

基于 WebGIS 与 SOLR 的地学可视化检索系统研究

孙洪亮¹, 王志宝¹, 孙相棋², 管泽礼¹

(1. 东北石油大学 计算机与信息技术学院, 黑龙江 大庆 163318;
2. 大庆油田有限责任公司第四采油厂 地质大队, 黑龙江 大庆 163511)

摘要:地球科学是数据密集型科学,信息检索是地学研究的必要步骤。文中针对目前主流的信息检索系统空间语义感知能力不足的问题,设计了顾及空间和语义的检索系统架构。在地学知识库的支持下,采用命名实体识别、实体消歧等自然语言处理技术,将非结构化文档空间化,采用 WebGIS 地理空间信息技术对查询过程与检索结果可视化。文中以石油勘探为例,在开源检索平台 SOLR 和空间数据库 PostgreSQL 的基础上实现了方法验证。经过系统测试和用户使用的验证,该系统提高了地球科学信息检索性能与用户体验。

关键词:地学信息检索; SOLR; 地球科学; 网络地理信息系统

中图分类号: P208; TP311

文献标识码: A

文章编号: 1673-629X(2016)06-0171-04

doi: 10.3969/j.issn.1673-629X.2016.06.038

Research on Geoscience Visualization Information Retrieval System Based on WebGIS and SOLR

SUN Hong-liang¹, WANG Zhi-bao¹, SUN Xiang-qi², GUAN Ze-li¹

(1. College of Computer & Information Technology, Northeast Petroleum University, Daqing 163318, China;
2. Geological Department, the Fourth Oil Production Plant of Daqing Oilfield Company, Daqing 163511, China)

Abstract: Earth science is data-intensive science, and information retrieval is a necessary step in earth sciences. As the mainstream information retrieval system can identify space semantic weakly, the architecture taking into account spatial entities and thematic information is put forward. The spatialization of unstructured documents are implemented with natural language processing including named entity recognition and disambiguation with the help of geoscience knowledge base. The visualization method of the query and retrieved results is put forward by using of WebGIS. A demo system based on SOLR, an open source information retrieval platform and PostgreSQL, a spatial database, is implemented to verify the method. Practices show that the system improves the performance and user experience by data test and validation.

Key words: geo-information retrieval; SOLR; earth sciences; Web GIS

0 引言

随着数字地球和智慧地球进程的推进,地球科学也进入大数据时代,在 Web 上和相关机构内部积累了海量的地球科学信息,信息检索是地球科学研究人员信息过滤和知识获取的必要工具。据统计,网络搜索引擎系统 19% 左右的检索词包含地理名词,15% 左右的检索词包含空间信息^[1]。由于相关的地理科学、地质科学、环境科学等领域的地球空间依赖性,地球信息检索系统要能主动感知用户查询和文档中的空间信息,并提供符合语境的可视化查询与结果浏览界面。

目前主流的信息检索系统,在用户查询阶段忽略了地名等关键词的特殊空间语义,在文档分析阶段也没有对文档中的空间信息进行识别和编码。例如,“河北南部的景区”切词成“河北”、“南部”和“景区”三个关键词,按照某种信息检索模型,实现查询和文档的最大匹配,发现最相关文档。这个过程不论是从语义角度,还是从检索可视化角度,都不能满足地球科学领域的特殊信息检索需求。

利用自然语言处理技术挖掘地球科学非结构化文档中的空间信息,采用地球空间信息技术实现对地学

信息检索的可视化,对提高地学信息检索的性能具有现实意义。

1 相关工作

为实现空间感知的网络信息检索,2002 年欧洲的卡迪夫大学、苏黎世大学等六所大学发起的 SPIRIT (Spatially-aware Information Retrieval on the Internet, 空间感知的网络信息检索) 项目,提出了地理空间信息检索的系统架构,以及文本空间解析、地理信息检索可视化、地理知识库、混合索引等研究主题,开辟了 GIR (Geographic Information Retrieval, 地理信息检索) 研究方向^[2]。

2006 年,西班牙巴伦西亚理工大学开发了面向网页搜索的地理信息检索系统 (Geographically-enhanced web search engine, GEOOREKA), 基于谷歌和雅虎搜索服务,集成了地理空间数据库,采用地图的方式允许用户对进行空间和主题双维查询,提高了网页地理搜索的质量^[3]。

2010 年,西班牙拉科鲁尼亚大学 (Local Search) 针对地理信息检索的需求,提出了一种支持查询扩展的混合索引结构,设计了地理信息检索系统架构,采用 TREC FT-91 和 TREC FT-94 数据集对原型系统进行了评测^[4]。

2012 年,弗吉尼亚理工学院与谷歌公司提出了时间、空间和文本集成检索框架,采用自然语言处理技术

对文本中的时空信息和主题信息建模,使用标签云和热度图方法对相关信息进行了可视化^[5]。

2010 年,德国海德堡大学使用 UIMA 文本管理平台和共现模型,提取了文档的时空轨迹,并采用 WebGIS 技术对文档的轨迹进行了可视化^[6]。

2013 年,北京大学针对当前量化的地理信息检索模型无法有效处理自然语义导致检索结果不理想的问题,以定性表达为基础,以推理方法为手段,实现 Web 文档中空间信息内容与查询请求的定性表达和信息提取,并使用实现了基于 WebGIS 的可视化信息检索系统—GeoSearch^[7-8]。

这些系统提取文本中的空间信息,通过空间和文本双重索引实现对非结构化文档的管理,采用信息可视化技术对检索结果进行视觉呈现。但是这些系统多是针对英文文档管理,主要抽取的是以地名为载体的地理空间信息,而地球科学还关注地层等地质空间特征。

文中从定性的角度出发,考虑中文环境下地球科学文档空间特征的复杂性,采用开源技术方案,实现对地球科学领域文档的可视化检索,提高了地球科学的信息检索体验与效率。

2 地学可视化信息检索总体架构

该系统的技术架构 (见图 1) 主要由两部分组成: 文档的预处理阶段和检索运行时阶段。

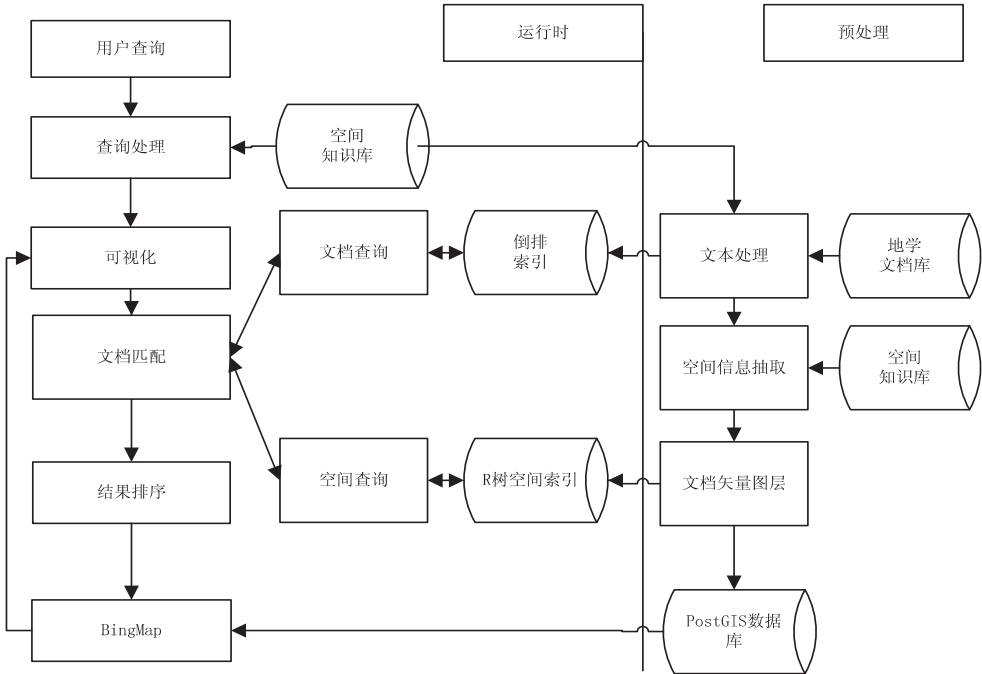


图 1 地学可视化信息检索系统架构

在预处理阶段,对地学文档进行归一化转换、领域分词,建立文本索引,文中采用 SOLR 平台作为文本处理平台;此外,还要做文档空间化处理,即采用空间命

名实体识别方法抽取文档中的地理空间信息与地质空间信息。文中采用空间知识库,结合空间实体在文档中的频率与排版样式,建立实体的重要性模型,在此模

型的基础上给出文档的空间主题,在空间知识库的支持下将文档的空间主题映射到地球空间的实际位置。文档空间矢量化根据每个文档的空间主题建立文档集的矢量图层,存储到 PostGIS 空间数据库,作为可视化引擎的输入。

在查询运行时阶段,用户输入主题查询和空间查询,表达自己的地质信息检索需求。如果用户的空间查询是一个空间实体,可以在地理视图上直接定位到该区域,进行查询可视化,用户也可以在地理视图上选择感兴趣的区域和地层作为空间查询输入。将业务查询与文本索引进行匹配,将空间查询与空间索引进行匹配,然后对两个结果集取交集运算,对最终的结果考虑主题和空间两个因子进行排序输出,将命中文档的空间范围叠置在 BingMap 展示,同时配合用户的交互操作,对文档列表视图和地理视图进行关联更新,实现对检索结果的可视化。

3 文档空间化技术实现

文档空间化主要包括领域分词、空间命名实体识别、语义消歧、空间编码、矢量化和索引化,如图 2 所示。其中,空间命名实体识别、语义消歧、空间编码是技术难点^[9]。

领域分词:地球科学对某一字符串有自身特定的语义理解,比如“松深 60”是一个探井井名,但是在一般的切词算法中容易被分割为“松深”+“60”两个词。文中采用基于领域词典和正向迭代最细粒度切分相结合^[10]的切词方法。

空间命名实体识别:识别文本中涉及的地理空间命名实体与地质空间命名实体,地球科学涉及地球各个圈层,文中采用 SWEET^[11]地球与环境科学本体中具有空间属性的概念作为命名实体分类,识别方法也是采用知识与机器学习结合的方法^[12-13]。除此之外,还解决了实体歧义问题,例如“萨尔图”是地理行政区域也是地质空间的底层。

语义消歧:包括地质空间实体与地质空间实体的歧义,例如“萨尔图”可能是松辽盆地的一个油田,也可能是指大庆市的一个行政区;还有地质空间实体与非地质空间实体的歧义,例如“铁人”可以是王进喜,也可以是铁人广场。文中系统采用基于本体的语义消歧方法,即在地球科学本体的统一约束下,根据词汇上下文的语义判定词汇的最大可能类别^[14]。

空间编码:文档可能会涉及水平空间和垂向空间的多个空间位置,文中认为这些空间位置中隐含着这个文档的空间主题,将文档中识别出来的空间命名实体做聚焦计算,将结果映射到地球空间,实现文档的空间主题编码。包括水平空间和垂直空间两个维度的映

射,例如一篇文档的水平空间信息为“松辽盆地北部”,垂直空间信息为“白垩纪”地层。

矢量化:遍历整个文档数据库,通过空间编码信息,根据地名数据库和地球科学领域空间数据,根据每个文档的 MBR 建立一个矢量图层,所建立的图层保存在 PostGIS 空间数据库中,以便于支持空间查询。

建立空间索引:采用 R 树索引建立海量文档矢量索引,提高文档空间查询的效率。然后,使用文中的倒排索引和 R 树建立组合索引,支持用户的主题和空间的组合查询。

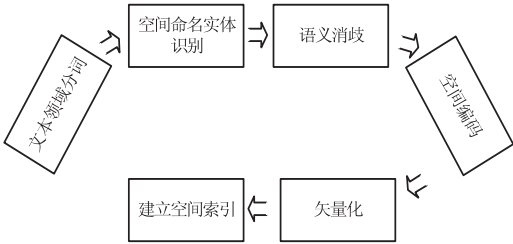


图 2 文档空间化流程图

地质空间本体:在地质文档空间化中多处用到地质空间本体,文中系统以 SWEET 本体为基础建立本体 TBox,ABox 来自于各类地质词典、叙词表和权威地质学数据库。地质空间本体采用融合式建立方法,对于 TBox 层的语义冲突主要还是人工消解,对于 Abox 的语义冲突采用半监督机器学习方法消解。

4 基于 WebGIS 的交互式地质信息检索可视化

信息检索可视化将文档、用户查询、信息检索模型、检索过程以及检索结果中各种语义关系转换成图形,在一个二维或者三维的空间中可视化,帮助地质研究人员理解检索结果、调整检索方向。WebGIS 在浏览器中实现了高效丰富的地质信息浏览与空间查询,是地质研究人员常用的信息管理工具,为抽象的地质信息检索相关对象提供了具体直观的地质上下文,是天然的地质信息检索可视化空间。

地质信息检索可视化原理如图 3 所示。

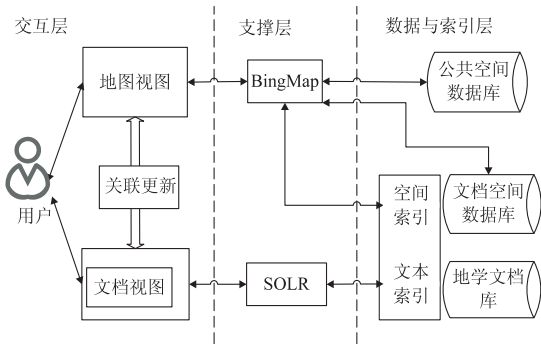


图 3 地质信息检索可视化原理

利用 WebGIS 先对用户空间查询进行可视化或者

定义用户的空间查询,WebGIS 内在的广角与聚焦信息可视化技术实现快速地定义用户的感兴趣区域;在结果浏览阶段,文档集视图与地理信息视图通过关联更新,在互动中实现对目标文档的定位。微软 BingMap 具有丰富的空间数据、友好的用户体验和丰富的二次开发功能。文中系统以 BingMap JS API 为基础,通过 OpenLayers 开源组件对 BingMap 进行尺度变换控制,使用文档 MBR 对文档空间特征进行表达,将检索结果在地图上呈现,并采用关联更新技术对文档视图进行动态更新。

5 原型系统实现与评测

文中系统采用开源软件 SOLR 5.1 作为检索平台,SOLR 的分词与文本分析都支持插件式的功能集成,文档查询与索引都支持 REST 风格的网络服务 API,可与其他模块实现快捷的松耦合通信,SOLR 的近实时索引技术和集群扩展技术为以后的大规模应用提供了保障。BingMap 是微软的在线地图,具有丰富的影像数据与空间矢量数据,是很好的在线 WebGIS。文档矢量图层与 BingMap 的客户端渲染采用 OpenLayers 实现,文档的关键词检索与列表显示采用 JQuery 库实现。为保证用户体验,浏览器与服务端的通信采用 XMLHTTP 协议,通信格式使用 JSON 格式。

系统测试数据来自 CNKI 下载的 2 000 篇文档,主要领域为地球科学石油勘探子领域。对这些文档进行归一化处理,形成 SOLR 可索引的格式,再让专业人员对文档进行标注,形成测试预料库。检索结果排序采用双维模型:语义维和空间维。弥补了传统文档检索忽略文档空间信息而导致的准确率降低的问题。通过实验证明,文中系统的 F_1 测度在 75% 以上,这是一个不错的结果,同时也是一个可应用的结果。

6 结束语

文中分析了地球科学领域对空间可视化信息检索的领域需求,分析了现有领域信息检索系统和地理信息检索系统的不足。考虑地球科学领域对信息检索的特殊需求,设计了地学可视化信息检索系统架构,给出了非结构化文档的空间化方法,实现了文档中空间信息的提取,实现了基于 WebGIS 的交互式地学信息检索可视化。最后以开源信息检索平台 SOLR 5.1 为基础平台,集成了 JQuery、EasyUi、OpenLayers 若干开源模块,实现了系统原型。该原型系统经过专业人员的测试和试用,可以提高地球科学专业人员的信息检索效率。

未来的工作,将进一步考虑地球科学含有三维空间信息的事实,进一步区分水平空间信息和垂向空间

信息,同时研发基于 3D 数字地球的可视化平台,以地球空间作为信息可视化空间,同时结合信息检索本源的语义空间检索可视化,进一步提升地球科学信息可视化检索的效果。

参考文献:

- [1] Sanderson M, Kohler J. Analyzing geographic queries [C]// Sanderson M, Jrvelin K, Allan J, et al. Proceedings of the 2004 workshop on geographic information retrieval, 27th annual international ACM SIGIR conference. New York: ACM Press, 2004: 245–246.
- [2] Jones C B, Purves R S. Geographical information retrieval [J]. International Journal of Geographical Information Science, 2008, 22(3): 219–228.
- [3] Buscaldi D, Rosso P. Geooreka: enhancing Web searches with geographical information [C]//de Antonellis V, Castano S, Catania B, et al. Proceedings of the seventeenth Italian symposium on advanced database systems. [s. l.]: [s. n.], 2009: 205–212.
- [4] Brisaboa N, Luaces M, Places A, et al. Exploiting geographic references of documents in a geographical information retrieval system using an ontology-based index [J]. GeoInformatica, 2010, 14(3): 307–331.
- [5] Wang B, Dong H, Boedihardjo A P, et al. An integrated framework for spatio-temporal-textual search and mining [C]// Proceedings of the 20th international conference on advances in geographic information systems. [s. l.]: ACM, 2012: 570–573.
- [6] Strötgen J, Gertz M, Popov P. Extraction and exploration of spatio-temporal information in documents [C]// Proceedings of the 6th workshop on geographic information retrieval. [s. l.]: ACM, 2010: 1–8.
- [7] 刘磊, 高勇, 林星, 等. 定性地理信息检索方法及其实现 [J]. 北京大学学报: 自然科学版, 2013, 49(6): 1017–1024.
- [8] 林星. 地理信息检索中的定性信息表达方法和检索模型研究 [D]. 北京: 北京大学, 2011.
- [9] 张毅, 王星光, 陈敏, 等. 基于语义的文本地理范围提取方法 [J]. 高技术通讯, 2012, 22(2): 165–170.
- [10] Lin Liangyi. ik-analyzer-java 开源中文分词器 [EB/OL]. 2015. <https://code.google.com/p/ik-analyzer/>.
- [11] Jet Propulsion Laboratory. SWEET overview [EB/OL]. 2015. <http://sweet.jpl.nasa.gov/>.
- [12] 鞠久朋, 张伟伟, 宁建军, 等. CRF 与规则相结合的地理空间命名实体识别 [J]. 计算机工程, 2011, 37(7): 210–212.
- [13] 唐旭日, 陈小荷, 张雪英. 中文文本的地名解析方法研究 [J]. 武汉大学学报: 信息科学版, 2010, 35(8): 930–935.
- [14] Buscaldi D. Approaches to disambiguating toponyms [C]// Proc of SIGSPATIAL. New York, NY, USA: ACM, 2011: 16–19.