XXXXXX(XXXXX) 第52卷 第6期 2016年11月

Acta Scientiarum Naturalium Universitatis Pekinensis, Vol. 52, No. 6 (Nov. 2016)

doi: 10.13209/j.0479-8023.2015.136

即时通信文本中地理信息提取  
—— 以微信为例

XXX1,2 XX1,2,† XXX1,2 XXX1,2

1. XXXXXXXXXXXXXXXX, 北京100871; 2. XXXXXXXXXXXXXXXXXXXXX, 北京 100871;   
† 通信作者, E-mail: XXX@pku.edu.cn

摘要提出一套面向即时通信文本中地理信息提取的技术方案, XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX, 可为相关研究和实践提供有力支持。

关键词XXXXXX; WebGIS/LBS; XXXXXX; XXXXXXXXXX; XXXXXX

中图分类号P208

Geographical Information Extraction from Instant Communication   
Messages: A Case Study of WeChat

XXX,2, XX1,2,†, XXX1,2, XXX1,2

1. Institute of Remote Sensing and Geographical Information System, Peking University, Beijing 100871;   
2. Beijing Key Laboratory of Spatial Information Integration and Its Applications, Beijing 100871;   
† Corresponding author, E-mail: tianyuanpku@pku.edu.cn

**Abstract** In order to provide synchronous map service based on message semantics in instant communication software, XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX. The synchronous message semantics-based image-text integrated service provided by the case study improves the user experience very well.

**Key words** instant communication message; WebGIS/LBS; Chinese text segmentation; spatial analysis pattern recognition; image-text integrated service

本世纪以来, XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXX。 XXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX, 可以大大提升交流信息的直观程度以及用户的交流感受和决策效率。

基于此, XXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXX, 基于实际的即时通信文本样本开展实例验证工作。

1 即时通信文本中地理信息提取总体方案

针对上述现存XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX, 为即时通信用户提供即时的、一体化的空间信息服务。

即时通信文XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX, 以专题地图的形式提供合适的餐厅信息。

本文提出的XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX分析

154w105.tif

图**1**  总体技术方案

Fig. 1 Overall technical solution

需XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX用户。

2 关键技术

本文提出XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX, 并以图文一体的方式将结果呈现给即时通信用户。

2.1 中文文本分词

中文文本分词指将一个汉字序列切分成单独的词, XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX, 尹锋[10]和何嘉等[11]分别以BP算法为基础提出改进算法。

提取即时通信文本中的地理信息时, 中文文本分XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX基础。

目前, XXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX, 并结合即时通信文本的具体特点对其进行调整和完善。

2.2 查询分析模式识别

查询XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX 展开。

表**1**  文本语义空间分析模式识别(据郭仁忠[16]扩展)

Table 1 Spatial analysis pattern recognition of text semantics (after Guo[16])

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 空间分析类别 | 含义 | 特点 | 文本语法关键词 | 形式化描述 | GIS操作 | 示例 |
| 空间位置 | 个体定位信息 | 定位信息 | 地理实体+“的”+地理实体 | Noun of Noun | Location | 颐和园东门  北京大学的未名湖 |
| 空间分布 | 同类空间事物  群体定位信息 | 分布中心/标准  距离/分布密度 | 地理实体+“沿线”+地理实体 | Distribution of Noun | Buffer on distance | 北大附近的车站 |
| 空间形态 | 空间物体几何  特征 | 走向/面积/周  长/曲面坡度 | 地理实体+“的”+“面积/周长/长度”+地理实体 | Noun Length/Area | Attribute for entity | 颐和园走一圈要多久 |
| 空间距离 | 几何上的接近  程度 | 欧氏距离 | 地理实体+“附近的”+地理实体/属性信息 | Near Noun | Distance of entities | 华联在北大东门500 m远 |
| 空间相关 | 空间方位 | 方位信息 | 地理实体+“东/南/西/北”+地理实体 | Noun Direction | Orientation for entity | 中关村/南边/的/电影院 |
| 空间拓扑 | 空间拓扑特征 | 地理实体+“内/外/里/  相邻”+地理实体 | Inside/Outside /Near  Noun | Overlay | 北大/里的/教学楼 |
| 相似相关 | 实体之间相似  相关关系 | 地理实体+“一样/相似的”+地理实体 | Noun Verb | Same attribute with entity | 和全聚德相似的烤鸭店 |

根据郭仁忠[16]提XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXX用户交互确认。

2.3 图文一体服务

图文一XXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXX, 大大提升用户交流体验。

目前, XXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX, 在既有分析结果的基础上执行更复杂或深入的查询分析操作。

3 实例验证

为了验证本文提出的即时通信文本地理信息提取技术方案的正确性、XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX即时通信环境下提供图文一体服务的效果。

首先进行中文分词, XXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXX, 在系统界面中实现图文同步服务。

实验中采用带有XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX准实时同步。

154w106.tif

图**2** 实例验证典型用例

Fig. 2 Typical verification cases

我们选取XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX、查询和决策。

图 2(a)中, XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX, 返回的图片显示北京大学校园内食堂的具体位置。

图 2(b)中, XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX加, 得到北京大学附近的车站信息, 以图片形式返回用户查看。

图 2(c)中, XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX, 并返回其面积信息。

图 2(d)中, XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX北京大学东门约500 m远的华联商厦。

图 2(e)中, XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX”、“里”和“教学楼”, 地图查询返回北大内部的教学楼信息。

图 2(f)中, XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX的路径和乘车信息。

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX求。

4 结语

针对当前移动终端即时通信与地图服务软件相互隔绝的问题, XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX, 可以为移动用户提供更好的应用体验, 这也是我们下一步的研究方向。

参考文献

1. 朱和平. 即时通信研究综述. 现代计算机: 专业版, 2006(12): 55–58
2. 李德仁. 论地球空间信息技术与通信技术的集成. 武汉大学学报: 信息科学版, 2001, 26(1): 1–7
3. Fritz J M. Provides intelligence in web-based tutors // North American Web Developers Conference. Frederiction, 1998: 10
4. 霍艳艳, 沈靖瑞. 即时通信软件的发展及现状研究. 河南科技, 2014(1): 8
5. 毛昕影. 基于GIS的智能手机旅游信息服务系统的研究与实现[D]. 成都: 电子科技大学, 2012
6. 刘涌泉. 再读词的问题. 中文信息学报, 1988, 2(2): 47–50
7. Guo J. Critical tokenization and its properties. Computational Linguistics, 1997, 23(4): 569–596
8. 刘源, 梁南元. 汉语处理的基础工程: 现代汉语词频统计. 中文信息学报, 1986, 1(1): 17–25
9. 李家福, 张亚非. 基于EM算法的汉语自动分词方法. 情报学报, 2002, 21(3): 269–272
10. 尹锋. 基于神经网络的汉语自动分词系统的设计与分析. 情报学报, 1998, 17(1): 41–50
11. 何嘉, 陈琳. 基于神经网络汉语分词模型的优化. 成都信息工程学院学报, 2006, 21(6): 812–815
12. Mark M D, Comas D, Egenhofer M J, et al. Evaluating and refining computational models of spatial relations through cross-linguistic human-subjects testing // Frank A U, Kuhn W. Spatial information theory: a theoretical basis for GIS. Berlin: Springer-Verlag, 1995: 553–568
13. 杜世宏, 王桥, 李治江. GIS中自然语言空间关系定义. 武汉大学学报: 信息科学版, 2005, 30(6): 533–538
14. 朱少楠, 张雪英, 张春菊. 地理空间关系描述的句法模式识别 // Proceedings of 2010 International Conference on Broadcast Technology and Multimedia Communication. Hong Kong, 2010: 355–357
15. Unwin D J. Introductory spatial analysis. London: Methuen, 1981
16. 郭仁忠. 空间分析. 武汉: 武汉测绘科技大学出版社, 1997
17. 王丹. 基于 Web 2.0 的信息服务研究[D]. 武汉: 华中师范大学, 2007
18. 李艳, 高扬. 基于地图API的Web地图服务及应用研究. 地理信息世界, 2010, 8(2): 54–57