北京大学学报(自然科学版) 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

张瑞洁[1,2] 田原[1,2,†] 刘思叶[1,2] 王雯夫[1,2]

1. 北京大学遥感与地理信息系统研究所, 北京 100871; 2. 空间信息集成与3S工程应用北京市重点实验室, 北京 100871;
† 通信作者, E-mail: tianyuanpku@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, 综合使用文本分词、空间分析匹配与图文一体服务等技术, 实现即时文本中地理信息的分析获取和同步地图服务, 可以在即时通信交流中提供同步的空间信息分析和主动的网络地图服务。以微信为例, 对上述技术方案进行实例验证。验证结果表明, 所提的技术方案正确、合理、可行。研究成果拓宽了 WebGIS/LBS 的应用领域, 增强了即时通信软件的服务能力, 可为相关研究和实践提供有力支持。

关键词 即时通信软件; WebGIS/LBS; 中文文本分词; 空间信息服务模式匹配; 图文一体服务
中图分类号 P208

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

ZHANG Ruijie[1,2], TIAN Yuan[1,2,†], LIU Siye[1,2], WANG Wenfu[1,2]

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, this paper proposes a technical solution, basically a comprehensive combination of Chinese text segmentation, pattern recognition, and image-text integrated service. A case study based on actual WeChat communication messages is carried out to verify the technical solution, which shows that the proposed solution is both feasible and practically effective. 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, 基于移动设备的应用迅速普及, 其中即时通信应用获得迅猛发展[1]。即时通信应用指通过互联网即时发送和接收消息的应用软件, 为人们日常交流提供方便快捷的工具, 如 QQ、微信和飞信等。即时通信中产生大量与空间相关的信息, 例如文本中出现的地名信息以及空间关系信息。在信息交互过程中, 用户的空间行为规划对这些空间信息存在强烈的依赖, 例如选择出行路线、方式及时间等[2]。目前, 运行在移动端的 WebGIS/LBS 软件已经普及, 可以为用户提供高精度的、在线的空间行为规划服务[3]。但是, 当前主流的基于移动设备的即时通信软件与 WebGIS/LBS 软件相互独立, 即时通信软件无法直接理解即时通信文本中的地理信息, 需要用户进行理解分析后, 再跳转

国家自然科学基金(41271385)资助
收稿日期: 2015–05–15; 修回日期: 2015–06–30; 网络出版日期: 2016–04–07

到移动 WebGIS/LBS 应用，将空间信息和服务需求重新输入，才能获得相符的空间信息服务。人工跳转和信息转录过程非常繁琐，当存在复杂地名和空间分析需求时，容易出现操作和录入失误，大大降低了即时通信用户的交流效率[4−5]。如果能够直接提取即时通信文本中的地理信息，并将同时运行在移动端的即时通信软件与 WebGIS/LBS 服务相结合，使用户在进行常规信息交流的同时获得同步的空间信息服务，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，本文针对即时通信文本中地理信息提取技术展开研究，以期提供一套切实可行的技术方案，使用户在即时通信中享受到同步的空间信息显示及查询分析服务。为验证相关技术方案的合理性、可行性和服务效率，选取微信和百度地图作为即时通信和 WebGIS/LBS 服务平台，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

针对上述现存问题和应用需求，本文提出一套即时通信文本中地理信息提取方案，以实现即时通信服务与地理服务集成的目标，为即时通信用户提供即时的、一体化的空间信息服务。

即时通信文本包含大量的地理信息及服务需求信息，如地名信息、POI (Point of Interest，兴趣点)、空间关系查询和路径分析需求信息等。以“晚上去西直门吃饭”为例，用户接收到该信息后会对其进行具体解析：其中包含的地名信息(即目的地)是“西直门”；出发地是用户的当前位置，可由移动设备直接获得；“去”表达了路径分析需求，即查找一条从当前位置到“西直门”的路径；时间信息是“晚上”，需根据晚间的路况信息对路径进行合理的规划；“吃饭”是专题信息检索条件，需要对西直门周边的餐饮信息进行检索，以专题地图的形式提供合适的餐厅信息。

本文提出的即时通信文本中地理信息提取方案是对上述自然过程的数字化模拟，其总体架构如图 1 所示。在该方案中，首先对即时通信文本进行语法结构分析，完整的即时通信文本被分割成基本的语义单元，以提取其中空间和专题信息关键词，包括时间、地点和查询分析关键字等；然后基于语法分析得到的空间分析关键词，对信息中的空间分析需求进行模式分析和匹配，确定符合即时通信语义的地理服务信息方案；最后在电子地图软件中调用并实现对应的地理信息服务，并将分析结果即时或同步地展示给用户。

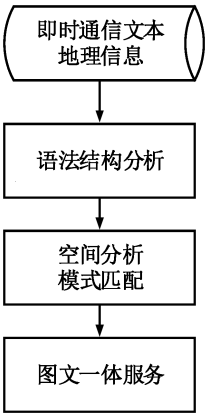

图 1 总体技术方案
Fig. 1 Overall technical solution

2 关键技术

本文提出的即时通信文本地理信息提取技术方案涉及的关键技术如下：针对信息文本语法结构分解需求，采用中文文本分词技术，将整个通信文本切割成基本的文本语义单元；针对空间分析模式匹配需求，基于既有的空间分析模式结果，对各类空间分析的语法模式进行研究，然后结合即时通信中常用查询分析关键字，实现通信文本中空间分析模式的匹配；针对图文一体服务需求，选用主流的电子地图软件，将得到的空间查询分析模式和关键字实现为具体的功能调用，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，其算法主要包括基于规则的分词方法、基于统计的分词方法和基于理解的分词方法[6]。基于规则的分词方法中，最常见的是最长词优先匹配法，Guo[7]对该算法的工作原理给予严格的形式解释，刘源等[8]将其大规模应用到汉语自动分词系统中。基于统计的分词方法主要包括基于期望最大值(expectation maximization)的方法和变长分词方法，李家福等[9]提出一种根据词语出现概率和基于极大似然原则构建的汉语自动分词的零阶马尔可夫模型。基于理解的分词算法是在分词的同时进行句法和语义分析，并利用语义和句法信息处理歧义现象，

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，中文文本分词能够对通信文本进行准确的语义分割，得到词组集合(即相互独立的关键词信息)，其中包括地理位置和空间分析需求信息以及在空间分析中可能用到的其他辅助信息，是后续分析的前提和基础。

目前，既有的很多中文分词系统都能满足即时通信文本中文分词需求。我们认为 ICTCIAS 分词系统具有较高的效率和准确率，同时由于其开源特性，便于集成开发。本文选用 ICTCIAS 作为中文文本分词的基础算法，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

查询分析模式识别指对中文文本分词输出的词组单元进行模式分析，提取其中的空间地物信息和空间查询分析需求，确定其对应的 GIS 分析方法和对应的分析要素。相关学者对于 GIS 空间分析的类别和内容开展了大量研究[12−16]。Unwin[15]将空间分析局限于点、线、面、曲面地图要素的参数描述和图形表述。郭仁忠[16]认为空间分析是基于地理对象的位置和形态特征的空间数据分析技术从空间信息内容出发，提出将空间分析分为 5 类: 空间位置、空间分布、空间形态、空间距离，以及空间相关(表 1)，此分类方法详细完整，与 GIS 系统结合紧密。本文的查询分析模式识别基于此分类方法展开。

根据郭仁忠[16]提出的空间分析类别、含义和特点，本文依次分析并设定各类空间分析的文本语法、形式化描述以及对应的 GIS 操作，表 1 给出分析结果和对应的示例。在技术实现中，需要基于中文文本分词输出的关键词信息，对关键词的词性及语法进行分类。将关键词中的空间地物信息、查询分析关键词及其组合模式依次与表 1 中形式化描述进行匹配，将得到的最佳匹配方案作为此文本对应的空间查询模式。将文本中的地物、时间等信息作为查询的要素信息，得到对应的 GIS 操作。由于即时通信文本常常是语法不规范的语言断片，在分析中需要给出必要的补充。例如“晚饭时候到北京大学东门集合吧”显然对应一个路径查询，但实际上并未给出起点，需要利用即时通信软件的自定位功能予以补充，或者要求用户交互确认。

2.3　图文一体服务

图文一体服务指将空间查询分析得到的基于地图的空间查询分析结果，在即时通信软件中与通信文本进行准实时的同步展示。图文一体的服务方式，可以为用户呈现与当前交流语义高度相关、丰富且直观的地图服务，大大提升用户交流体验。

目前，大量网络地图服务提供了 API 函数接口，用户可以在线提交查询分析需求，并得到对应的结果[17−18]，为实现即时通信中的图文一体服务提供了直接而有力的支持。本文基于主流网络地图服务系统，将查询分析模式识别中得到的 GIS 操作直接转化为网络地图服务对应的 API 函数，并将返回

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

的结果以图片的方式与即时通信文本进行同步显示。在网络地图服务系统的支持下，用户也可以通过点击图片进入地图系统，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、可行性以及运行效率，我们设计了相应算法，采用 C#和 JavaScript，在 Visual Studio 2012 平台上开发了验证系统，其中集成了 ICTCIAS 分词系统组件以及百度地图开发组件，实现对 ICTCIAS 分词系统以及百度地图服务的调用。验证系统的总体界面采用典型的即时通信软件风格，以便模拟和验证在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，输出即时通信文本的分词信息，提取其中出现的关键词，包括动词、空间地物信息、时间和其他限定词等；利用查询分析模式识别对分词信息进行正确的解析，形成地图服务调用方案；将地图服务调用方案提交百度地图服务进行查询分析，得到图片格式的返回结果，在系统界面中实现图文同步服务。

实验中采用带有地理信息的微信文本 216 例，均来自北京大学地球与空间科学学院 GIS 班 30 位同学的实际微信数据。研究发现，实例数据完全涵盖了表 1 给出的 5 类空间分析模式。其中，空间位置关系 186 例，主要表现为单独的地理实体或由“的”连接的两个地理实体；空间分布关系 23 例，多包含“沿着、附近”等关键词；空间形态关系 5 例，文本中存在“多大、多长”等关键词；空间距离关系 17 例，文本中存在表示距离的关键词，如“多远”等；空间方位、拓扑、相似和相关关系 31 例，文本中存在表示“以东、南侧”及“里/外/旁边”等关键词。

实验中根据微信群的具体特点，对部分群落方言进行翻译，比如“搓饭”等价于“吃饭”，单独出现的“学校”等价于“北京大学”，保证了相关分析和模式匹配的正确进行。

经本文所有作者人工验证，所有实例数据均得到正确处理，相关的分词结果、查询分析模式匹配和网络地图函数调用方案均与其语义相匹配。在北京大学校园网环境下，整体运行时间均在秒级，可以实现与即时文本通信的准实时同步。

我们选取部分典型用例来说明实例验证效果，如图 2 所示。可以明显看出，在即时通信中加入地理信息同步服务，将通信信息中文字的地理信息和分析需求以图片形式同步显示，大大提升了交流用户的直观体验，方便了交流、查询和决策。

图 2(a)中，通信文本为“北京大学的食堂好吃嘛?”。该例属于空间位置分析，关键词“北京大学”和“食堂”都为地理名词，根据地名库匹配为地图中地理实体的位置，返回的图片显示北京大学校园内食堂的具体位置。

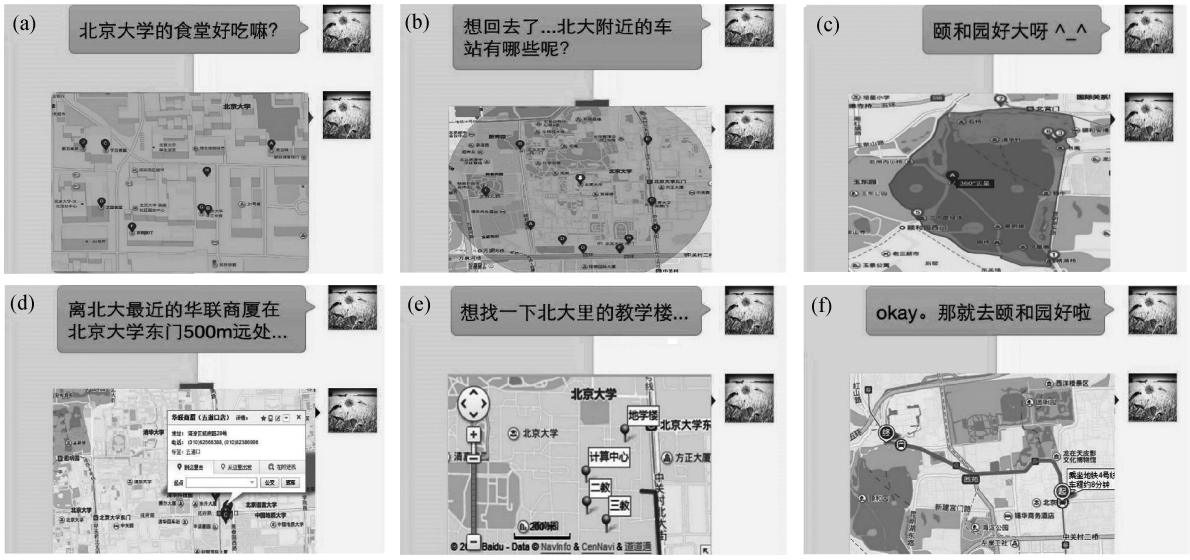

图 2 实例验证典型用例

Fig. 2 Typical verification cases

图 2(b)中，通信文本为“想回去了…北大附近的车站有哪些呢”。该例属于空间分布分析，包含的关键词是“北大”、“附近”和“车站”，其中“北大”使用别名信息解析为“北京大学”，车站图层与北京大学缓冲区图层进行叠加，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，通信文本为“颐和园好大呀^_^”，该例属于空间形态分析，句中的关键词是“颐和园”，地图中高亮显示“颐和园”的边界范围，并返回其面积信息。

图 2(d)中，通信文本为“离北大最近的华联商厦在北京大学东门 500 m 远处…”，该例属于空间距离类别，关键词是“离”、“北大”、“华联商厦”、“北京大学东门”和“500 m”，地图查询标注了距离北京大学东门约 500 m 远的华联商厦。

图 2(e)中，通信文本为“想找一下北大里的教学楼…”。该例属于空间拓扑分析，关键词是“北大”、“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，通信文本为“okay。那就去颐和园好啦”。该例属于空间距离分析，关键词是“去”和“颐和园”，空间分析返回从当前位置去颐和园的路径和乘车信息。

上述实例研究说明，本文提出的面向即时通信文本的地理信息提取技术方案可以顺利地予以编程实现，说明该方案具有良好的可行性。针对实例数据中的各类空间分析需求，输出结果全部通过人工验证，证明了该技术方案的正确性。在校园网环境下，验证系统秒级的反应速度符合即时通信软件的界面交互需求，运行效率符合实际需求。

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，本文提出一套综合使用文本分词、空间分析模式识别与图文一体服务等技术的即时通信文本地理信息提取技术方案，以实现即时通信与地图服务软件的集成应用，为移动用户提供更为智能、直观和便捷的应用服务。以微信和百度地图为例展开实例验证，实验结果证明该技术方案是合理、正确和可行的。本文成果进一步拓宽 GIS 应用领域，实现 WebGIS/LBS 地图服务增值，也增强了即时通信软件的空间服务能力。目前，基于移动设备的语音识别技术正在逐渐得到重视，如果将本文提出的技术方案与语音通信结合，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