

文章编号:1001-9081(2010)04-1132-03

基于网页信息检索的地理信息变化检测方法

曾文华,黄桦

(浙江省测绘科学技术研究院,杭州310012)

(mapz@sohu.com)

摘要:针对地理信息变化频繁,难以及时发现的问题,提出了一种基于网页信息检索的地理信息变化检测方法,通过设计搜索条件在互联网上收集符合条件的网页,设计评价方法评价搜索结果的可信度,并对最终搜索结果进行统计和空间分析,实现基于网页信息检索技术的地理信息变化检测。以杭州地区为例,开发了基于Web的杭州地区地物变化检测系统,验证了该方法的可行性和有效性,为区域的地物变化检测提供了新方法。

关键词:Web;变化检测;可信度评价;信息检索;地理信息

中图分类号:TP311 **文献标志码:**A

Method for detecting changed geographical information based on information retrieval of Web pages

ZENG Wen-hua, HUANG Hua

(Zhejiang Institute of Surveying and Mapping, Hangzhou Zhejiang 310012, China)

Abstract: For the difficulty to detect the frequent changes of geographic information, a method based on information retrieval of monitoring changed geographic information was presented. By designing the search condition, estimating the reliability ranking and analyzing all results by statistics and spatial analysis, the geographic information monitoring based on information retrieval has been realized. Finally the system for monitoring the geographic information change of Hangzhou area was developed, which verified the feasibility and validity of this method and provided a new method to detect the changed geographic information.

Key words: Web; change monitoring; reliability rank; information retrieval; geographical information

0 引言

地理空间数据信息的现势性是地理信息系统(Geographical Information System, GIS)的核心。一方面地理信息应用的日益频繁,与国民经济、人民生活密切相关;另一方面社会发展迅猛,使得地表变化加剧,地理信息变化加快。地理信息的现势性,即在计算机中记录的与位置有关的信息与现实空间的一致性,却不尽人意,远远赶不上地物目标的实地变化速度;而要保持城市地理空间信息的现势性更是难上加难。因此实现对变化的地理信息快速高效的检测是一个十分紧迫的问题。

长久以来遥感影像是检测地理信息变化的主要信息源和主要手段,影像的来源主要有卫星遥感和飞行器航摄,它们都有投入过大、费用过高的特点,不适合频繁、随时的获取,而且影像上不能得到定性的信息,如新修道路、水库的名称信息等。如今互联网的迅猛发展,使得它已经成为能够与报纸、电视和广播齐名的四大传媒之一。与传统媒体相比,互联网新闻在时效性上有着不言而喻的优势。如今越来越多表示地物发生变化的信息正以互联网新闻的形式展现出来,以网页为检测地理信息变化的信息源与遥感影像为检测信息源相比,有时效性好、描述信息丰富、自动识别率高和信息获取的廉价性等优点。

随着信息检索(Information Retrieval, IR)技术的发展,近年来GIS研究者对从海量、非结构化的文档(主要是网页)中

获取地理信息资源表现出了极大的兴趣^[1-3],形成了新的研究热点——地理信息检索(Geographical Information Retrieval, GIR)^[4],把IR技术和GIS空间分析思想结合起来,来检索网页中人们感兴趣的地理信息。

利用互联网上的网页作为地理信息变化的检测信息源,国内在这方面的研究较少。本文提出的基于网页信息检索的地理信息检测方法就是通过借鉴现有IR技术和GIR技术,对特定的网站进行信息检索,来发现表示地理信息发生变化的新闻,从而得到检测区域内发生变化的地理信息。它为检测特定区域地物的变化提供了一种全新的方法。

1 基于IR技术的地理信息变化检测方法

1.1 总体设计思想

总体设计思想如图1所示,它由4个部分组成。构造搜索条件,包括定义能表示地理信息发生变化的关键字、关键字组合和被搜索的网站;使用成熟的商业搜索引擎基于该搜索条件进行搜索;分析返回的搜索结果,并对每条新闻都做可信度评价,即该新闻表示地理信息发生变化的可信程度;存储搜索结果并进行统计分析。

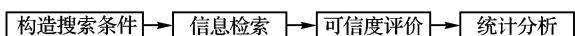


图1 总体设计

1.2 搜索条件设计

1.2.1 限定搜索网站

收集检测区域各级政府的新闻网以及政府机构网,比如

收稿日期:2009-10-15;修回日期:2009-12-30。 基金项目:地理空间信息工程国家测绘局重点实验室经费资助项目(200962)。

作者简介:曾文华(1976-),男,浙江建德人,高级工程师,主要研究方向:地理信息系统、电子地图; 黄桦(1982-),男,浙江杭州人,助理工程师,硕士研究生,主要研究方向:地理信息系统。

交通、水利、林业、国土、规划等政府机构网站。通过限定搜索网站,一方面可以限定搜索区域(行政区域),从而最大限度地减少由于地名相同而造成搜索结果的差错;另一方面也能确保网站信息的权威性和时效性,从而保证地物检测搜索结果的可靠性和时效性;最后可以加快搜索的速度,搜索目标更有针对性,减少对搜索结果无谓的分析和排查。

1.2.2 设计搜索关键字

现有成熟的搜索引擎都是基于关键字的查询,因此关键字的设计对搜索结果非常重要。本文按照查询关键字是否包含空间关系把查询关键字分为两类:第一类由“地名+地理要素名+动词”组成,表示在什么地方什么地理要素发生了变化;第二类由“地理要素名、动词+空间关系(包含)+地理位置”组成,表示在什么检测区域内什么地理要素发生了变化。

对于第一类搜索来说,地名关键字可以从检测区域内街道、乡、镇名称得到;地理要素名称可由国标地理要素分类中得到^[5];收集表示各种地理要素发生变化的动词。表1是地理要素名和表示发生变化的动词示例。

表1 地理要素名和动词

地理要素分类	地理要素名称	表示变化的动词
交通	公路	通车、建成、拓宽
	隧道	贯通
	隧道	通车、建成
...		
水系	河流	改造
	运河	通航
...		

对于第二类搜索来说,由于限定了搜索网站,所以能够尽可能地保证搜索得到的结果是发生在被检测区域内的。因此“地理要素名、动词+空间关系(包含)+地理位置”可以简化成“地理要素名+动词”或“动词”来进行搜索。

1.3 可信度评价

现在市场上成熟的搜索引擎能够保证大部分用户在搜索结果的第一页上找到想要的结果。之所以能做到这点很大程度上是采用了PageRank算法,它和其他排序算法一起,共同决定着搜索结果的排序^[6]。为了检测到更多可靠的表示地物变化的新闻,本文在现有搜索引擎提供的排序结果之上,对搜索结果进行可信度评价。影响可信度的因素主要有关键字本身的重要程度、关键字出现在网页的位置、关键字出现密集程度和地物发生变化的准确时间。进行可信度评价,要经历以下4个步骤。

1)设定权重。由于同一个关键字出现在网页的不同位置,它所表达网页内容的能力是有差别的。文本引入结构层次权重系统的概念,对一个文本文档,按照其结构分层,依次为标题、正文等,按照不同的结构域在文档的重要程度,对不同域的特征项赋予不同程度的加权^[7]见表2。其中 $\alpha, \beta > 1$ 且 $\alpha > \beta$,标题、正文、层次结构权重依次减少,具体的取值需要通过实验得到,由于其他层次对排序影响有限,一律设置为0,从而减少计算的复杂度。

对于由“地名+地理要素名+动词”组成的查询关键来说,它们对检测地物变化的作用也是不同的。因此本文对它们也设定权重见表3,其中 $C > B > A$, 动词、地理要素名、地方权重依次减少。具体的取值需要通过实验得到。

关键字出现在文档中的疏密程度也影响着网页表现地物发生变化的可信程度。比如3个关键字同时出现在一段话中和3个关键字分别出现在正文第一段、中间一段和最后一段中对检测地物变化的效果是不同的。本文把关键字同时出现在正文一段话中定义为密集出现,其他情况视为非密集出现。本文对这两种情况设定相关系数如下:密集出现为 X , 非密集出现为 Y , 其中 $X > 1 > Y$ 。

表2 不同域的特征项的权重

域	权重系数
标题	α
正文	β
其他	0

表3 关键字权重

关键字	权重系数
地名	A
地理要素名	B
动词	C

2)权重计算。对于第一类查询来说:权重 $W = \sum_{i=1,2,3} K_i * Y_i$;如果关键字都出现在正文,权重 $W = W * m$ 。

K_i 表示第 i 个关键字的权重系数, Y_i 表示第 i 个关键字出现域的权重系数, m 表示关键字是否集中出现的相关权重系数。

对于第二类查询来说:不管是一个关键字还是两个关键字,本文只关注关键字是否同时出现在标题中,如果是则认为它的权重 W 和第一类查询权重最大值相同。

3)获取地物变化时间。通过日期范式能从网页正文中得到网页发布的时间,但有时候标题中会出现“年底、两年内、月底”等表示将来的词汇,而地物发生变化的真实时间是在网页发布时间的“年底、两年内、月底”出现。准确计算地物在现在到底是否发生变化,对检测结果至关重要。因此在获取网页发布时间的同时,如果网页标题中出现了表示将来的词汇,则需要分析、修正地物发生变化的真正时间。

4)入库。为了能够对搜索结果进行统计和空间分析,需要对得到的每条搜索结果进行处理。根据搜索地理要素名称,把每条搜索结果归类到9大地理要素分类中;根据搜索地名,为每条搜索结果进行空间位置定位,以标识表明发生变化的地理要素的位置。

1.4 统计和空间分析

对搜索结果按照地理要素分类进行统计,以表明在该区域各种地理类型变化的频繁程度。根据每个搜索结果的经维度坐标,把它们定位到地图上,通过空间分析来确定各个区域地理要素变化的频繁程度。

2 应用实例

2.1 检测系统的实现

本文以浙江杭州地区为检测范围,用Google搜索API为搜索引擎,用MySQL存储搜索结果,利用上文所描述的方法,建成了基于Web的杭州地区地物变化检测系统。该系统的建设主要由以下步骤组成。

首先构建搜索条件。本系统收集杭州地区各级政府网站及新闻网站共41个;收集杭州地区所有街道、乡、镇名称共210个^[8]。

然后利用 Google 搜索 API 对限定的网站进行关键字搜索，并进行可信度排序。域的权重系数分别为： $\alpha = 5, \beta = 3$ ；关键字权重系数为： $A = 2, B = 4, C = 6$ ；相关权重系统为： $X = 1.2, Y = 0.8$ 。对每个搜索结果计算得到权重后，进行可信度分级并通过分析得到地物发生变化的时间，分级标准如表 4 所示。

2.2 结果与分析

本检测系统在一台千兆带宽、Windows XP 操作系统下分别于 2009 年 4 月 14 日和 2009 年 5 月 14 日进行两次搜索。

表 5 部分搜索结果

标题	内容	时间	关键字	可信度
二棉路西伸通车了	“老杭二棉厂附近新修了条路，前两天好像通车了。”昨日，网友“微澜”反映，她周末出行时，发现工人路旁边的河上建了新桥，车可以直接上桥开往萧杭路…	2009 年 6 月 8 日	桥 通车	1
萧山网新闻中心河庄网	江东工业园区二期闸北安置区块，已做好道路等用地的报批工作，河庄大道拓宽工程已列入区政府 2008 年实施项目，九桥东接线一标段施工顺利，有望年底竣工通车	2008 年 9 月 7 日	道路 通车	2
江东二期安置小区建设建站顺利	2009 年 5 月 8 日…江东二期安置小区两户联建区块位于河庄大道以西、江东大道南北两侧的闸北和同二村，规划总用地面积 840 亩，将安置农户 1241 户。多层安置区块位于河庄…	2009 年 5 月 8 日	河庄 小区 规划	3
龙坞乡村茶文化休闲旅游区将再度扩容	2009 年 3 月 23 日，其中，溪涧景观带以溪水甘冽、蜿蜒流淌的龙门溪为主线，两侧设绿带、临溪茶楼，与青龙山水库、听松亭、横山松径、横山茶园等景观呼应。建成后，辅以龙…	2009 年 3 月 23 日	龙门 水库 建成	4

表 6 2009 年 4 月 14 日搜索结果

年份	可信度	记录数
2007	1	94
	2	6
	3	13
	4	207
2008	1	213
	2	12
	3	37
	4	425
2009	1	102
	2	15
	3	18
	4	249

经过人工判断，可信度为 1 的准确度超过 60%，可信度为 4 的准确性超过 30%。之所以在第二次搜索中还会出现 2007 年、2008 年的搜索结果，是因为 Google 对各个网站网页的抓取时间和频率不同而造成的。对搜索结果按照 9 大要素进行统计，发现交通类中公路变化最大。而把每条搜索结果都定位到杭州地图上，发现杭州市城区地理要素发生变化最为频繁。

3 结语

本文探讨了一种基于网页信息检索的地理信息变化检测方法，该方法是以特定网站中海量、持续更新的网页为信息源，利用信息检索技术对这些网站进行信息检索，并对检索结果进行可信度评价，从而获取最有可能表示检测区域内地物发生变化的新闻。由于新闻网站内容的高准确性和高时效性，使得这种方法能够更快更准确地找到发生变化的地理信息。

本文提出的方法只针对文本信息的解释，没有涉及到图

完成第一次搜索大约需要半天时间，完成第二次搜索大约需要 5 h。表 5 是这两次搜索到的部分结果，表 6、7 是这两次搜索的结果统计。

表 4 可信度分级

权重(m)	等级
$m \geq 52$	1
$48 \leq m < 52$	2
$40 \leq m < 48$	3
$m < 40$	4

片、地图等其他形式信息的检测，随着互联网表述地理信息的地图网站、工程示意图等信息的丰富，对这些信息的检测对地理信息变化的发现变得十分重要，应该成为下一个研究的方向。另一方面，本方法得到的信息都是针对变化的描述信息，要绘制到地图上，还需要配合地理底图、实地调绘、影像判读等资料和手段，最终实现地理信息的更新，保持其现势性。随着信息化程度的提高，以及更优的可信度评价算法出现，基于网页信息检索的地理信息变化检测将有越来越广阔的发展空间。

参考文献：

- [1] PURVES R S, CLOUGH P, JONES C B. The design and implementation of SPIRIT: A spatially aware search engine for information retrieval on the Internet [J]. International Journal of Geographical Information Science, 2007, 21(7): 717–745.
- [2] ARAMPATZIS A, van KREVELD M, JONES R I, et al. Web-based delineation of imprecise regions [J]. Computers, Environment and Urban Systems, 2006, 30(4): 436–459.
- [3] McCURLEY S K. Geospatial mapping and navigation of the Web [C]// Proceedings of the 10 th International WWW Conference. New York: ACM Press, 2001: 221–229.
- [4] JONES C B, PURVES R S. Geographical information retrieval [J]. International Journal of Geographical Information Science, 2008, 22(3): 219–228.
- [5] 国家测绘局测绘标准化研究所. GB/T 13923—2006. 基础地理信息要素分类与代码[S]. 北京: 中国标准出版社, 2006.
- [6] 黄德才, 戚华春. PageRank 算法研究[J]. 计算机工程, 2006, 32(4): 145–162.
- [7] 李凡, 林爱武, 陈国社. 一种基于 VSM 文本分类系统的设计与实现[J]. 华中科技大学学报: 自然科学版, 2005, 33(3): 53–55.
- [8] 浙江地名. 浙江省乡、镇、街道名录[EB/OL].[2009-07-01]. <http://www.zjsdmw.com/sites/main/template/0001.aspx?id=147&Page=1>.