农产品市场名称地理定位的设计与实现①

Design and Implementation of Market Location System

胡宜敏[1,2] 宋良图[1] 黄 河[1,2] 武民民[1,2] 黄 伟[1,2] (1.中科院合肥智能机械研究所 安徽 合肥 230031; 2.中国科学技术大学 信息科学技术学院 安徽 合肥 230026)

摘 要: 为了对农产品市场价格信息进行地理分析，将网络上采集的大量农产品市场名称与其地理位置关联起来成为需要。本文通过地理信息库的匹配和网络搜索引擎的搜索结果的统计，将农产品市场名称转化为该市场所在县及其以上级的地名，并映射成该地名唯一的行政区划代码。结果表明该设计能实际应用于实际市场名的地理位置标定。

关键词: 农产品市场 地理定位 地理信息库 搜索引擎 皮尔逊卡方检验

1 引言

网络中有大量农产品价格信息，为了将价格信息的区域性变化直观的表现出来，把这些数据和地理分布联系起来就是一件非常有意义的事情，农产品市场的地理定位对价格信息的智能检索，消除一个市场因书写习惯不同而判定为多个市场所造成的数据的重复有重要意义。网络上发布的农产品价格信息后会带有市场名称，于是通过对市场名称的地理定位就可以将价格信息同地理位置关联起来。但网络上农产品价格信息数据量十分巨大，而且市场名称不断变化，对市场名的地理位置进行人工标定工作量大、效率低，仅靠人眼难于发现而且十分枯燥。于是本文提出了一种由计算机完成市场名地理定位的设计。

本文中的地理定位是将农产品市场名称映射到市场所在地的行政区划代码。我国地名有很多别称和重复，但地名所对应的行政区划代码却是唯一的，通过将市场名称映射成行政区划代码就为市场名添加了唯一的标识。我国行政区划代码覆盖县及县级以上地名，于是可将市场名最小定位到县级地名。

在网络上的农产品市场名称没有统一标准，书写灵活度大而且不规范。这也使对市场名称的地理定位变得困难，不能只通过一种方法达到目的。本文设计和实现了一种通过本地处理和网络搜索相结合的方式来解决这个问题，过程分三个步骤：(1)基于地理信息库索引的市场名地理定位。(2)基于搜索引擎搜索结果项的市场名地理定位。(3)基于搜索引擎搜索结果数量的市场名地理定位。首先第一步对市场名中的地名信息进行抽取，通过与本地地理信息库进行匹配。在第一步不能定位的情况下，将市场名称提交给搜索引擎，判断搜索结果中的地名信息进行地理定位。在第二步得到多个候选地名出现频率相近时进行第三步操作，将候选地名和市场名称再次提交搜索引擎得到搜索结果的网页数量，通过皮尔逊卡方检验进行评价，定位市场名的地理位置。

2 基于地理信息库的市场名地理定位

从农产品市场名中直接提取县(市)级地名，和地理信息库进行匹配。对农产品市场名中的词与在建成索引形式的地理信息库进行检索，以匹配所需要的地名信息。

2.1 从网络上抓取的农产品市场名的的一些特征

网络上的农产品市场名没有统一规范，书写灵活度大[1]。例如：

a. “合肥周谷堆批发市场”

b. “周谷堆农贸市场”

c. “安徽合肥周谷堆蔬菜副食品批发市场 ”

① 基金项目:国家自然科学基金项目(60774096);国家十一五科技支撑计划项目(2006BAD10A1410);国家 863 计划项目(2006AA10Z237)
收稿时间:2008-08-27

这是同一个市场，却又多种写法，但对他们的处理方式是不一样的。最为理想的市场名为：省级地名＋市级地名＋县级地名＋乡镇名＋市场名，而从网络上采集的农产品市场名，某些部分是缺失的，大体可以分为两类，(1)市场名中本身含有我们需要县（市）级地名信息；(2)市场名中没有能直接提取的明确的地名信息，如只有省份，或者乡镇一级的地名信息，或仅有市场名，还有一些习惯写法书写的市场名和错别字。对于第(1)类可以使用基于地理信息库索引的地理定位，而对于第(2)类，使用网络搜索引擎的搜索结果中的地名和搜索结果的网页数量信息进行定位。

2.2 地理信息库索引

本文的设计有两个地名表：(1)省级地名表，包括省级地名名称、简称、别名和行政区划代码；(2)全国县(市)级地名表，包括县(市)级地名名称、别称和其行政区划代码。地名表的建立是一项非常重要的工作，地理定位的成功率很大程度上取决于地理信息库的建立状况。对于别称的建立，例如："北京市" 别称 "北京"，本文利用中文分词的词性标注功能，对一个地名字符串，从左到右逐次增加一个字符，直到分词系统识别出第一个地名，便存为别名。当然有些地名的别称是计算机无法处理的，必须人工建立，例如："紫云苗族布依族自治县" 别称 "紫云县"。在地名识别过程中，需要大量从地理信息库匹配地名和行政区划代码信息，从数据库中匹配地理信息速度对于大规模进行市场名定位来说比较慢，无法达到使用要求，所以为了提高匹配速度，将地理信息库建成索引形式便成为一种好的选择，本文采用 Lucene[2]建立索引，从索引文件中检索地理信息的速度达到了使用要求。

2.3 基于地理信息库的市场名地理定位过程描述

(1) 提取农产品市场名中的所有两个字以上的连续词。对市场名进行循环取词，将市场名中含有的地名信息提取出来，例如：地名 "新疆米泉"，就会取出 "新疆"、"新疆米"、"新疆米泉"、"疆米"、"疆米泉"、"米泉" 6 个词。对于长度较短的市场名，词组的生成速度较快，而对于名字较长的市场名，则相对较慢，但这种方式的最大优点是能全面的提取到地名信息，不会出现遗漏的现象。但当市场名中出现多个地名时就难于定位，由于中国地名重复的很多，有些乡镇的名字可能和城市名字相同，城市道路的名字可能是其它城市的名字，这必须通过后续处理屏蔽或者减小这种干扰。

(2) 提取市场名所在省级地名。人们习惯在书写市场名时总会先写省份的名字，通常在市场名字符串最左面，所以省份名称最先提取出来。将第(1)步取出的词组按从左到右的顺序依次匹配省份地名库，找到第一个省份地名后，匹配停止。读取得到省份的行政区划代码，并取出前两位。因为同一个省内的地名的行政区划代码的前两位是相同的，通过它来限定后面县(市)级地名匹配的范围，可以消除其他省份的地名对该市场名定位的噪声。

(3) 取出市场名中可能的县(市)级地名。这里可以分为有省份和无省份两种情况：如果省份缺失，就从左到右提取市场名称字符串第一个地名作为定位地名。因为根据书写习惯，越靠近市场名字符串左侧的地名行政级别越大，虽然有可能会没有提取到最小的县(市)级地名，但不会出现地理定位的错误，例如："马鞍山安民市场" 如果从左到右提取最后一个地名，便是 "鞍山"，而不是 "马鞍山"；如果有省份，则从右向左提取市场名字符串第一个地名作为定位地名，取到地名后和刚取到的省份的地理信息库进行匹配，并且保证提取到的地名后不会带 "路"、"街" 等字样，这样可以保证取到县级地名而又消除其他省份的地名的噪声影响[3]。

(4) 经过上述步骤，如果得到符合规则的地名，则定位成功，反之进行基于搜索引擎搜索结果的市场名地理定。

其具体过程如图 1 所示：

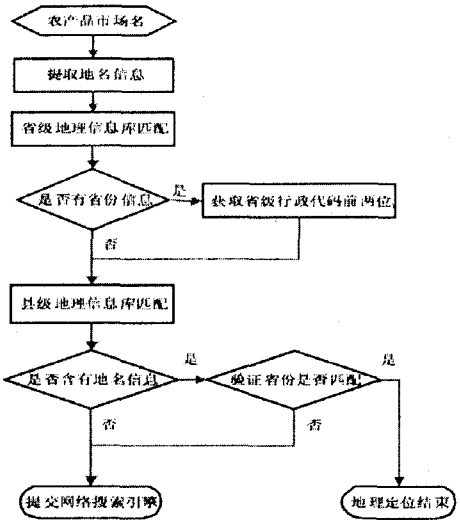

图 1　基于地理信息库的市场名地理定位过程

3　基于搜索引擎搜索结果项的市场名地理定位

从网页上采集的市场名中绝大多数地理信息比较完整，这种情况下只要建立一个比较完备的地理信息库，进行词的匹配就可以对市场名进行地理定位，然而由于有的网站用户在某一地域较多，发布信息时有默认的地理位置，从而导致市场名没有相应的地理信息，这种情况下对市场名的地理定位就通过网络搜索引擎来进行检索。通过搜索引擎对市场名检索，市场名和它所在的地名一起出现在搜索结果的同一网页中的可能性很大。搜索结果就形成了由许多独立网页组成的语料库，通过分析语料库中的地理信息来判断市场名和地名之间的联系，从而定位市场的所在地。

3.1 中文分词技术

本文的设计采用中文分词系统 ICTCLAS，相对逐字循环取词，分词处理速度更快，对一些地名的切分比较准确，所以在处理大文本的语料时，相当有优势，但由于地名非常多而且复杂，有些地名中文分词无法正确切分，影响了定位的准确性。循环取词再和地理信息库进行匹配，理论上能更准确的提取出地名，但对大文本来说速度慢，达不到使用要求。在对市场名称匹配地理信息库时采用了逐字循环取词的方法，而对于网络搜索引擎检索的结果得到的大文本的分析，采用了中文分词。

3.2 基于搜索引擎搜索结果的市场名地理定位过程描述

(1) 将该市场名提交给搜索引擎，返回关于该市场的搜索结果，提取一定数量的结果的文本信息，并将其获取到本地。

(2) 通过规则表达式去除非搜索项相关的其他信息，如广告等，并且去除非文字信息，以减少对后续分析噪声干扰。

(3) 对以上得到的语料库进行中文分词，对地名进行标定，将文本中的地名按出现频率从高到低进行排序，得到在语料库中出现地名的频率表。

(4) 按地名出现次数从高到低，将获取的地名和地理信息库进行匹配，如果市场名中存在省级地名，就只和该省的地名信息进行匹配，加快了处理速度，并滤除该省以外地名的干扰。

(5) 经上步操作得到经过筛选的出现频率由高到低的一组地名，如果这时出现频率最高的地名的频率高于后面地名出现频率达一个阈值，则判定该地名为该市场所要定位的结果，反之就进行基于搜索引擎搜索结果数量的市场名地理定位。其具体过程如图 2 所示：

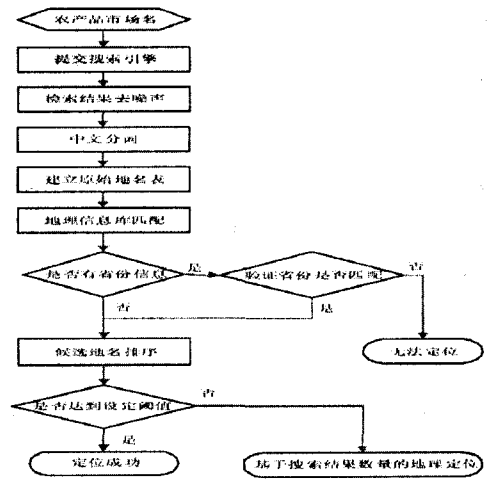

图 2　基于搜索引擎搜索结果的市场名地理定位过程

4　基于搜索引擎搜索结果数量的市场名地理定位

通过搜索引擎检索不仅能得到简单的包含检索内容的搜索结果，还能够进行高级搜索，比如在网页上同时包含若干个词和不出现某个词的检索，并且提供了搜索到的网页数量。为了判断在市场名称和地名同时出现在一个网页上是不是仅仅是偶然现象，本设计使用假设检验来进行评价。在经过上步处理得到的候选地名如果出现的频率相近没有达到阈值要求，便无法准确判断结果，这时利用搜索引擎提供的高级搜索服务得到的检索结果数量，再通过皮尔逊卡方检验来对各个地名和市场名的独立性进行评价。

4.1 皮尔逊卡方检验[4]

X^2 统计量计算了观测值和期望值之间差别的总和，并且将期望值作为比例因子，得到：

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \qquad (1)$$

其中 i 表示出现次数依赖关系表的行变量，j 表示列变量，O_{ij} 表示表单元(i，j)的观测值，E_{ij} 表示期望值。通过计算边缘分布可以得到期望频度 E_{ij} 的值，计算方法不复杂，将出现频率转化为比例值后分别按行和列计算总和即可。表 1 给出了候选地名和市场名的分布关系。

表 1 地名和市场名出现次数之间的依赖关系的 2×2 的表

	W1=market	W1≠market
W2=place	O_{11}	O_{12}
W2≠place	O_{21}	O_{22}

在表 1 中 place 是候选地名，market 是市场名。在搜索引擎搜索得到的结果中，市场名和候选地名同时出现在同一网页中的数量为 O_{11}，是 place∩market 的搜索结果数量；在同一网页中出现候选地名但不出现市场名的网页数量为 O_{12}，是 place∩$\overline{market}$ 的搜索结果数量；在同一网页中出现市场名但不出现候选地名的网页数量为 o_{21}，是 $\overline{place}$ ∩market 的搜索结果数量；o_{22} 是在同一网页中既不出现市场名又不出现候选地名的网页数量。

根据公式(1)的推导，x^2 检验应用于 2×2 表的表达形式为：

$$X^2=\frac{N(O_{11}O_{22}-O_{12}O_{21})^2}{(O_{11}+O_{12})(O_{11}+O_{21})(O_{12}+O_{22})(O_{21}+O_{22})} \quad (2)$$

N 为该搜索引擎所有的中文网页数量。

4.2 基于搜索引擎搜索结果数量的市场名地理定位过程描述

(1) 统计出在文本中出现频率最高的前若干个地名{ place1、place2、…placeN}和市场名 market，将{place1∩market,place2∩market，…placeN∩market},{place1 ∩ $\overline{market}$,place2 ∩ $\overline{market}$ ，… placeN ∩ $\overline{market}$ },{ $\overline{place1}$ ∩ market, $\overline{place1}$ ∩ market，… $\overline{place1}$ ∩market}提交给搜索引擎进行检索，返回搜索结果的数量。

(2) 由公式(2)根据相应的网页数量计算各个地名的 X^2，选择对应 X^2 值最大的地名作为定位结果。X^2 的值越大，就以越高的置信度确认农产品市场和该候选地名的相互独立性的零假设是不成立的，将该农产品市场定位到该地名是合理的。

5 实验结果

对从全国农业网站上采集的 10000 个农产品市场名中随机抽取 1000 个进行测试，地名频率阈值采用 0.8,即当 f2/f1<0.8 时，确定频率数最高的地名为所需要的地名，反之则进行基于搜索引擎搜索结果数量的市场名地理定位，其中 f2 为排名第二的地名的频率数，f1 排名第一的候选地名的频率数。市场名地理定位的正确率(P)，召回率(R)和 F 值分别定义为：

P=(正确定位的市场数/定位的市场数)×100%；

R=(正确定位的市场数/实际市场数)×100%；

$$F=\frac{(1+\beta^2)RP}{R+\beta^2P}\times 100\%;$$

F 是计算正确率和召回率的综合指标，为正确率和召回率的平衡因子,通常认为正确率和召回率同等重要,因此取 $\beta=1$[5]。以下为分别经过基于地理信息库索引的市场名地理定位、基于搜索引擎搜索结果项的市场名地理定位、基于搜索引擎搜索结果数量的市场名地理定位三个步骤后的结果见表 2:

表 2　农产品市场名地理定位实验结果

	实际市场数	定位市场数	正确定位数	P/%	R/%	F/%
步骤 1	1000	910	910	100%	91.0%	95.3%
步骤 2	1000	971	953	98.1%	95.3%	96.7%
步骤 3	1000	1000	974	97.4%	97.4%	97.4%

该结果基本满足使用要求，在依次经过三个步骤召回率不断提高但正确率有所下降，错误定位的市场名主要是含有错别字，地理位置书写不够明确而导致地名重名的市场名称，如果将类似的市场名滤除，还会提高定位的正确率。在将要搜索的内容提交给网络搜索引擎时，可能会提交给不同的服务器，而不同的服务器上的数据量不一定一致，所以在选择搜索引擎时选取一些服务器间数据量变化小的也会提高定位的正确率。

6 结束语

本文介绍了在基于地理信息库匹配和网络搜索引擎检索的结果及其数量的基础上，综合各种方法的优点，提出了一种对农产品市场名称进行地理定位的设计。实验结果满足了对农产品市场名称地理定位的使用要求。

参考文献

1 黄德根,孙迎红.中文地名的自动识别.计算机工程,2006,32(3):220－222.

2 高红,黄德根,杨元生. 汉语自动分词中中文地名识别.大连理工大学学报,2006,46(4):576－581.

3 Manning CD, Sch ütze H. Foundations of Statical Language Processing: The MIT Press, 1999:95－116.

4 曾文,鄢军霞.城市 GIS 地名定位工具的设计与应用.中国地质大学学报,2006,31(5):725－728.

5 Hatcher E, Gospodnetic O. Lucene in action: New York Manning, 2005:68－230.