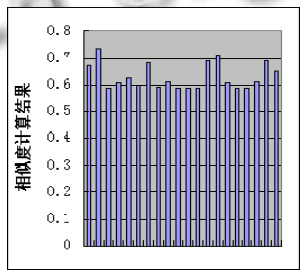

图 5 The Jaro 方法

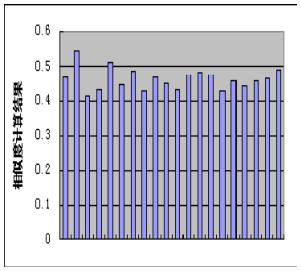

图 6 本文改进方法

Jaccard 方法和 Levenshtein Distance 方法由于考虑相同词长度因素的影响，计算结果偏低，而 The Jaro-Winkler Distance 方法相比其他两个算法利用匹配 Jaccard 方法和 Levenshtein Distance 方法由于考虑相同词长度因素的影响，计算结果偏低，而 The Jaro-Winkler Distance 方法相比其他两个算法利用匹配窗口

控制相似度匹配, 减少了相同词本身因素对相似度影响.

2) 取 POI 点“西安工程大学图书馆”计算该 POI 点与本体树节点相似度结果如图 7 所示:

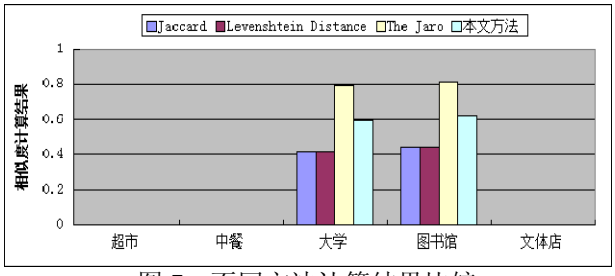

图 7 不同方法计算结果比较

“西安工程大学图书馆”映射结点可以是“大学”和“图书馆”. 利用本文提供的改进方法进行语义相似度计算, 考虑两者与原 POI 实例的相似词匹配得出“西安工程大学图书馆”这一 POI 实例与“图书馆”相似度更高, 同时也与用户的标注习惯相同.

3) 再以“西安中医药研究院”为例, 利用 Jaccard 方法计算其与本体树结点的相似度, 结果如图 8 所示. 计算结果表明“西安中医药研究院”与本体树结点“中医院”和“研究院”相似度较高且相似度相同, 而“中医院”这一映射点信息偏差较大, 不能够用于信息标注, 这是由于 Jaccard 方法只考虑了相同字词匹配的影响. 引入 Levenshtein Distance 方法和 The Jaro-Winkler Distance 方法, 利用相似字词变换操作和匹配窗口的作用, 修正相同词位置和顺序的影响, 得出“研究院”的相似度更高, 计算结果如图 9 所示.

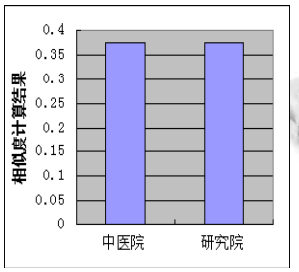

图 8 Jaccard 方法结果

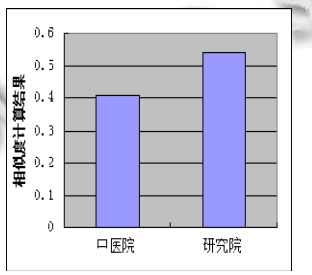

图 9 改进方法结果

综上所述改进后的方法将三种方法结合, 弥补各自的不足, 提高了语义相似度的精度, 为用户标注信息提供更可靠地依据.

2 总结

针对当前移动用户兴趣点标注没有统一的语义本体, 本文从 POI 分类方面构建了概念本体层次模型, 将用户标注语义与本体树节点通过 WordNet 语义计算得到相似语义集, 再在语义集合中利用 Jaccard 相似算法、Levenshtein Distance 算法以及 The Jaro-Winkler Distance 三者加权的匹配算法弥补各算法不足来建立映射关系, 得到映射关系对, 提高了映射的准确性. 在获取兴趣点映射之后, 为兴趣点标注提供了统一的规范语义, 提高人们分享信息的有效性, 为移动用户兴趣点推荐提供参考.

参考文献

1 张玲.POI 的分类标准研究.测绘通报,2012,(10):82–84.

2 李瑞珊.基于自然语言处理的多源 POI 数据融合的研究[学位论文].青岛:中国海洋大学,2013.

3 Ramprasath M, Hariharan S. Using ontology for measuring semantic similarity for question answering system. 2012 IEEE International Conference on Advanced Communication Control and Computing Technologies (ICACCCT). 2012. 218–223.

4 He W, Yang XP, Huang DP. WNOntoSim: A hybrid approach for measuring semantic similarity between ontologies based on WordNet. 2011 Eighth Web Information Systems and Applications Conference (WISA). 2011. 73–77.

5 Lavanya S, Arya SS. An approach for measuring semantic similarity between words using SVM and LS-SVM. 2012 International Conference On Computer Communication and Informatics (ICCCI). 2012. 1–4.

6 林学民,王炜.集合和字符串的相似度查询.计算机学报, 2011,34(10):1853–1862.

7 Chowdhury SD, Bhattacharya U, Parui SK. Online handwriting recognition using levenshtein distance metric. 2013 International Conference on Document Analysis and Recognition (ICDAR). 2013. 79–83.

8 Celik D, Elci A. A broker-based semantic agent for discovering Semantic Web services through process similarity matching and equivalence considering quality of service. Science China (Information Sciences), 2013, 56(012102): 1–24.