



海量语义数据处理与知识服务

Scalable Semantic Data Processing and Knowledge Services

荷兰阿姆斯特丹自由大学

黄智生

Zhisheng Huang

VU University Amsterdam

The Netherlands

huang@cs.vu.nl



海量语义数据处理计算策略

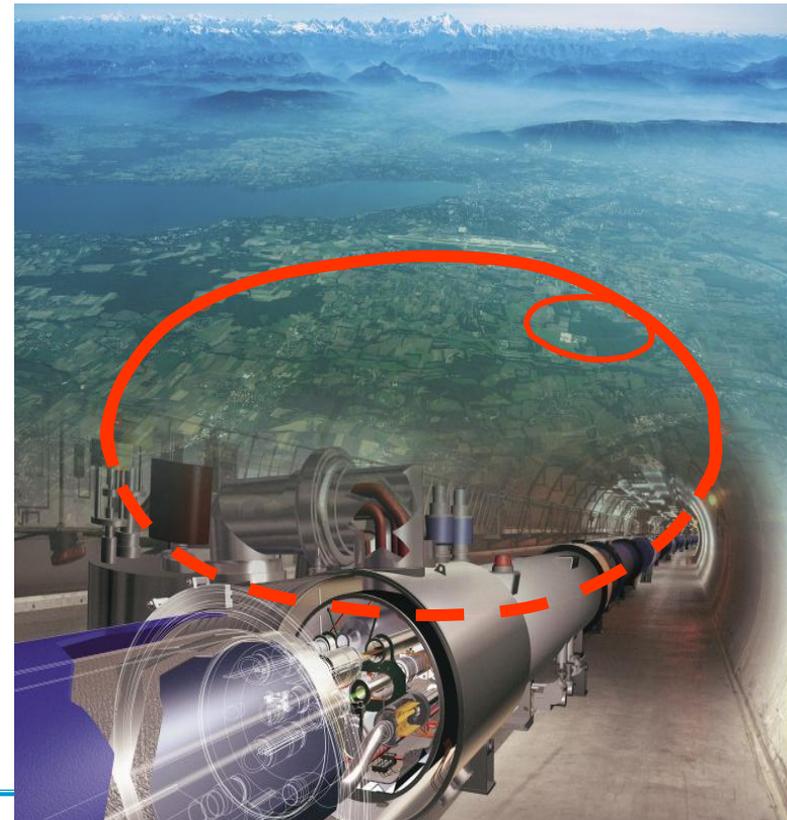
- 并行计算 (parallel computing)
- 分布式计算 (distributed computing)
- 云计算 (cloud computing)
- 启发式方法 (Heuristic Approach)
- 组合式方法 (Configurable Approach)

LarKC: 一个海量语义数据处理平台

<http://www.larkc.eu>

- The Large Knowledge Collider (大型知识对撞机)

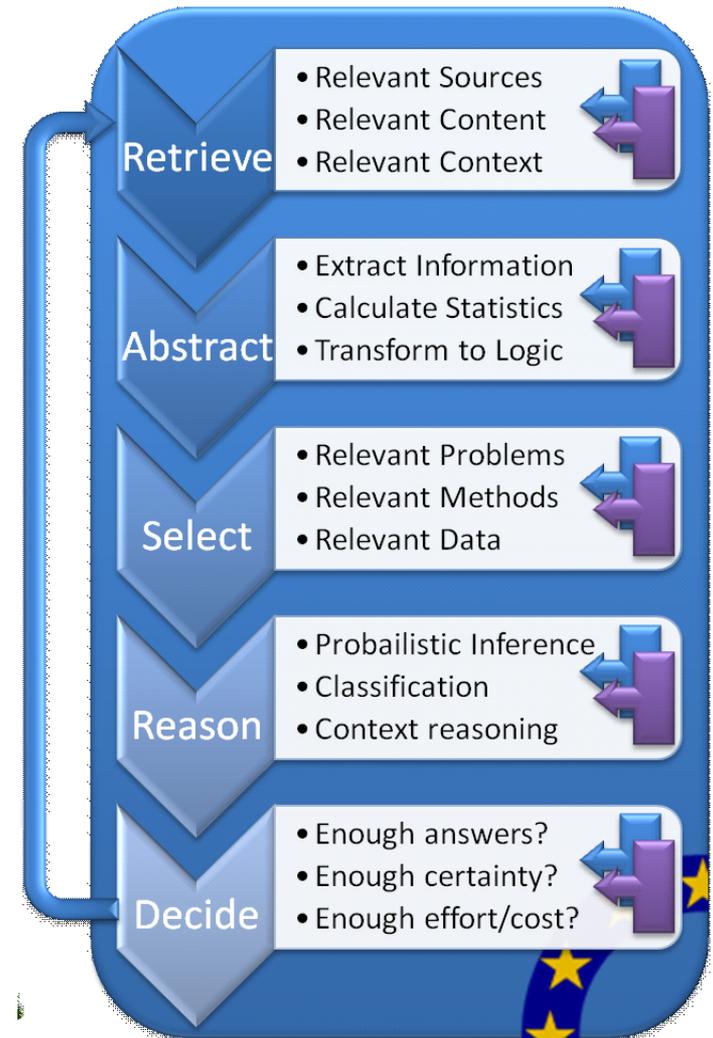
A configurable platform
for experimentation
by others



“Configurable platform”

“a configurable platform for infinitely scalable semantic web reasoning”.

Enrich current logic-based Semantic Web reasoning with methods from information retrieval, machine learning, information theory, databases, and probabilistic reasoning



欧盟第七框架研究课题: LarKC

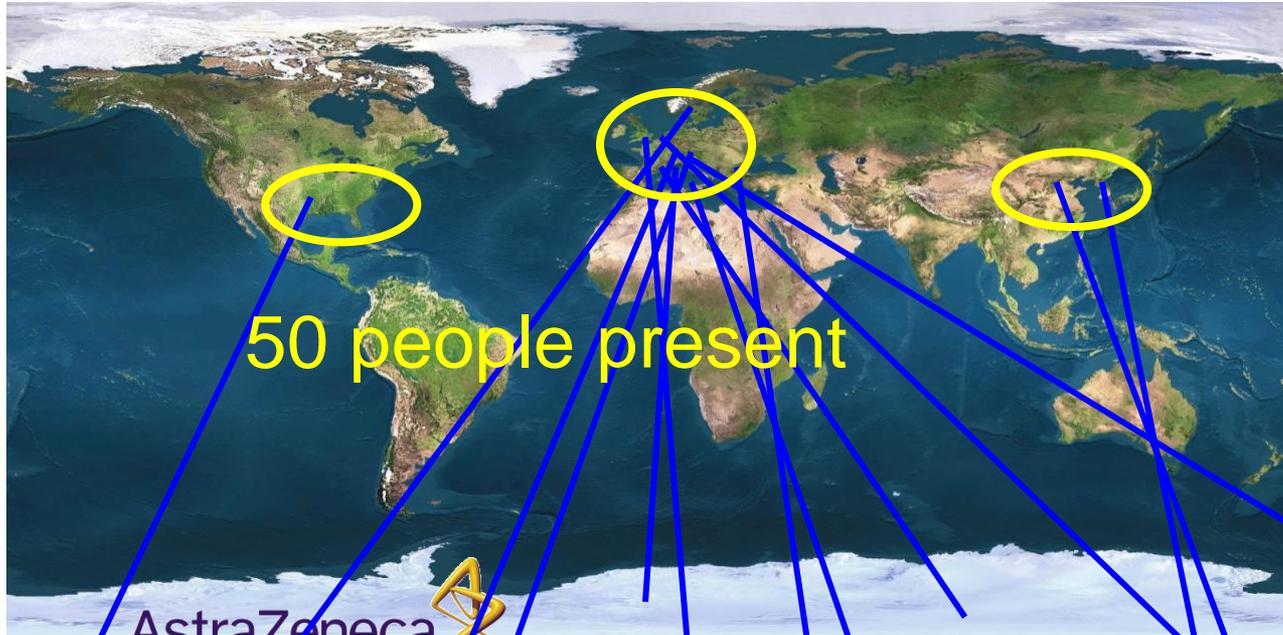
EU 7th framework Project

- 总预算1千万欧元: 10M€ budg
- 历时3年半: 3.5 years
- 八十个人年: 80 person years
- 3个实例研究: 3 case studies
- 14个合作单位: 14 partners,
来自12个国家: 12 countries,
来自3大洲: 3 continents

- project nr. FP7 – 215535



The consortium



50 people present



AstraZeneca



H L R I S



SIEMENS | ontotext
Semantic Technology Lab



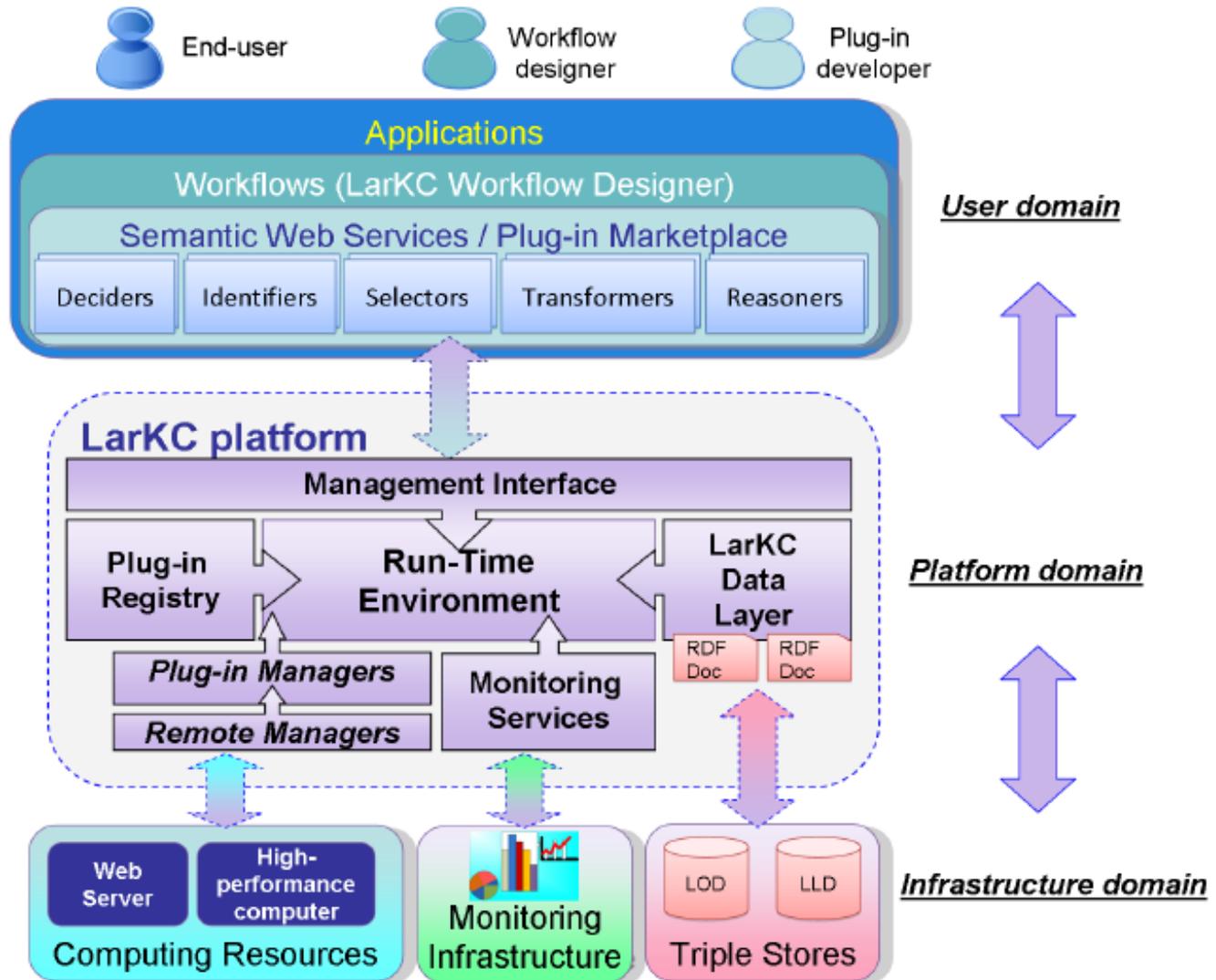
Saltlux



CEFRIL
FORGING INNOVATION | 2014-2017



LarKC Architecture





应用开发：物联网与智能交通

- 意大利米兰交通预测系统（2009）
- 韩国首尔路标管理系统（2010）
- 智能手机城市信息服务系统（2011）
- 北京智能交通管理与决策系统（2012）
- 中国路标管理系统（2013）
- 智慧楼宇监控系统（2014）
- 智慧公园知识服务平台（2015）

- 关联生命数据集及其语义检索系统（2009）
- 全基因组关联研究（癌症研究）（2010）
- Openphacts药物研发平台（2011）
- 临床试验知识管理与决策平台（2012）
- 抗菌药物合理用药系统（2013）
- 脑信息学知识平台（2014）
- 智慧病房（2015）

基于语义技术的智慧北京知识管理与决策支持系统



The screenshot shows a web browser window with the following elements:

- Browser Address Bar:** `http://localhost:8182/workflow`
- Page Title:** 基于语义技术的手机定位识别系统
- Page Header:** VU UNIVERSITY AMSTERDAM logo and "Powered by LARKG" logo.
- Navigation Controls:** "选择时段: 开始日期 20100907", "时间 0000", "结束日期 20100914", "时间 2359", "地图初始化", "地图更新".
- Map:** A Google Map of Beijing with a blue rectangular overlay indicating a specific geographic area. The map shows major roads, parks, and landmarks.
- Footer:** Windows taskbar showing the start button, open windows, and system tray with the time 5:41 PM.

天通苑地区移动数据指标验证

调查问卷表格及现场



天通苑地区居民出行问卷调查

您好！本调查为北京工业大学的基础科研工作之一，为促进天通苑地区交通服务水平提升，占用您一点宝贵时间。此次调查的数据保证不会给您带来任何麻烦，感谢合作！

请您结合实际详细、认真填写调查表中的有关问题，在填写后我们会赠送精美礼品！

1 性别 A.男 B.女
2 年龄 A.20-30岁 B.30-40岁 C.40-50岁 D.50-60岁 E.60岁以上
3 家庭成员数 A.1 B.2 C.3 D.4 E.5 其他
4 家庭拥有小汽车数 A.0 B.1 C.2 D.3 其他
5 家庭所在区域
A.天通北苑（一区 二区） B.天通中苑 C.天通苑东区
D.天通东苑（一区 二区 三区） E.天通西苑（一区 二区 三区）
F.其他
6 您的手机属于哪个移动通信运营商 A.移动 B.联通 C.电信 其他
7 您家庭成员的每日累计出行次数约为 A.2 B.3 C.4 D.5 其他

出行基本信息

第①次出行：
1 出行时间
□2:00-6:00 □6:00-10:00 □10:00-14:00
□14:00-18:00 □18:00-22:00 □22:00-2:00
2 出行目的：A.上班 B.上学 C.公务 D.探亲访友 E.文化娱乐 F.购物
G.其他
3 出行起点 昌平区天通苑北三区（如：朝阳区，平乐园100号）
4 出行终点 昌平区天通苑北三区（如：朝阳区，平乐园100号）
5 出行方式：
(1) 如果不需要换乘：采用的交通方式为：
A.地铁 B.公交 C.小客车 D.步行 E.班车 F.出租 G.黑车
H.摩托 I.校车 J.电动自行车 K.货车 L.自行车 M.其他
总共花费您的时间为：(120)分钟
(2) 如果需要换乘：
您首先采用的交通方式为(公交) 花费了(20)分钟
随后您换乘(地铁) 花费了(40)分钟
之后您换乘() 花费了()分钟
之后您换乘() 花费了()分钟，最终您到达了目的地。
A.地铁 B.公交 C.小客车 D.步行 E.班车 F.出租 G.黑车
H.摩托 I.校车 J.电动自行车 K.货车 L.自行车 M.其他

第②次出行：1 出行时间
□2:00-6:00 □6:00-10:00 □10:00-14:00
□14:00-18:00 □18:00-22:00 □22:00-2:00

北京工业大学 Beijing University of Technology

2 出行目的：A.上班 B.上学 C.公务 D.探亲访友 E.文化娱乐 F.购物
G.其他
3 出行起点 昌平区天通苑北三区（如：朝阳区，平乐园100号）
4 出行终点 昌平区天通苑北三区（如：朝阳区，平乐园100号）
5 出行方式：
(1) 如果不需要换乘：采用的交通方式为：
A.地铁 B.公交 C.小客车 D.步行 E.班车 F.出租 G.黑车
H.摩托 I.校车 J.电动自行车 K.货车 L.自行车 M.其他
总共花费您的时间为：(40)分钟
(2) 如果需要换乘：
您首先采用的交通方式为(地铁) 花费了(40)分钟
随后您换乘(公交) 花费了(20)分钟
之后您换乘() 花费了()分钟
之后您换乘() 花费了()分钟，最终您到达了目的地。
A.地铁 B.公交 C.小客车 D.步行 E.班车 F.出租 G.黑车
H.摩托 I.校车 J.电动自行车 K.货车 L.自行车 M.其他

第③次出行：
1 出行时间
□2:00-6:00 □6:00-10:00 □10:00-14:00
□14:00-18:00 □18:00-22:00 □22:00-2:00
2 出行目的：A.上班 B.上学 C.公务 D.探亲访友 E.文化娱乐 F.购物
G.其他
3 出行起点 昌平区天通苑北三区（如：朝阳区，平乐园100号）
4 出行终点 昌平区天通苑北三区（如：朝阳区，平乐园100号）
5 出行方式：
(1) 如果不需要换乘：采用的交通方式为：
A.地铁 B.公交 C.小客车 D.步行 E.班车 F.出租 G.黑车
H.摩托 I.校车 J.电动自行车 K.货车 L.自行车 M.其他
总共花费您的时间为：()分钟
(2) 如果需要换乘：
您首先采用的交通方式为() 花费了()分钟
随后您换乘() 花费了()分钟
之后您换乘() 花费了()分钟
之后您换乘() 花费了()分钟，最终您到达了目的地。
A.地铁 B.公交 C.小客车 D.步行 E.班车 F.出租 G.黑车
H.摩托 I.校车 J.电动自行车 K.货车 L.自行车 M.其他

感谢您的积极配合

- 出行时间
- 出行次数
- 出行目的
- 出行距离及耗时
- 出行方式

手机数据与出行需求分析

- 手机信令中蕴含着移动用户的位置信息，能够对移动用户进行连续追踪，具有**实时性、高样本量**等特性，完全吻合进行交通信息分析的需求。
- 手机信令数据的采集对象为手机终端，而交通出行人群的主体也是手机的主要使用群体
- 因此，基于手机信令数据分析得到的交通信息更能够反映**主体人群的交通特征和规律**，更具有研究的必要性及意义。

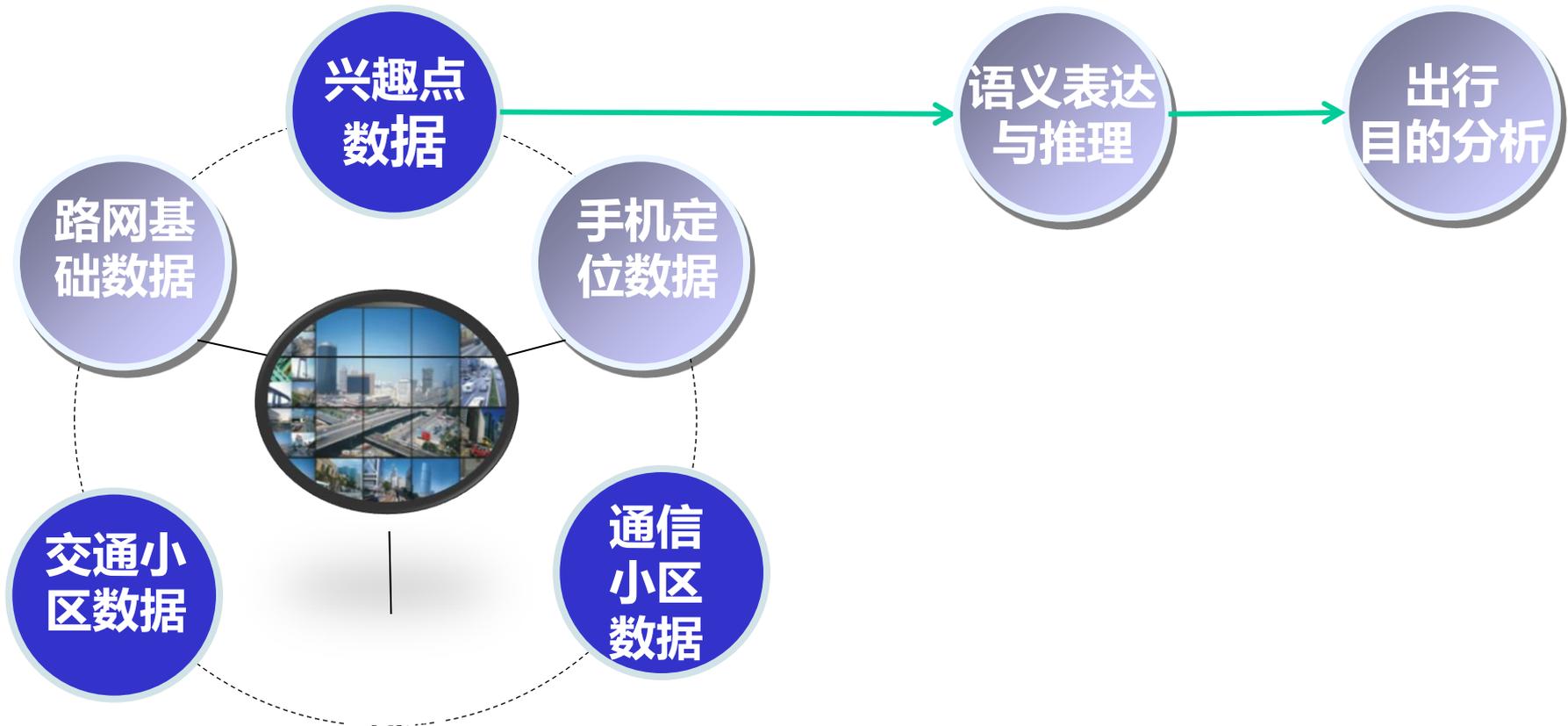
研究背景概述

- 单一的手机定位数据分析在数据处理可靠性，指标丰富性方面存在一定不足
- 利用语义技术在处理多源数据，知识推理等方面良好扩展性
- 发挥现有异构数据的统一表达实现多源数据融合分析，针对手机数据开展纵深分析
- 掌握北京市内居民出行、通勤的规律，分析进出京人口的数量、OD，实时监控热点区域实际客流数量
- 开展基于多源数据融合分析的出行目的分析
- 为北京城市交通规划、城市交通管理提供科学的数据支持

语义数据优势

- 语义技术就是通过使用一种新的**知识表示方式**让信息**包含更多语义**，使计算机在一定程度上理解知识的含义，从而实现机器对知识的自动处理及逻辑推理。
- 语义技术能够充分利用万维网上大量的**语义资源**，并且易于集成各种异质的数据源，同时具有良好的**扩展性能**。

1.3 研究技术路线



3.2 多源数据的语义表达

- 判定一个手机持有者的家庭位置，是出行分析的基本信息需求。它可以通过分析手机位置来估计。这个基本推理规则是：

如果一个人在午夜12点到早晨6点连续呆在同一个位置，则基本上可以判定这个位置就是这个手机持有者的家庭位置，除非他/她是在上夜班。

- SPARQL 查询

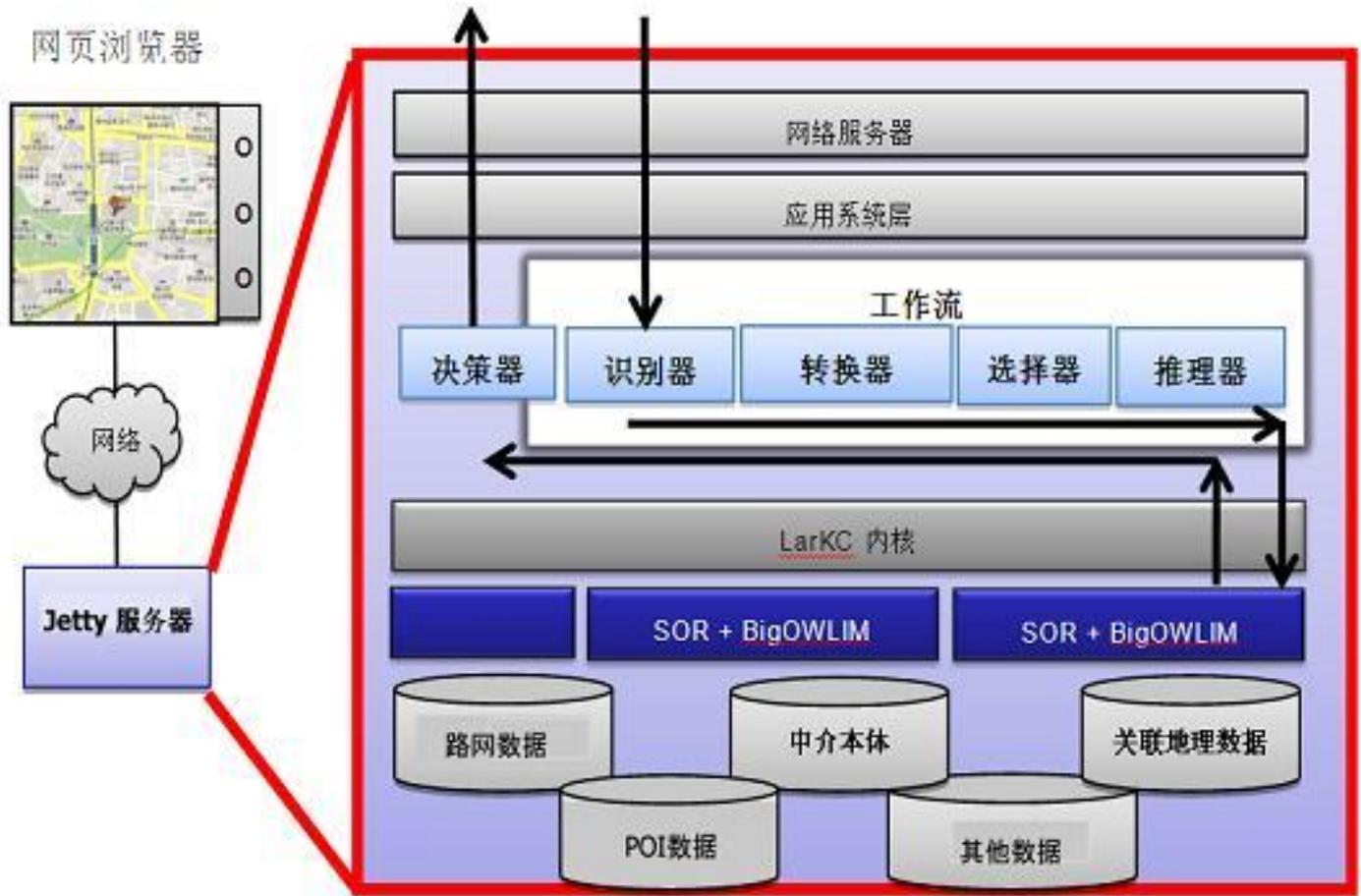
```
select distinct ?phoneNumber ?lat ?long
where {?record1 rdf:type ucc:MobileRecord.
      ?record1 ucc:MobilePhoneNumber ?phoneNumber.
      ?record1 ucc:TimeStamp ?timestamp1.
      ?record2 ucc:TimeStamp ?timestamp2.
      FILTER(?timestamp1 <= '20100908010000' && ?timestamp2 >= '20100908060000').
      ?record2 ucc:MobilePhoneNumber ?phoneNumber.
      ?record1 wgs:lat ?lat.
      ?record1 wgs:long ?long.
      ?record2 wgs:lat ?lat.
      ?record2 wgs:long ?long.
```

语义查询 (SPARQL)

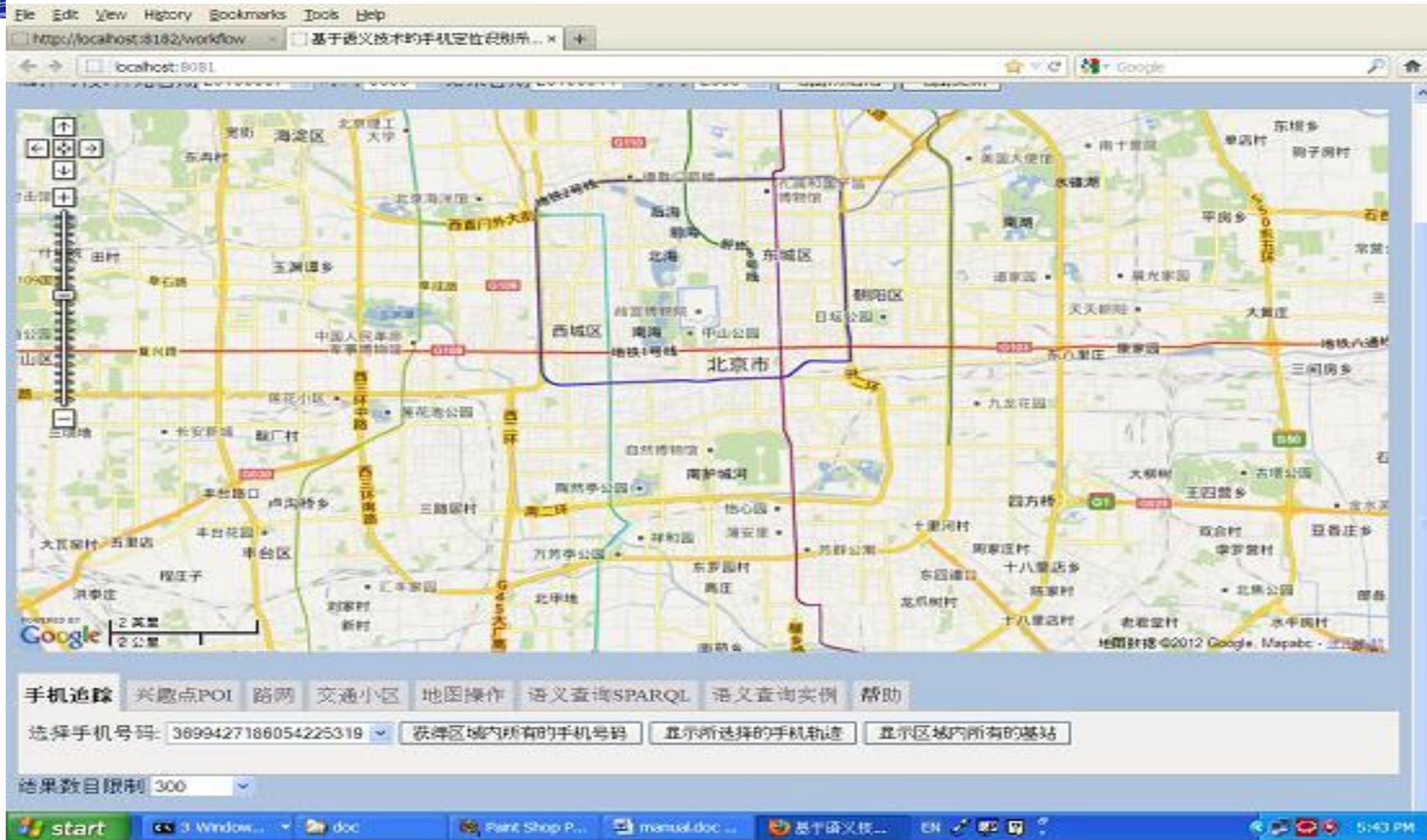
- 列出特定手机用户的轨迹数据，及其所经过的邻近的交通小区的名字， 并以其时间标排序：

```
select distinct ?onenumber ?timestamp ?lat ?long ?name
where {
?record rdf:type ucc:MobileRecord.
?record ucc:MobilePhoneNumber "3699427186054225319".
?record wgs:lat ?lat.
?record wgs:long ?long.
?record ucc:TimeStamp ?timestamp.
?tr rdf:type ucc:TrafficRegion.
?tr wgs:lat ?lat1.
?tr wgs:long ?long1.
FILTER (?lat1 > ?lat - 0.005 &&
?lat1 < ?lat + 0.005 &&
?long1 > ?long - 0.005 &&
?long1 < ?long +0.005).
?tr ucc:name ?name.
FILTER( ?lat <= 40.064409681221484 &&
?lat >= 39.74837783143156 &&
?long <= 116.81419372558594 &&
?long >= 116.26487731933595).}
ORDER BY ?timestamp
```

系统结构



用户界面



手机轨迹追踪

File Edit View History Bookmarks Tools Help

http://localhost:8182/workflow

localhost:8081



手机追踪 兴趣点POI 路网 交通小区 地图操作 语义查询SPARQL 语义查询实例 帮助

选择手机号码: 3890427573232070159

I	TimeStamp	Lat	Long	Time Diff	Shift Speed	StayTime(sec)	Region(under test)	Status
[0]	2010-09-09,06:59:20	39.889506	116.345398	0	0	0	五塔寺社区	Staying
[1]	2010-09-09,07:34:31	39.887537	116.340818	2111	0.7638	0	中央财经大学	New departure point
[2]	2010-09-09,08:06:12	39.891738	116.350768	1901	1.8349	0	中央音乐学院	New departure point
[3]	2010-09-09,08:06:18	39.891738	116.350768	6	0	6	中央音乐学院	Staying
[4]	2010-09-09,08:08:06	39.891738	116.350768	108	0	114	中央音乐学院	Staying
[5]	2010-09-09,08:08:09	39.891738	116.350768	3	0	117	中央音乐学院	Staying

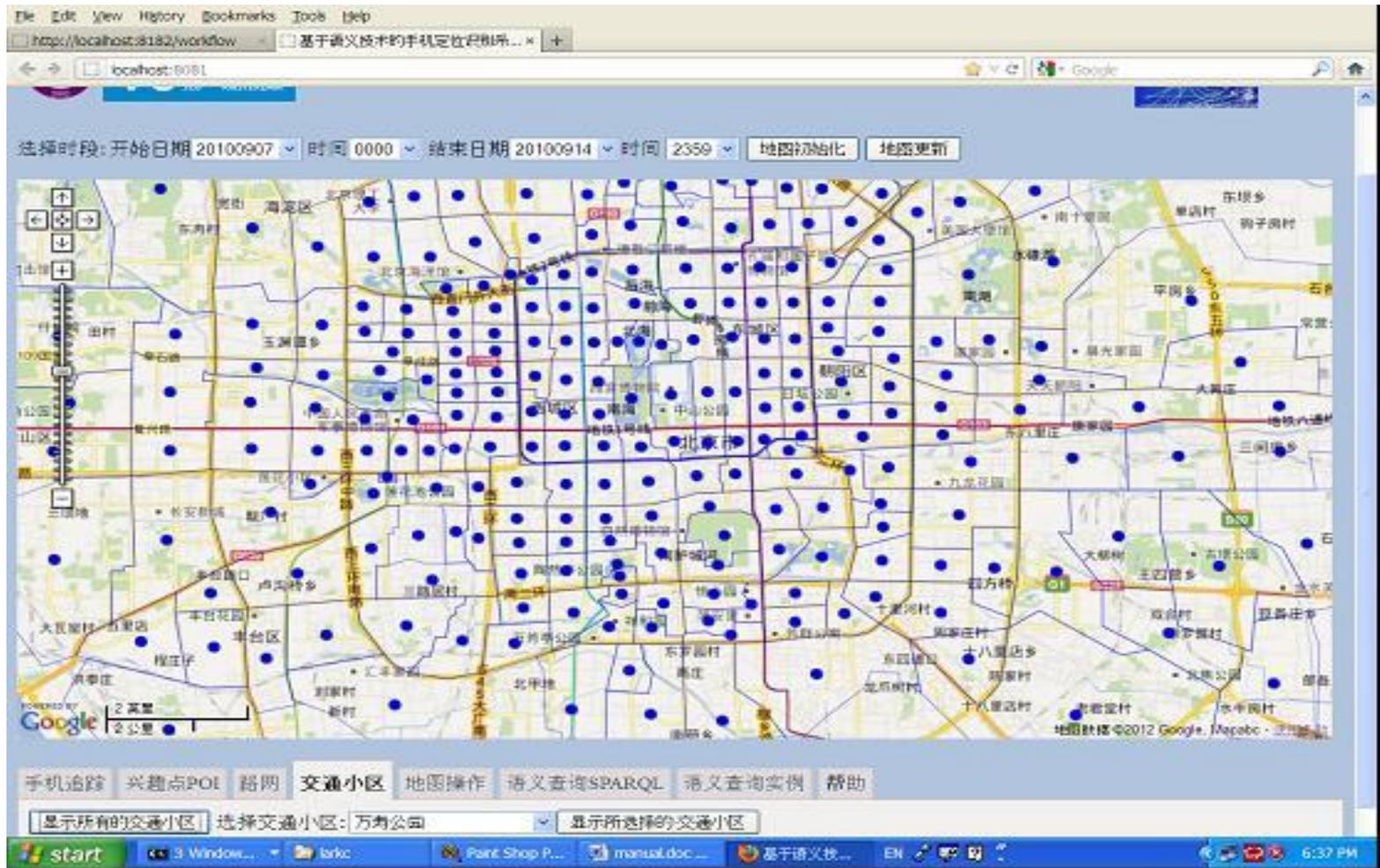
start 3 Window... larkc Paint Shop P... manual.doc... 基于语义技... EN 6:07 PM

北京POI语义数据



The screenshot displays a web application interface for querying POI data in Beijing. The main area is a Google Map showing a network of roads and several blue location pins. A search bar at the bottom left contains the text "选择POI类别: 桥" (Select POI category: Bridge). Below the search bar, there are two buttons: "显示区域内所有的兴趣点" (Show all POIs in the area) and "显示区域内所选类别的兴趣点" (Show POIs of the selected category in the area). The results limit is set to "500". The browser window shows the URL "http://localhost:8182/workflow" and the page title "基于语义技术的手机定位识别系...". The Windows taskbar at the bottom shows the time as 6:33 PM.

北京交通小区语义数据



The screenshot displays a web application interface for visualizing semantic data of traffic zones in Beijing. The browser window shows the URL `http://localhost:8182/worldow` and the page title "基于语义技术的手动定位识别系...". The application interface includes a navigation bar with the following elements:

- Navigation: "手机追踪", "兴趣点POI", "路网", "交通小区", "地图操作", "语义查询SPARQL", "语义查询实例", "帮助"
- Time Selection: "选择时段: 开始日期 20100907", "时间 0000", "结束日期 20100914", "时间 2359", "地图初始化", "地图更新"
- Map: A map of Beijing with numerous blue dots representing traffic zones. The map includes labels for districts like "海淀区", "朝阳区", and "丰台区", and various landmarks and roads.
- Map Controls: A vertical toolbar on the left side of the map for navigation and zooming.
- Footer: "显示所有的交通小区" | "选择交通小区: 万寿公园" | "显示所选择的交通小区"

The Windows taskbar at the bottom shows the system time as 6:37 PM and includes icons for the Start button, taskbar, and several open applications like "Paint Shop P...", "manual.doc...", and "基于语义技...".

总结：语义技术的优越性

- 采用独立于具体应用系统的统一数据表达格式，便于融合他人的现有数据，也有利于未来的系统功能的扩充
 - 便于知识提取和知识表达，代替现有的大量的人工干预的枯燥工作，
 - 引入知识处理，提高了处理问题的精度和效率
 - 提供知识管理与推理，对宏观把握信息系统提供决策支持
-

Questions and Discussions

