(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 103605752 A
(43)申请公布日 2014.02.26

(21)申请号 201310596806.7

(22)申请日 2013.11.21

(71)申请人 武大吉奥信息技术有限公司
地址 湖北省武汉市东湖开发区庙山小区江夏大道武大科技园

(72)发明人 黄俊韬 魏延峰 吴杰 赵雷雷 刘琳 刘勇 肖豪 邓跃进 宋爱红 范业稳 朱伟奇 张龙 陈胜鹏 程方 贺楷锴 许振华

(74)专利代理机构 北京天奇智新知识产权代理有限公司 11340
代理人 刘黎明

(51)Int.Cl.
G06F 17/30(2006.01)

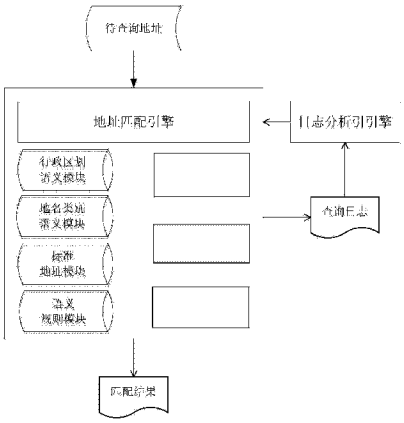
权利要求书2页 说明书7页 附图1页

(54)发明名称

一种基于语义识别的地址匹配方法

(57)摘要

本发明公开了一种基于语义识别的地址匹配方法，所述方法包括地址匹配引擎和日志分析引擎，地址匹配引擎包括行政区划语义模块、地名类别语义模块、标准地址模块、语义规则模块、中文分词模块、语义识别模块、查询模块。所述方法根据用户输入的待检索地址通过基于语义识别的地址匹配引擎快速、准确地查找到匹配的地址，并以在线服务的形式返回用户检索结果，日志分析引擎记录并分析查询日志，根据日志分析结果优化地址匹配引擎。



CN 103605752 A

1. 一种基于语义识别的地址匹配方法，其特征在于，所述方法包括地址匹配引擎和日志分析引擎，所述地址匹配引擎根据用户输入的待检索地址基于语义识别快速、准确地查找到匹配的地址，并以在线服务的形式返回用户检索结果，所述日志分析引擎记录并分析查询日志，根据日志分析结果优化地址匹配引擎。

2. 根据权利要求 1 所述的方法，其特征在于：所述地址匹配引擎包括行政区划语义模块、地名类别语义模块、语义规则模块、标准地址模块、语义规则模块、中文分词模块、语义识别模块、查询模块，所述地址匹配引擎的运行步骤如下，

（1）利用行政区划语义模块建立行政区划语义库；

（2）利用地名类别语义模块建立地名类别语义库；

（3）利用标准地址模块建立规范化的具体地址库；

（4）利用语义规则模块建立基于语义的地址检索方法；

（5）利用中文分词模块对待查地址进入中文分词；

（6）利用语义识别模块对分词得到的词元进行语义识别；

（7）利用查询模块对识别后的词元基于语义方法进行查询。

3. 根据权利要求 2 所述的方法，其特征在于：在步骤（1）中，所述行政区划语义库模块中，行政区划以其国家标准编码为基本信息，建立相应的行政区划语义库，用于地址匹配后续过程的行政区划语义匹配，行政区划是界定地址行政范围的主要属性之一，也是地址的基本语义之一，在进行地址匹配时，行政区划语义是优先考虑的语义匹配。

4. 根据权利要求 2 所述的方法，其特征在于：在步骤（2）中，所述地名类别语义库模块，描述了地名地址所归属的分类，是地址的基本语义之一，其中，地名类别编码是多级编码的方式。

5. 根据权利要求 2 所述的方法，其特征在于：步骤（3）中，所述具体地址库模块对不同区域的地址数据进行规范化设计，准确地进行地址匹配。

6. 根据权利要求 2 所述的方法，其特征在于：在步骤（4）中，所述基于语义的地址检索方法模块，建立基于语义的地址检索规则和规则间的关系，形成地址检索的知识库，具体方法为

a、当词元有且仅有一个，并且该词元类型为“行政区划”时，进行行政区划查询，返回行政区划的信息；

b、当词元有多个且所有词元都是“行政区划”时，进行行政区划查询，返回地区级别最小的那个行政区划信息；

c、当词元有多个且所有词元都是“行政区划”时，并且行政区划不是上下级关系，应取第一个行政区划做范围，其他行政区划做关键字进行普通地名查询，返回位置在行政区划内，且名称包含指定关键字的地名；

d、当词元有且仅有一个，并且该词元类型为“类别”时，按照类别进行查询；

e、当词元中既有“类别”，也有“行政区划”的时候，按照规则 c 得到行政区划和关键字，在指定类别中查找地名；

f、当词元既不包含“行政区划”类的词，也不包含“类别”类的词，将这些词当做关键字查找地名；

g、当词元既包含“行政区划”类的词，同时包含关键字，则在行政区划内按指定关键字

进行查询；

h、当词元既包含多个“行政区划”类的词，并且行政区划不是同一区域，或者是第二个行政区划大于第一行政区划，包含关键字，第一个当行政区划，后面当做关键字处理；

i、当词元有多个且所有词元都是“地名类型”时，按照排列的先后顺序，排在最后的为“地名类型”，其他词作为关键字进行查询

g、当包含“门牌”类型词元时，按以上规则构建查询条件，并进行地址查询。

7. 根据权利要求 2 所述的方法，其特征在于：步骤（5）中对待查地址进入中文分词模块，对于用户输入的以自然语言形式表示的待查地址采用成熟的中文分词算法，并将行政区划语义库、地名类别语义库中的数据纳入到用于中文分词的字典中，以改进中文分词算法的查准性、查全性，适当减少中文分词算法中存在的语义歧义问题，分词过程中支持同义词典，分词后得到地址词元。

8. 根据权利要求 2 所述的方法，其特征在于：步骤（6）中，对分词得到的词元进入语义识别模块，对于分词后得到的地址词元，根据行政区划语义库、地名类别语义库以及专家知识对词元进行语义识别，确定每个词元匹配的数据库表范围，避免大范围地检索无关联的地址数据，以减少数据库检索的时间。

9. 根据权利要求 2 所述的方法，其特征在于：所述步骤（7）中，对识别后的词元基于语义规则进入查询模块，在具体地址库中采用数据库查询语言对识别后的词元基于语义规则进行查询，返回查询结果给用户。

10. 根据权利要求 1 所述的方法，其特征在于：所述日志分析引擎记录地址匹配日志，并分析日志，将分析结果反馈给地址匹配引擎，优化地址匹配引擎，也能够导入其他外部系统的日志库或知识库，并利用其优化地址匹配引擎。

一种基于语义识别的地址匹配方法

技术领域

[0001] 本发明属于对地观测与导航技术领域，具体涉及一种基于语义识别的地址匹配方法。

背景技术

[0002] 地址检索、查询服务是网络地图在线服务的重要功能。地址匹配是将文字性的描述地址与其空间的地理位置坐标建立起对应关系的过程，其目的是要根据用户输入的待检索地址快速查找到匹配的地址，并以在线服务的形式返回用户检索结果。传统的方法通常采用基于关键词的精确或模糊匹配方法，这对于大规模或大范围的地名地址数据，不仅查找的速度慢，很难满足网络地图在线快速服务的需要，也没有顾及地址的语义信息，导致查找的准确性比较低，查找结果多样且往往不是用户所需要的结果。例如，当用户在互联网上查找“上海南京路”时，期望的返回结果应该是位于上海，名称为南京路的道路，但普通的查找方法可能会将南京的上海路和上海的南京路都作为结果返回。因此，针对上述两个问题，发明了一种基于语义识别的地址匹配方法，可以有效地提高地址数据查找的速度和准确性，从而提高网络地图在线服务质量，为用户提供良好的服务检验。

发明内容

[0003] 针对上述现有技术中的不足，本发明的目的在于提供一种基于语义识别的地址匹配方法。对于以自然语言形式表示的地址信息，通过中文分词技术，并顾及地址的语义建立用于地址匹配的语义库或知识库，然后根据地址数据表达的语义特点，建立地址匹配的规则，通过适当的匹配算法提高地址检索的速度和准确性。

[0004] 为了实现上述发明目的，本申请提供了以下技术方案：

[0005] 一种基于语义识别的地址匹配方法，所述方法包括地址匹配引擎和日志分析引擎，所述地址匹配引擎根据用户输入的待检索地址基于语义识别快速、准确地查找到匹配的地址，并以在线服务的形式返回用户检索结果，所述日志分析引擎记录并分析查询日志，根据日志分析结果优化地址匹配引擎。地址匹配引擎包括行政区划语义模块、地名类别语义模块、语义规则模块、标准地址模块、语义规则模块、中文分词模块、语义识别模块、查询模块。所述方法步骤如下：

[0006] （1）利用行政区划语义模块建立行政区划语义库；

[0007] （2）利用地名类别语义模块建立地名类别语义库；

[0008] （3）利用标准地址模块建立规范化的具体地址库；

[0009] （4）利用语义规则模块建立基于语义的地址检索规则；

[0010] （5）利用中文分词模块对待查地址进入中文分词；

[0011] （6）利用语义识别模块对分词得到的词元进行语义识别；

[0012] （7）利用查询模块对识别后的词元基于语义规则进行查询；

[0013] （8）利用日志分析引擎记录地址匹配日志，分析日志，将分析结果反馈给地址匹配

引擎，优化地址匹配引擎。

[0014] 在步骤（1）中，所述行政区划语义模块中，行政区划以其国家标准编码为基本信息，建立相应的行政区划语义库，用于地址匹配后续过程的行政区划语义匹配，行政区划是界定地址行政范围的主要属性之一，也是地址的基本语义之一，在进行地址匹配时，行政区划语义是优先考虑的语义匹配。

[0015] 在步骤（2）中，所述地名类别语义模块，描述了地名地址所归属的分类，是地址的基本语义之一，其中，地名类别编码是多级编码的方式。

[0016] 步骤（3）中，所述标准地址模块对不同区域的地址数据进行规范化设计，准确地进行地址匹配。

[0017] 在步骤（4）中，所述语义规则模块，建立基于语义的地址检索规则和规则间的关系，形成地址检索的知识库，具体方法为

[0018] a、当词元有且仅有一个，并且该词元类型为“行政区划”时，进行行政区划查询，返回行政区划的信息；

[0019] b、当词元有多个且所有词元都是“行政区划”时，进行行政区划查询，返回地区级别最小的那个行政区划信息；

[0020] c、当词元有多个且所有词元都是“行政区划”时，并且行政区划不是上下级关系，应取第一个行政区划做范围，其他行政区划做关键字进行普通地名查询，返回位置在行政区划内，且名称包含指定关键字的地名；

[0021] d、当词元有且仅有一个，并且该词元类型为“类别”时，按照类别进行查询；

[0022] e、当词元中既有“类别”，也有“行政区划”的时候，按照规则 c 得到行政区划和关键字，在指定类别中查找地名；

[0023] f、当词元既不包含“行政区划”类的词，也不包含“类别”类的词，将这些词当做关键字查找地名；

[0024] g、当词元既包含“行政区划”类的词，同时包含关键字，则在行政区划内按指定关键字进行查询；

[0025] h、当词元既包含多个“行政区划”类的词，并且行政区划不是同一区域，或者是第二个行政区划大于第一行政区划，包含关键字，第一个当行政区划，后面当做关键字处理；

[0026] i、当词元有多个且所有词元都是“地名类型”时，按照排列的先后顺序，排在最后的为“地名类型”，其他词作为关键字进行查询

[0027] g、当包含“门牌”类型词元时，按以上规则构建查询条件，并进行地址查询。

[0028] 在步骤（5）中，对待查地址进入中文分词模块，对于用户输入的以自然语言形式表示的待查地址采用成熟的中文分词算法，并将行政区划语义库、地名类别语义库中的数据纳入到用于中文分词的字典中，以改进中文分词算法的查准性、查全性，适当减少中文分词算法中存在的语义歧义问题，分词过程中支持同义词典，分词后得到地址词元。

[0029] 步骤（6）中，对分词得到的词元进入语义识别模块，对于分词后得到的地址词元，根据行政区划语义库、地名类别语义库以及专家知识对词元进行语义识别，确定每个词元匹配的数据库表范围，避免大范围地检索无关联的地址数据，以减少数据库检索的时间。

[0030] 所述步骤（7）中，对识别后的词元基于语义规则进入查询模块，在具体地址库中采用数据库查询语言对识别后的词元基于语义规则进行查询，返回查询结果给用户。

[0031] 所述步骤(8)中，日志分析引擎记录地址匹配日志，并分析日志，将分析结果反馈给地址匹配引擎，优化地址匹配引擎，也能够导入其他外部系统的日志库或知识库，并利用其优化地址匹配引擎。

[0032] 优选方案为：

[0033] 首先，为了保证地址查找服务能够匹配到所需要的结果，保证检索的查全性，需要建立符合标准的完整地址库。为了达到良好的地址匹配效果，本发明对地址库进行了以下优化设计：

[0034] 1)对于以自然语言形式表示的地址进行语义分析，将语义信息分为行政区划语义、地名类别语义和具体地址三类；行政区划语义表示了地址所归属的行政区划范围，如湖北省。地名类别语义表示了地址所属的类别，如行业性质类别，比如快餐、超市、大学。具体地址为地址信息中不能归于行政区划语义、地名类别语义的地址语义部分，如测绘大厦。

[0035] 2)依据上述语义信息类别，分别建立多层级的语义数据库表，包括行政区划语义库、地名类别语义库、和根据行政区划与地名类别分类的多个具体地址库。

[0036] 3)对上述语义库进行了规范化设计，例如，行政区划采用国家标准编码，可支持到街道、村级的编码。地名类别和具体地址参考测绘行业标准《地理信息公共服务平台地理实体与地名地址数据规范》(CH/Z9010-2011)及相关的国内、国标标准进行规范化设计，给出了设计原则和数据库表结构，具体设计在系统实现时完成，可以满足不同系统的地址检索需求。

[0037] 其次，在传统的地址检索方法的基础上，根据地址检索的经验，通过访问专家和典型用户建立了基于语义的地址检索规则和规则间的关系，形成地址检索的知识库。

[0038] 然后，对用户输入的需要查找的地址进行中文分词和语义识别。中文分词采用较成熟的分词算法，但需要顾及上面描述的语义信息分类，也就是基于行政区划语义、地名类别语义和具体地址对用户描述的地址进行分词，划分成为基本的地址词元或关键词。然后对地址词元进行语义识别，判断地址词元属于行政区划语义、地名类别语义还是具体地址。语义识别依据地址词元的性质、词元的关系等知识，通过语义匹配算法实现。

[0039] 最后，通过建立地址匹配规则，分别对语义识别后的地址词元进行匹配，返回查找结果，记录地址匹配日志，并分析日志，将分析结果反馈给地址匹配引擎，优化地址匹配引擎。

[0040] 有益效果

[0041] 1、先通过对地址进行语义分析与识别，对识别后的地址词元快速定位于相应的语义库，并在该语义库中进行关键词匹配。由于分类后的语义库规模较没有进行分别的地址规模小，提高了关键词匹配的速度，而定位语义库的时间很短，从而整体上可以获得较高的地址查找速度。同时，由于在进行中文分词时顾及了地址的语义，使得查找的结果更能体现用户的意思，有利于提高查找的准确性。

[0042] 2、本发明通过建立基于语义的匹配规则，充分采用地址匹配的经验知识，提高了地址匹配算法的效率。

附图说明

[0043] 图 1 是基于语义识别的地址匹配方法示意图。

具体实施方式

[0044] 具体实施方式如下：

[0045] <一>建立行政区划语义库

[0046] 行政区划是界定地址行政范围的主要属性之一，也是地址的基本语义之一。在进行地址匹配时，行政区划语义是优先考虑的语义匹配。

[0047] 行政区划以其国家标准编码为基本信息，建立相应的行政区划语义库，用于地址匹配后续过程的行政区划语义匹配。行政区划语义库的表结构如表 1 所示。

[0048] 表 1 行政区划语义库的表结构

[0049]

字段名称	数据类型	允许为空	描述
GBCODE	VARCHAR2(11)	×	行政区划国标码
GBNAME	VARCHAR2(255)	×	行政区名称
GB_LI_NAME	VARCHAR2(255)	√	行政区划简称
ZIP	VARCHAR2(6)	√	邮政编码
X	NUMBER(38,8)	√	地名位置 X 坐标
Y	NUMBER(38,8)	√	地名位置 Y 坐标

[0050] <二>建立地名类别语义库

[0051] 地名类别描述了地名地址所归属的分类，也是地址的基本语义之一。本发明设计了地名类别语义库的数据表结构如表 2 所示。其中，地名类别编码可以是多级编码的方式。不同用户在系统实施时可根据设计的规则自定义具体的地名类别库，以满足不同的系统要求。

[0052] 表 2 地名类别语义库的数据表结构

[0053]

字段名称	数据类型	允许为空	描述
ID	NUMBER(38)	×	主键
CODE	NUMBER(16)	×	编码
NAME	VARCHAR2(128)	×	分类名称

[0054] <三>建立规范化的具体地址库

[0055] 参考测绘行业标准《地理信息公共服务平台地理实体与地名地址数据规范》（CH/Z9010-2011）和其它相关国内、国际标准，对不同区域的地址数据进行规范化设计，以有利于准确地进行地址匹配。规范化的具体地址库表结构如表 3 所示。

[0056] 表 3 规范化的具体地址库表结构

[0057]

字段名称	数据类型	允许为空	描述
OID	NUMBER(38)	×	要素信息编号
DOMAINNAME	VARCHAR2(512)	×	简要地址信息（兴趣点名）
ADDCODE	VARCHAR2(20)	√	地址代码
X	NUMBER(38,10)	×	地名位置 X 坐标
Y	NUMBER(38,10)	×	地名位置 Y 坐标
CLASSID	VARCHAR2(16)	√	地名类别编码
STANDARDNAME	VARCHAR2(512)	×	详细地名信息
GBCODE	VARCHAR2(11)	×	行政区划国标码
ADDRESS	VARCHAR2(512)	√	地址信息（含门牌

[0058]

			号）
CONTINENT	VARCHAR2(255)	√	洲
COUNTRY	VARCHAR2(255)	√	国家
PROVINCE	VARCHAR2(255)	√	省
CITY	VARCHAR2(255)	√	市
DISTRICT	VARCHAR2(255)	√	区
TOWN	VARCHAR2(255)	√	镇、乡
STREET	VARCHAR2(255)	√	街道
STREET_NUMBER_PREFIX	VARCHAR2(255)	√	街道门牌号前缀
STREET_NUMBER	NUMBER(6)	√	门牌号
STREET_NUMBER_SUFFIX	VARCHAR2(255)	√	门牌附加信息
BUILDING_NUMBER	VARCHAR2(255)	√	楼牌号
BUILDING_SUFFIX	VARCHAR2(255)	√	楼牌附加信息

[0059] ＜四＞建立基于语义的地址检索规则

[0060] 根据地址检索的经验，通过访问专家和典型用户建立了基于语义的地址检索规则和规则间的关系，形成地址检索的知识库。部分规则如下：

[0061] （1）当词元有且仅有一个，并且该词元类型为“行政区划”时，进行行政区划查询，

返回行政区划的信息；

[0062] 例如：湖北省，湖北。

[0063] （2）当词元有多个且所有词元都是“行政区划”时，进行行政区划查询，返回地区级别最小的那个行政区划信息；

[0064] 例如：湖北省武汉市，湖北武汉。

[0065] （3）当词元有多个且所有词元都是“行政区划”时，并且行政区划不是上下级关系，应取第一个行政区划做范围，其他行政区划做关键字进行普通地名查询，返回位置在行政区划内，且名称包含指定关键字的地名；

[0066] （4）当词元有且仅有一个，并且该词元类型为“类别”时，按照类别进行查询；

[0067] 例如：快餐，超市。

[0068] （5）当词元中既有“类别”，也有“行政区划”的时候，按照规则 3 得到行政区划和关键字，在指定类别中查找地名；

[0069] 例如：武汉超市。

[0070] （6）当词元既不包含“行政区划”类的词，也不包含“类别”类的词，将这些词当做关键字查找地名；

[0071] 例如：眼镜。

[0072] （7）当词元既包含“行政区划”类的词，同时包含关键字，则在行政区划内按指定关键字进行查询；

[0073] 例如：武汉眼镜。

[0074] （8）当词元既包含多个“行政区划”类的词，并且行政区划不是同一区域，或者是第二个行政区划大于第一行政区划，包含关键字，第一个当行政区划，后面当做关键字处理；

[0075] 例如：武汉湖南眼镜，武汉湖北眼镜。

[0076] （9）当词元有多个且所有词元都是“地名类型”时，按照排列的先后顺序，排在最后的最为“地名类型”，其他词作为关键字进行查询。

[0077] 例如：酒店停车场。

[0078] （10）当包含“门牌”类型词元时（门牌前缀、门牌、门牌后缀、楼牌），按以上规则构建查询条件，并进行地址查询。

[0079] ＜五＞对待查地址进行中文分词

[0080] 对于用户输入的以自然语言形式表示的待查地址采用成熟的中文分词算法，并将行政区划语义库、地名类别语义库中的数据纳入到用于中文分词的字典中，以改进中文分词算法的查准性、查全性，适当减少中文分词算法中存在的语义歧义问题。分词过程中支持同义词典。分词后得到地址词元。

[0081] ＜六＞对分词得到的词元进行语义识别

[0082] 对于分词后得到的地址词元，根据行政区划语义库、地名类别语义库以及专家知识对词元进行语义识别，确定每个词元匹配的数据库表范围，避免大范围地检索无关联的地址数据，以减少数据库检索的时间。

[0083] ＜七＞对识别后的词元基于语义规则进行查询

[0084] 在具体地址库中采用数据库查询语言对识别后的词元基于语义规则进行查询，返回查询结果给用户。

[0085] ＜八＞根据日志分析结果优化地址匹配引擎

[0086] 根据查询结果和用户反馈信息，记录地址匹配日志，并分析日志，将分析结果反馈给地址匹配引擎，优化地址匹配引擎。也能够导入其他外部系统的日志或知识库，并利用其优化地址匹配引擎。

[0087] 最后应说明的是：显然，上述实施例仅仅是为清楚地说明本申请所作的举例，而并非对实施方式的限定。对于所属领域的普通技术人员来说，在上述说明的基础上还可以做出其它不同形式的变化或变动。这里无需也无法对所有的实施方式予以穷举。而由此所引申出的显而易见的变化或变动仍处于本申请型的保护范围之中。

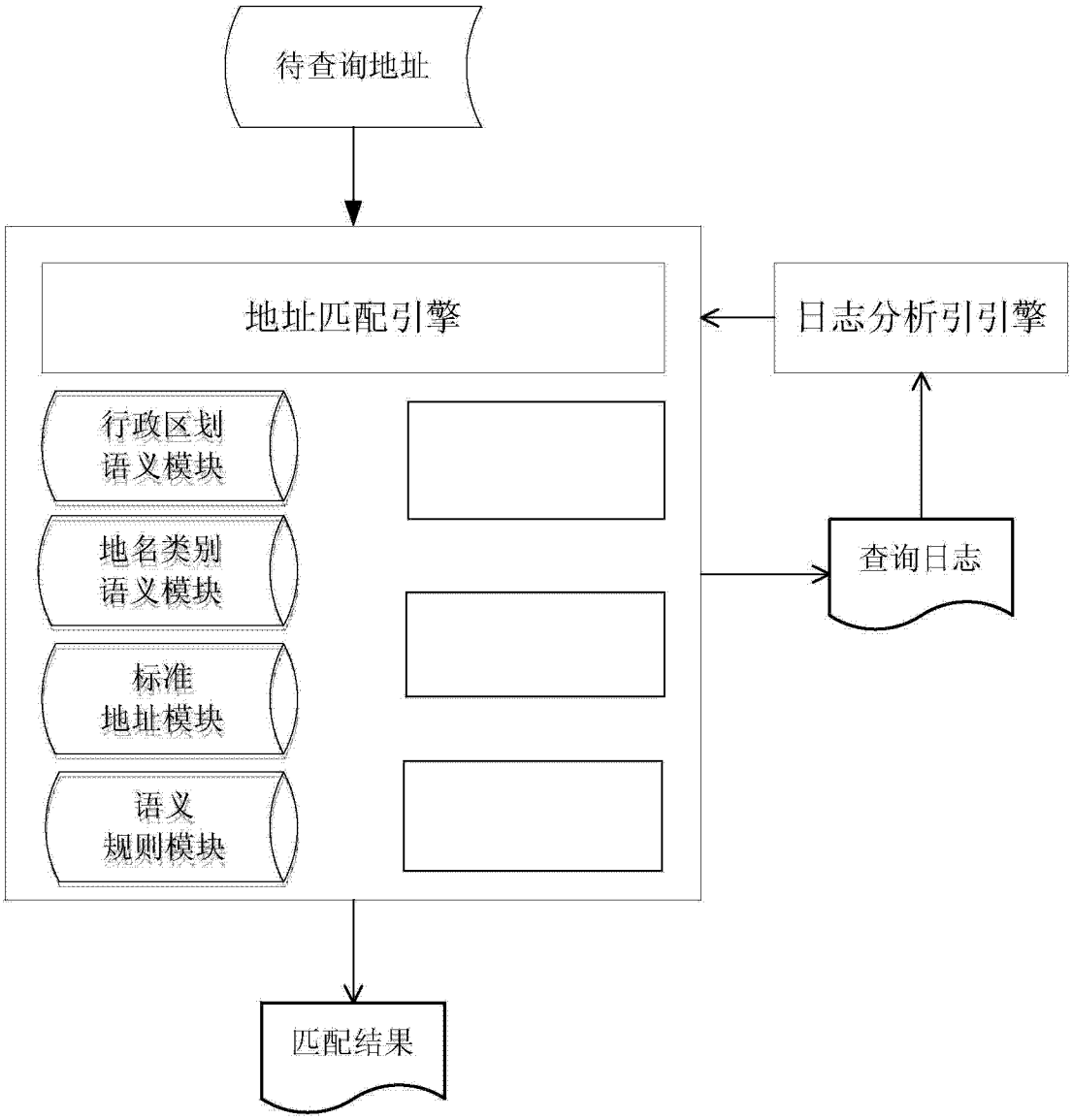

待查询地址
地址匹配引擎
日志分析引引擎
行政区划
语义模块
地名类别
语义模块
标准
地址模块
语义
规则模块
查询日志
匹配结果

图 1