引文格式：郭俊枫,赵仁亮,郑娇龙.面向网页文本的地理要素变化发现[J].地理信息世界,2015,22(1):52-56.

面向网页文本的地理要素变化发现

郭俊枫[1]，赵仁亮[2]，郑娇龙[1]

(1.中国矿业大学(北京)地球科学与测绘工程学院，北京 100083；2.国家基础地理信息中心，北京 100830)

作者简介：
郭俊枫(1990-)，男，山西运城人，地图制图学与地理信息工程专业硕士研究生，主要研究方向为地理要素变化发现。
E-mail:
guojfgis@163.com

收稿日期：2014-10-14

【摘要】地理要素变化发现是地理信息数据库动态更新的重要组成部分。互联网在信息传播中扮演着越来越重要的角色，网页文本中蕴含着一些现势性很强的地理要素信息，可作为地理要素变化发现的数据源。本文结合网络爬虫和朴素贝叶斯分类模型，提出并实现了一种面向网页文本的地理要素变化发现方法。首先，本文在收集分析地理要素变化新闻锚文本的基础上，构建了网络地理要素变化新闻关键词库，并基于关键词库设计了适于地理要素变化发现的网络爬虫，实现了候选网页文本的主动获取；接着为了提取地理要素变化新闻，本文训练构造了适于地理要素变化发现的朴素贝叶斯分类器，对候选网页文本进行筛选。最后通过实验对比了本文方法与现有方法在准确性和全面性上的表现。

【关键词】地理要素；动态更新；变化发现；网页文本；贝叶斯分类
【中图分类号】P2;TP3 【文献标识码】A 【文章编号】1672-1586(2015)01-0052-05

Changing Information Search of Geographic Features Based on Web Page

GUO Junfeng[1], ZHAO Renliang[2], ZHENG Jiaolong[1]
(1.College of Geoscience & Surveying Engineering, China University of Mining & Technology, Beijing, 100083; 2.National Geomatics Center of China, Beijing, 100830)

Abstract: Searching for Changing information is the essential part for dynamic updating of geographic database. Internet has been more and more important for information dissemination. Web page contains up-to-date geographic feature information and it can be used for the aid of changed detection. A new method is presented and realized. Fisrtly, a key words base of geographic changed news is established by analyzing and summarizing anchor texts of geographic changed news. To acquire candidate web page forwardly, a web crawler is designed and realized that is suitable for geographic changed detecting, and then we constructed a Naive Bayes classifier, and fetch geographic changed news using the classifier, finally, we compared existing methods with ours through experiment. Results indicate that this method improved accuracy of detection and equalled in comprehensiveness with the existing methods.

Key words: dynamic udpdating; geographic features; changed detecting; internet web page text; bayes classification

0 引 言

地理信息数据库的更新是地理信息服务的保障。现阶段，更新已由全面更新转变为“更新内容灵性化，更新周期适时化，更新方式多元化”[1]的动态更新方式，需要快速及时地发现地理要素的变化。互联网在信息传播中扮演着越来越重要的角色，网页文本中蕴含着一些现势性很强的地理要素信息，可作为地理要素变化发现的数据源。以网页文本作为数据源发现地理要素变化实质上是从互联网中发现以地理要素变化为主题的新闻。该方法具有快速、廉价的优点，是一个颇具研究价值的变化发现方法。

以往学者们在进行特定主题的网络地理信息发现时，主要采取网络爬虫结合布尔模型或向量空间模型的方法：闫会杰等人直接通过关键词匹配的方式来发现描述地理要素变化的新闻[2]。王曙、吉雷静、曾文华等人，为了发现地理要素变化新闻，首先利用google api获取与设定关键词（如：公路 通车）相关性较大的候选网页文本，再通过关键词匹配来进一步筛选提取，而且是基于一定的规则进行匹配的[3-5]。张春菊、武昊等人分别针对新地名和地理信息服务的发现，采用向量空间模型计算网页文本与各自地理信息主题的相关度，设定阈值，大于阈值的则认为是所需的信息[6-7]。但以上研究所提出的方法主要侧重于对于相关信息的发现，而对于发现信息的进一步准确抽取方法的研究较少，然而

在实际生产中，准确地抽取所发现的相关信息从而过滤噪声对于提高生产效率至关重要。本文针对这一问题，研究实现了一种新的面向网页文本的地理要素变化发现方法。

1 总体流程

面向网页文本的地理要素变化发现，实质是从互联网中发现以地理要素变化为主题的新闻（总体流程如图1所示）。该任务可以拆解为以下两部分：

1）候选网页文本的获取：为了发现地理要素变化新闻，首先需要从互联网中按照一定的方式主动地获取网页。获取到的候选网页不便于提取，故需要对其解析，得到网页的标题和正文。

2）地理要素变化新闻的提取：候选网页文本集中包含许多与地理要素变化无关的网页文本，需要对其筛选，以提取地理要素变化新闻。

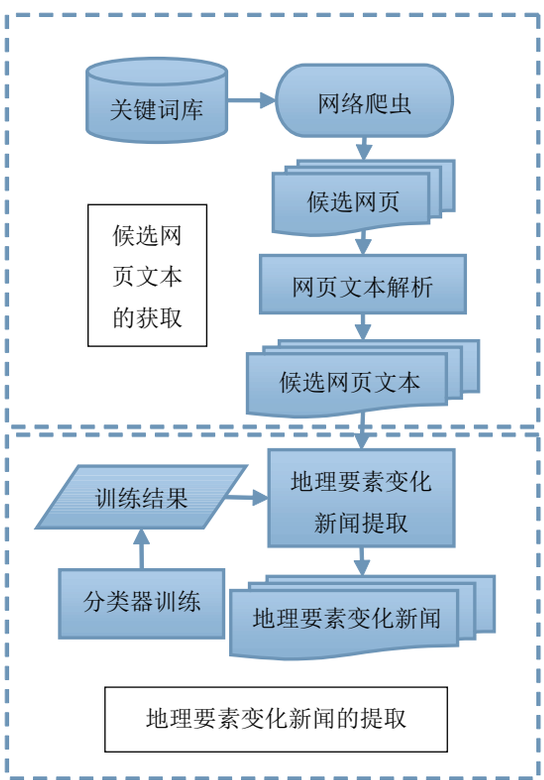

图1 总体流程
Fig.1 The overall flow

2 候选网页文本的获取

本部分的任务是从互联网中主动获取网页并解析其文本。网络爬虫是网页获取的一种常用工具，它可以按照一定的方式自动抓取网页。因此，本文构建了一个适于地理要素变化发现的网络爬虫来获取候选网页。该爬虫的工作流程如图2所示。

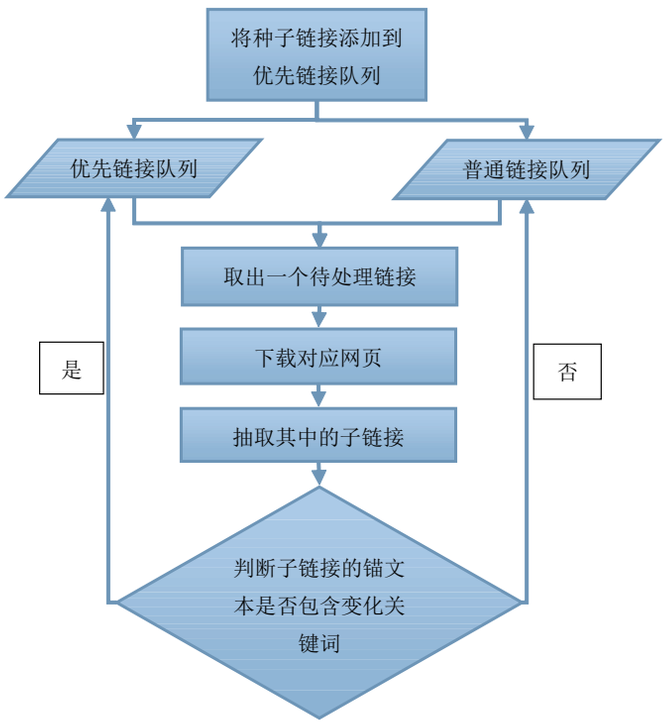

图2 爬虫工作流程
Fig.2 The workflow of crawler

候选网页获取的算法流程如下：

1）人工收集地理要素变化新闻网页在父页面中的锚文本，并总结归纳其中表达地理要素和变化的词，构建地理要素变化关键词库。

2）构建两个队列用于存储链接，分别命名为优先链接队列和普通链接队列。

3）人工收集富含地理要素变化新闻的网站，并将其网址添加到优先链接队列中。

4）从链接队列中取出一个链接，取出规则是优先取优先链接队列中的链接，只有当优先链接队列为空时才取普通链接队列中的链接。

5）按照一定的网络协议下载4)中取出链接所对应的网页并存储。

6）抽取5)中所下载网页中包含的子链接，若子链接在父页面中的锚文本包含1)中收集到的地理要素变化关键词，则将其添加到优先链接队列，否则将其添加到普通链接队列。

7）跳转到4)，循环执行直至达到人为设定的终止条件。

为了提高发现的效率，本算法在步骤6)中进行了链接分析，依据链接对应锚文本是否包含地理要素变化关键词来判断链接对应网页以地理要素变化为主题的可能性，从而确定是否需要优先处理。

为了便于依据网页内容进一步提取地理要素变化新闻，需要对候选网页进行解析。Joup是一款JAVA语言编写的HTML解析器，它提供了一套快捷的API，可以使用DOM或CSS选择器轻松地查找、取出数据，还可以操作HTML元素。本文利用Jsoup提供的API实现对网页标题和正文文本的解析。

3 地理要素变化新闻的提取

本部分的任务是筛选候选网页，提取地理要素变化新闻。以往学者在提取特定主题网络地理信息，多采用布尔模型和向量空间模型，但提取准确性有待改善。

分类是将一个未知样本分到几个预先已知类的过程。在众多分类模型中，朴素贝叶斯分类模型是应用最为广泛的模型之一。虽然朴素贝叶斯分类模型基于特征独立性假设，且该假设在实际中往往不成立，但由于其最大后验概率决策的规则，致使朴素贝叶斯分类模型在实践中往往取得良好的分类效果[8]。

提取地理要素变化新闻可看作是一个二分类问题（地理要素变化类、非地理要素变化类），因此，本文基于朴素贝叶斯分类技术构建了适于地理要素变化发现的朴素贝叶斯分类器。并利用该分类器筛选候选网页，提取出地理要素变化新闻。

地理要素变化新闻提取的算法流程如下：

1）人工从网络中收集5 000条新闻标题并判断其类别（地理要素变化类、非地理要素变化类），称之为训练数据。

2）对训练数据进行分词，将出现的所有词语存储到一个表中，称之为词表。

3）分别统计词表中词语在两个类别的训练数据中的词频，根据式(1)和式(2)计算两个类别的概率以及词表中词语在两个类别中出现的概率，并将计算结果存储。

$$P(\text{类别}c)=\frac{\text{训练数据中属于类别}c\text{的条数}}{\text{训练数据的总条数}} \tag{1}$$

$$P(\text{词}a\,|\,\text{类别}c)=\frac{1+\text{类别}c\text{的训练数据中词 }a\text{的词频}}{\text{词表大小}+\text{类别}c\text{的训练数据中所有词的 词频和}}$$

$$\text{词}a\in\text{词表}\quad \text{类别}c\in\{\text{地理要素变化类，非地 理要素变化类}\} \tag{2}$$

4）从候选网页解析出的网页标题中取出一个，对其进行分词并统计分词结果，构成备选词集。

5）从备选词集中移除不包含在词表中的词语，得到词集。

6）基于3)中的计算结果，根据式(3)分别计算两个类别确定的条件下待分类标题出现的条件概率与该类别概率的乘积。

$$P(\text{待分类标题}\,|\,\text{类别}c)\bullet P(\text{类别}c)=P(\text{词}w_1\,|\,\text{类别}c)\bullet P(\text{词}w_2\,|\,\text{类别}c)...P(\text{词}w_3\,|\,\text{类别}c)\bullet P(\text{类别}c)$$

$$\text{词}w\in\text{词集}\quad \text{类别}c\in\{\text{地理要素变化类，非地 理要素变化类}\} \tag{3}$$

7）比较6)中针对两个类别的乘积计算结果，较大项对应的类别即为待分类标题的类别。

8）跳转到4)，循环执行直至将所有候选网页处理完。

4 实 验

本文在Windows环境下，以Eclipse为开发平台，集成开源爬虫Heritrix、HTML解析器Jsoup，开发了网络地理要素变化新闻发现原型系统（如图3所示）。系统中的中文分词功能是利用自然语言处理与信息检索共享平台提供的NLPIR系统的API实现的。

图3 系统界面

Fig.3 System interface

为了对比本文方法与现有方法在发现的准确性和全面性上的表现，本文分别以发现公路的变化信息和发

现变电站的变化信息为例，设计了两组对比实验。发现的准确性和全面性分别以查准率和查全率来衡量。为了便于统计查全率，实验略去了候选网页的获取部分，在预先收集的测试数据集上进行发现测试。测试数据集来源于网络随机收集，包含5 000条网页文本数据，其中以公路变化为主题的数据168条，以变电站变化为主题的数据217条。

第一组实验的任务是公路要素变化发现，包含三个实验。实验一参照文献[4]中的方法，记为方法*A*，是基于改进的布尔模型来发现公路变化新闻。实验二参照文献[7]中的基于向量空间模型的方法，记为方法*B*。实验三则按照本文的发现方法，记为方法*C*。第二组实验的任务是变电站要素变化发现，实验设计与第一组类似。

第一组实验的结果见表1。

表1 第一组实验的结果

Tab.1 Result of the first group experiments

	实验一	实验二	实验三
自动提取到的公路变化信息数据	81	265	145
实际包含的公路变化信息数据	44	131	124
查准率	54.3%	49.4%	85.5%
查全率	26.2%	78.0%	73.8%

第二组实验的结果见表2。

表2 第二组实验的结果

Tab.2 Result of the second group experiments

	实验四	实验五	实验六
自动提取到的变电站变化信息数据	527	276	268
实际包含的变电站变化信息数据	181	126	176
查准率	34.3%	45.7%	65.7%
查全率	83.4%	58.1%	81.1%

统计每种方法在两组实验中的平均查准率和平均查全率，结果如图4所示。

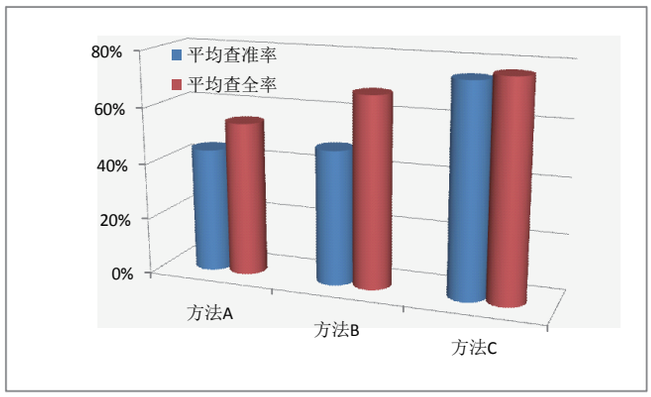

图4 三种方法的平均查准率和平均查全率对比

Fig.4 Comparison of three methods

分析实验一和实验四的结果可知，文献[4]中的提取方法查准率较低的主要原因在于其关键词的匹配方法对主题的锁定作用有限，如“2014年/*t*全区/*n*道路/*n*春运/*n*工作/*vn*圆满/*ad*完成/*v*”，既包含公路变化关键名词“道路”，又包含关键动词“完成”，但并非公路变化信息。该方法的查全率在两组实验中浮动较大，主要原因是关键词集由人工收集得到，词集的完备程度对查全率影响很大。

实验二和实验五参照文献[7]中的方法，该方法的指导思想是:在一个文本中出现次数很多的单词，在另一个同主题文本中出现次数也会很多，反之亦然。所以，如果特征空间坐标系取特征词的词频作为测度，就可以体现同主题文本的特点。但并不能说符合该特点的文本就一定恰好符合该主题，可能只是与该主题相关的文本，故其查准率不高。

实验三和实验六所采用方法的查准率查全率表现较前两种方法有所提高。其原因在于朴素贝叶斯分类模型的最大概率决策规则。只要正确类的概率比其他类高就可以得到正确的分类，所以尽管概率估计轻度的甚至是严重的不精确，但往往不影响其得到正确的结果。

两组对比实验表明，本文方法的平均查准率较现有方法有所提高，而在平均查全率上与表现较好的向量空间模型持平。

5 结束语

本文以网页文本作为数据源，提出并实现了一种地理要素变化发现的方法，改善了现有面向网页文本的发现方法在准确性上的不足。但该方法中也仍存在着一些不足：①本文中朴素贝叶斯分类器的训练文本需要人工收集和标定类别，工作量大。②方法在发现的准确性和全面性上有小幅度的起伏。后续研究将针对这两方面做进一步的改进。

参考文献

[1] 陈军, 赵仁亮, 王东华. 基础地理信息动态更新技术体系初探[J]. 地理信息世界, 2007, 14(5):4-9.

[2] 闫会杰, 赵巍. 服务于基础地理信息数据动态更新的网络蜘蛛[J]. 测绘技术装备, 2012(2):21-22.

[3] 王曙, 吉雷静, 张雪英, 等. 面向网页文本的地理要素变化检测[J]. 地球信息科学学报, 2013, 15(5):625-634.

[4] 吉雷静. 面向网页文本的地理信息变化语义检测方法研究[D]. 南京:南京师范大学, 2013.

[5] 曾文华, 黄桦. 基于网页信息检索的地理信息变化检测方法[J]. 计算机应用, 2010, 30(4):1132-1134.

[6] 张春菊, 张雪英, 朱少楠, 等. 基于网络爬虫的地名数据库维护方法[J]. 地球信息科学学报, 2011, 13(4):492-499.

[7] 武昊, 廖安平, 何超英, 等. 基于主题相关度的地理信息Web服务爬虫研究[J]. 地理与地理信息科学, 2012, 28(2):27-30.

[8] 郑炜, 沈文, 张英鹏. 基于改进朴素贝叶斯算法的垃圾邮件过滤器的研究[J]. 西北工业大学学报, 2010, 28(4):622-627.