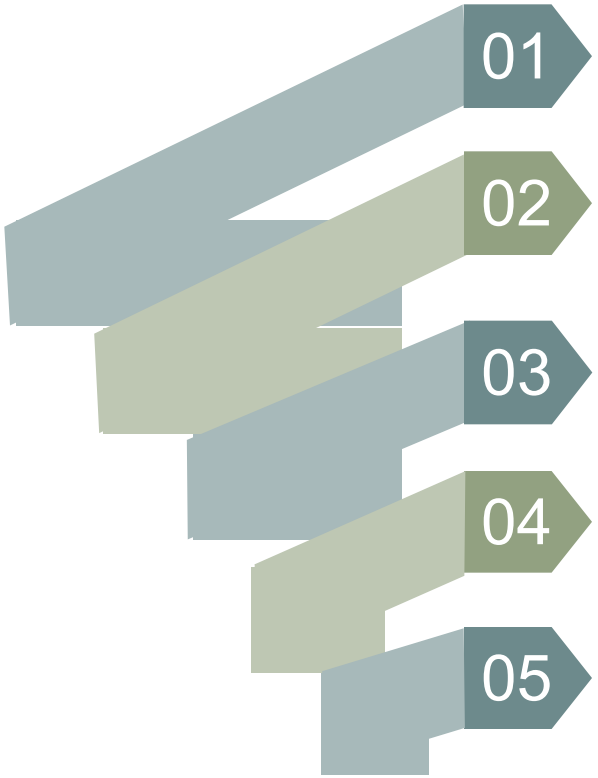
Web Information Acquisition, Data Integration and Its Application in Crisis Management

Jiping Liu

Chinese Academy of Surveying and Mapping

☞Outline



01 **Background**

02 **Current Situation and Problems**

03 **Research Methods**

04 **Typical experiments**

05 **Conclusion and Ongoing work**

☞Background

Disaster: affect the world every day

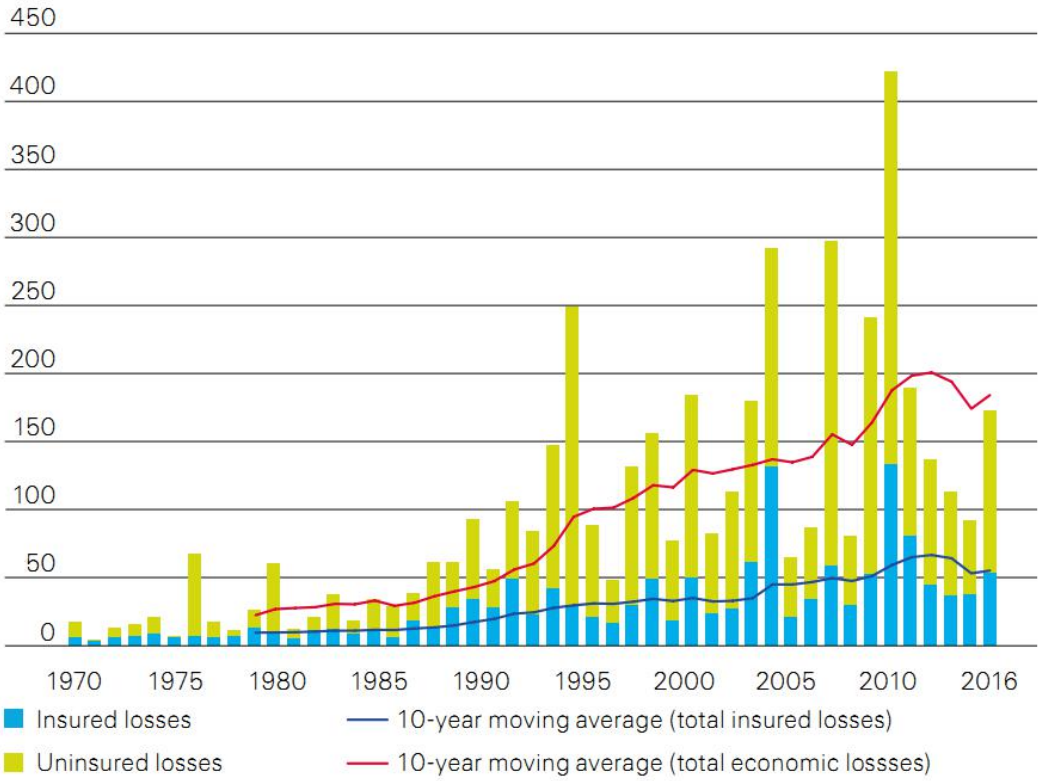
- A grave occurrence having ruinous results. (Webster's Dictionary)
- It could be **natural or man-made**.
- such as flood, hurricane, fire, earthquake, plane crash, etc.

☞Background

- Swiss Re's sigma study reveals the number of catastrophic events and victims and the economic losses from 1970 to 2016:

The estimated total economic losses were USD 175 billion in 2016.Catastrophe losses in 2016 were 0.24% of global GDP, again in line with he 10-year average.

Economic losses, 1970–2016, USD billion in 2016 prices

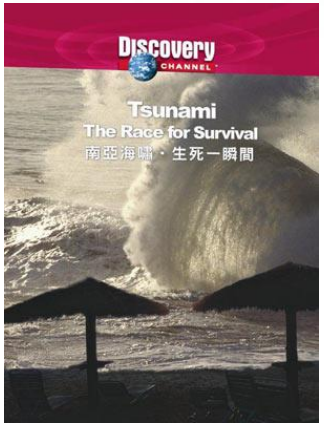


Total losses = insured + uninsured losses.

☞Background

Terrorist attacks

Flood



Tsunami

Hurricane

Earthquake

In china, only the natural disasters caused 1432 people died and direct economic losses stood at 503.29 billion yuan in 2016.

☞Background

How to respond better to the disaster?

How to deal with emergency in a short time?

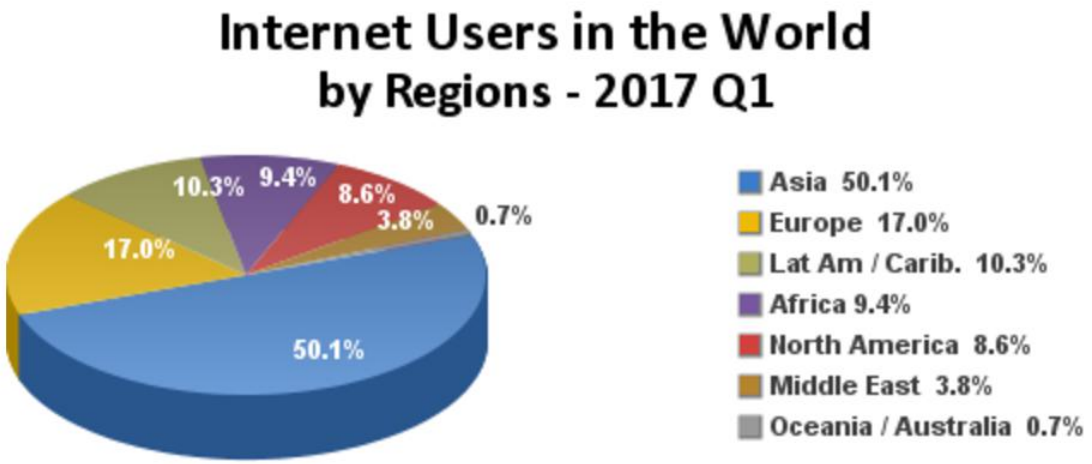
Policymakers need comprehensive and effective information to make decisions

A continuous process of planning and coordinating for disaster management.

☞Background

Internet of Things: The new era is coming

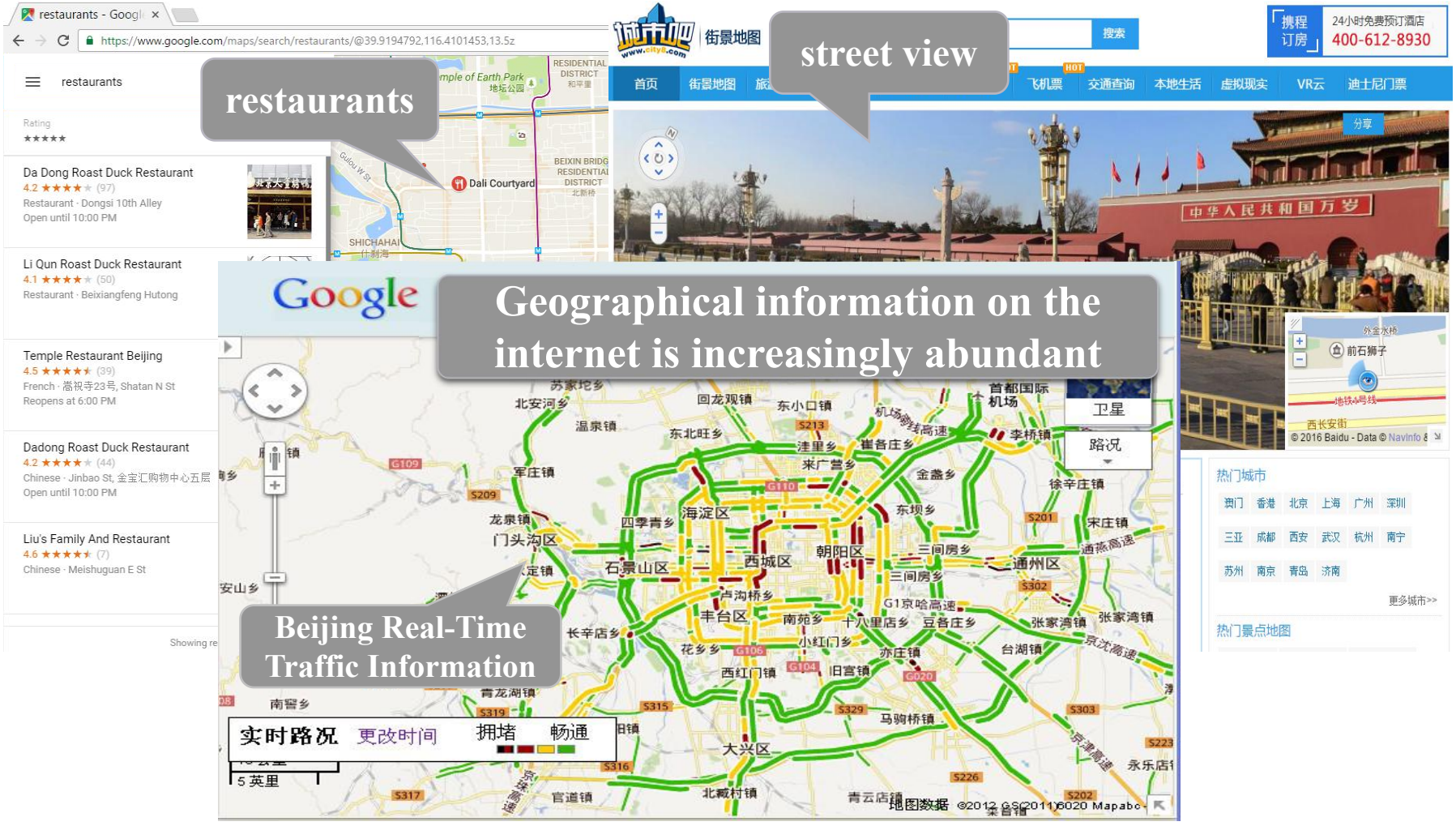
- Up to March 31, 2017, the number of internet users worldwide was 3.7 billion, up from 3.4 billion in the previous year.



- The Asian region ranked first, which was 50.1% of the total
- The number of netizens in China has reached 731 million. Penetration is reached 52.7%

☞Background

- As the important channels of information release, dissemination and exchange , internet contains a large number of rich geographical information.



☞Background

As a kind of geographic information public service, Large volume of Point of Interest (POI) are provided by many geographic information services on the Internet, and the number will increase rapidly in future.

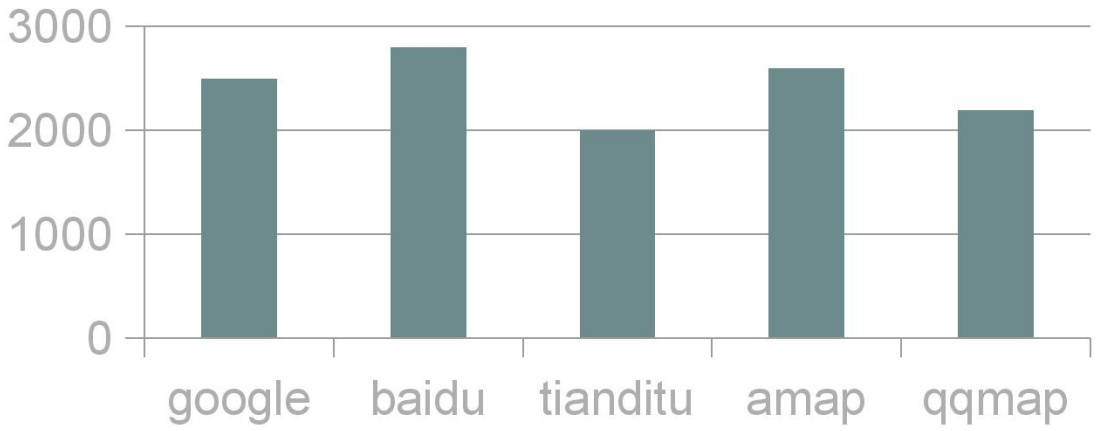
- **POI (Point of Interest) is mostly used in GIS and LBS**
- Large volume of POI are provided by many web sites
- POIs are updated periodically

More than 1000 geographic information service websites

☞Background

As a kind of geographic information public service, Large volume of Point of Interest (POI) are provided by many geographic information services on the Internet, and the number will increase rapidly in future.

- POI (Point of Interest) is mostly used in GIS and LBS
- **Large volume of POI are provided by many web sites**
- POIs are updated periodically



The average number of POI for each website is more than **20 million**, and the total of POIs from the five web sites is up to **121 million**.

☞Background

As a kind of geographic information public service, Large volume of Point of Interest (POI) are provided by many geographic information services on the Internet, and the number will increase rapidly in future.

- POI (Point of Interest) is mostly used in GIS and LBS
- Large volume of POI are provided by many web sites
- **POIs are updated periodically**

Name	Frequency of updates
Google Map	About once a month
Baidu Map	Once every quarter
Tianditu Map	Twice a year
AMap	Once every quarter
QQMap	Once every quarter

☞Background

Sources of internet geographical information:

criteria	Data from authoritative and professional	Public service data	Data of GPS	Data spontaneously	Data derived from Web2.0
In					netizens
In					Geotags
Di					vague
Cr					lower
G					lower
In					weaker
Up					no
Up				crowdsourcing、passive	crowdsourcing、passive	crowdsourcing、passive

☞Background

Sources of internet geographical information:

Maps from Internet service enterprises

Update mode	initiative、regular

☞Background

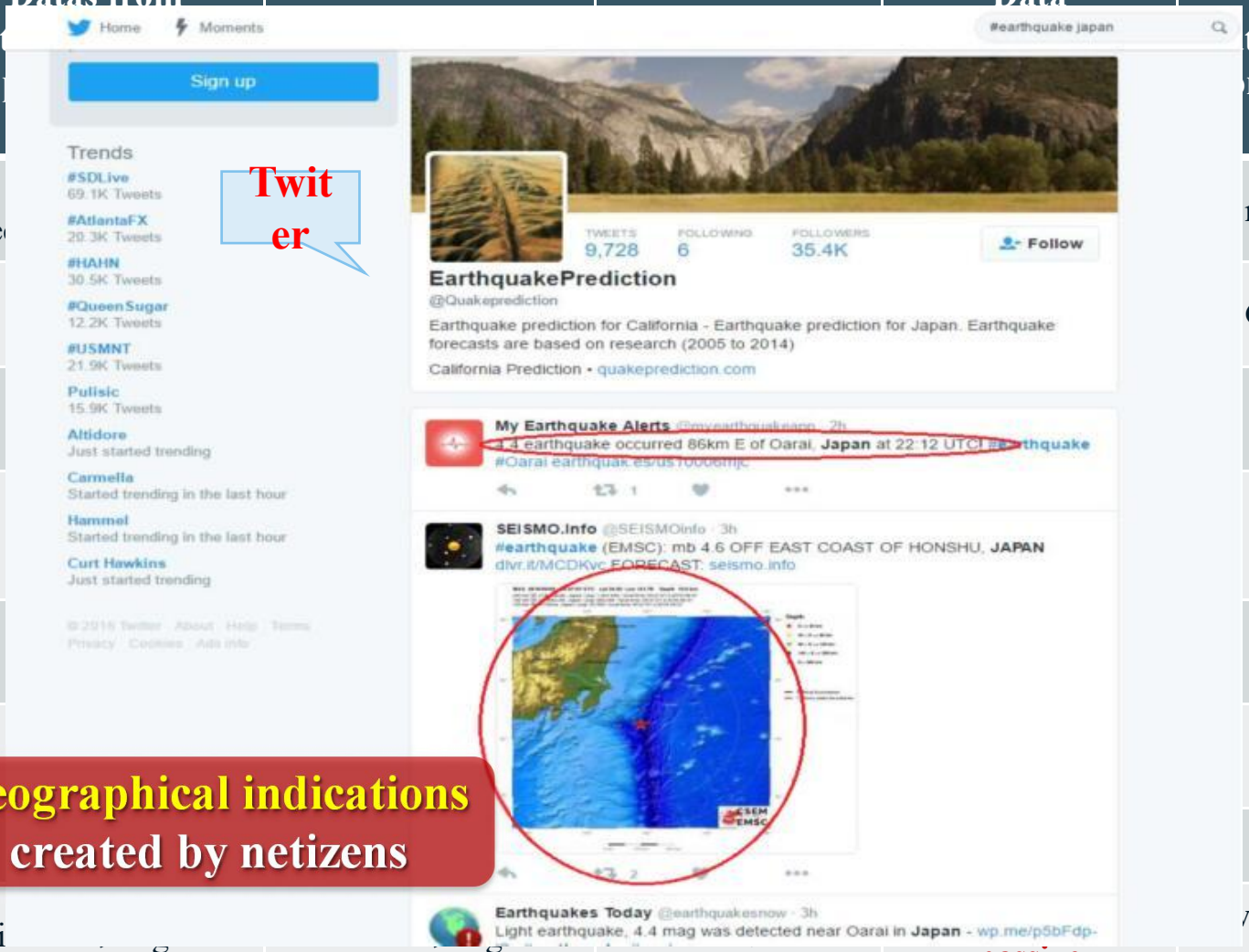
Sources of internet geographical information:

criteria	datas from auth… p… d…			Data …	Data derived from Web2.0
Information sources	C… spec…				netizens
Information forms					Geotags
Directivity of topic					vague
Credibility of content					lower
Geometric richness					lower
Integrity of attribute					weaker
Update subject					no
Update mode	init…				…owdsourcing、passive

☞Background

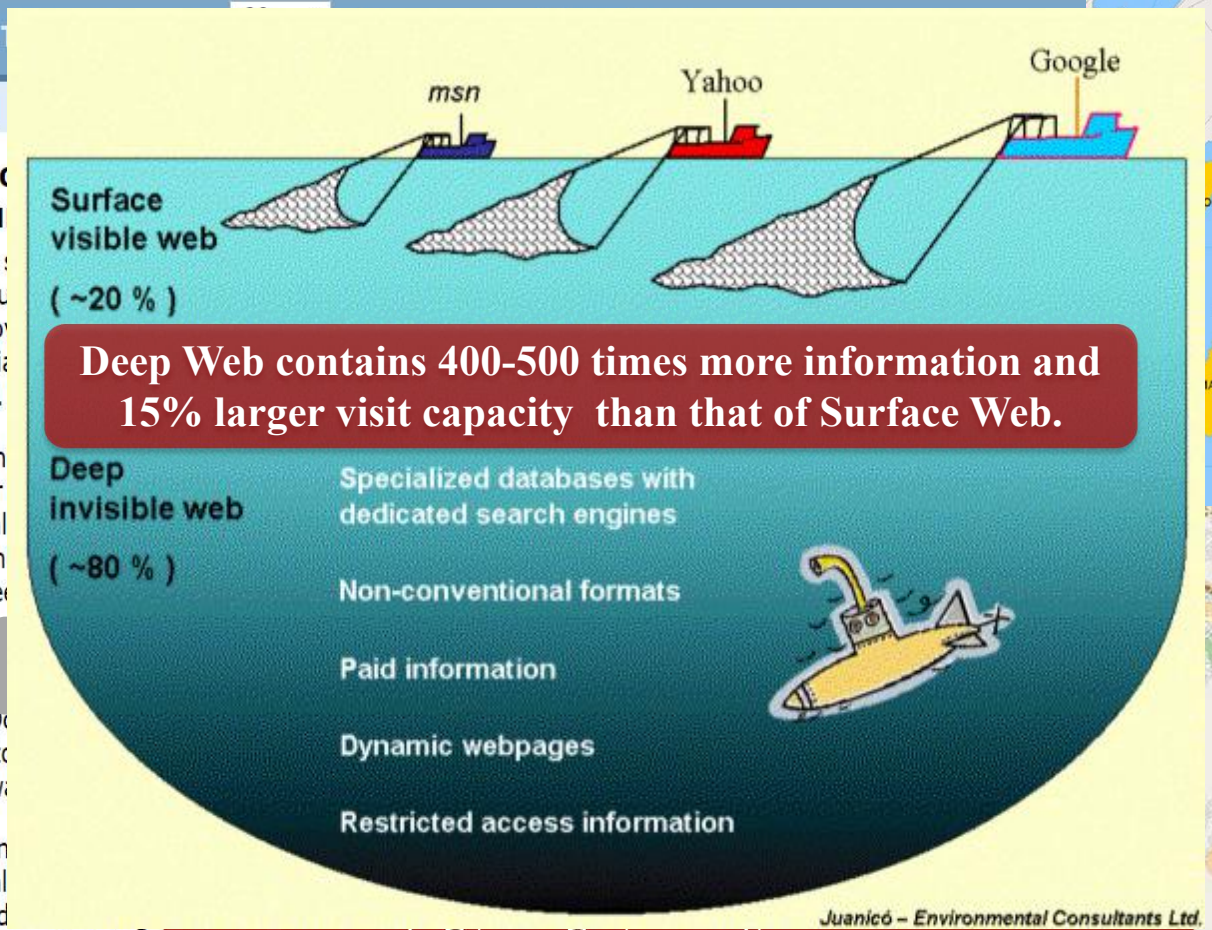
Sources of internet geographical information:

criteria	aut…	Datas from …			Data	…ta derived …m Web2.0
Information sources	spe…					netizens
Information forms						Geotags
Directivity of topic						vague
Credibility of content						lower
Geometric richness						lower
Integrity of attribute						weaker
Update subject						no
Update mode	ini…			passive	**passive**	…/dsourcing、passive



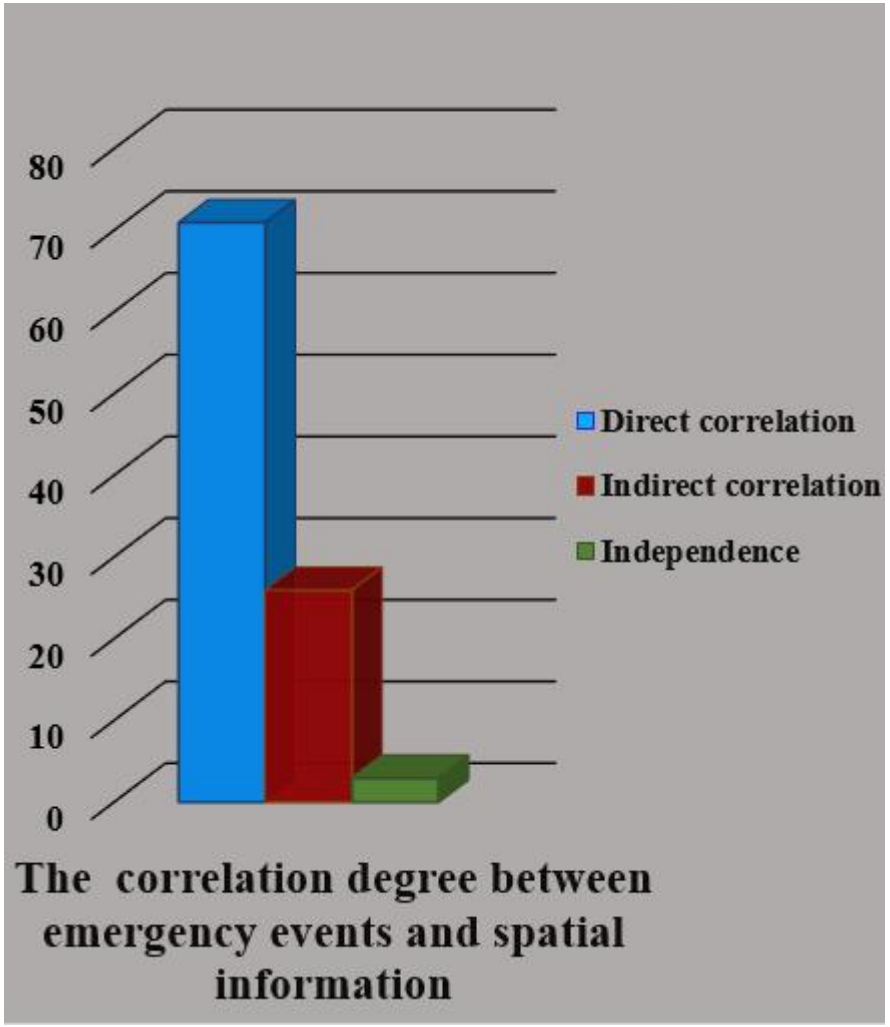
☞Background

Characteristics of internet geographical information:



Background

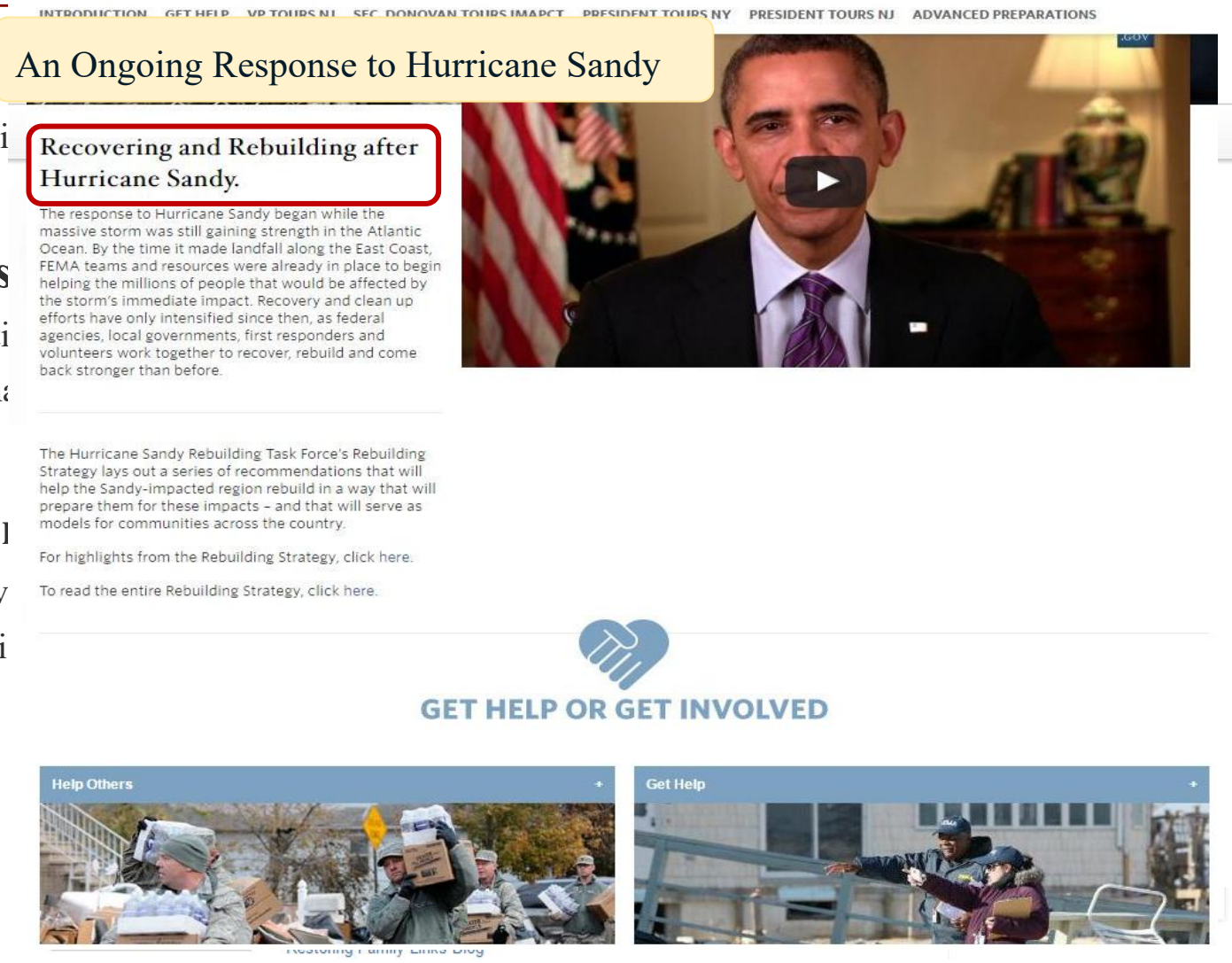
- The direct correlation degree between emergency events and spatial information is 71%, and the indirect correlation degree is 26%.



☞Background

Geographical information on the Internet has played a very important role in all aspects of

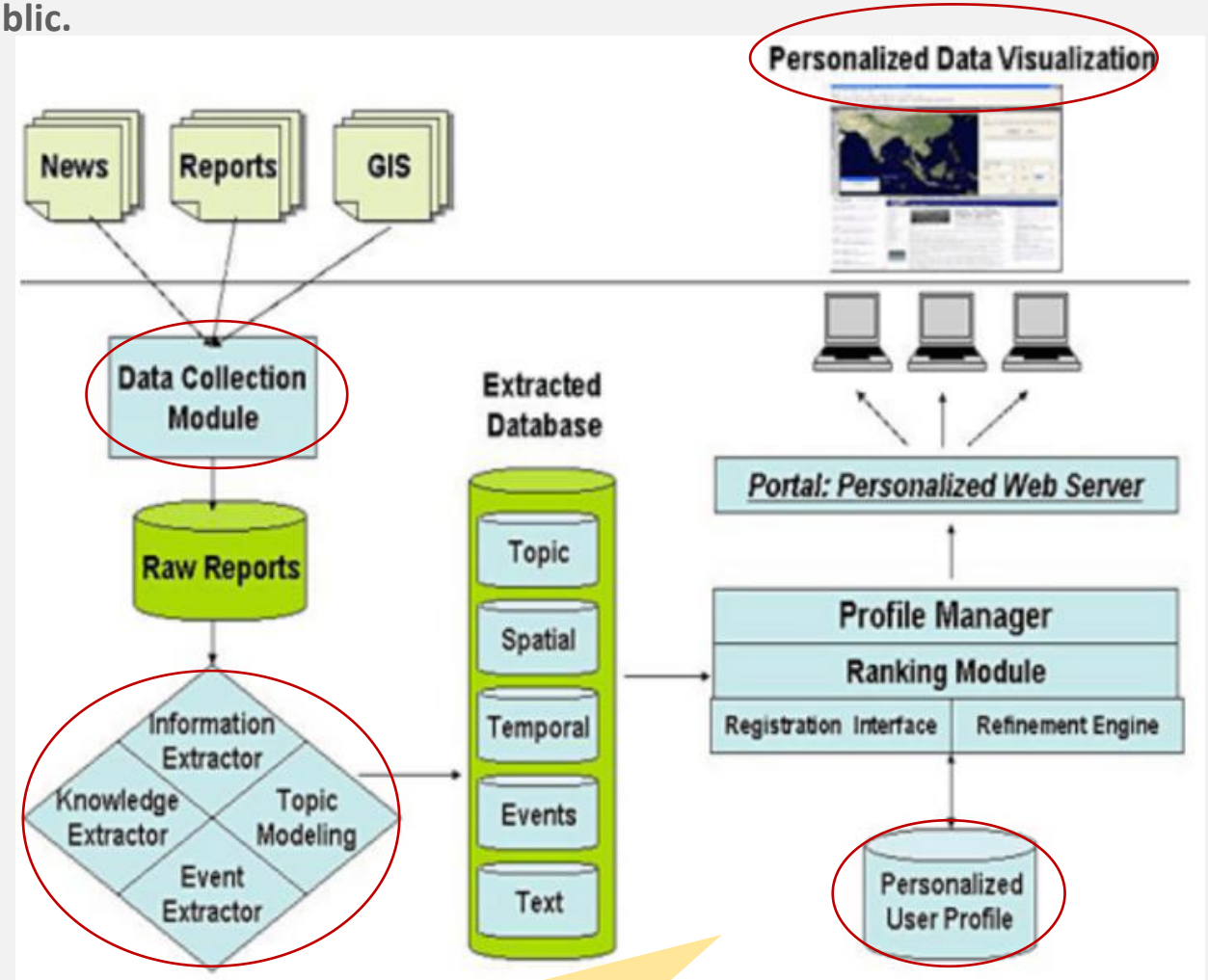
- **pre-disaster**
  - --provide the earli
  - --early warning
- **during disas**
  - --providing real-ti
  - --helping coordina
  - --finding missing
- **post-disaste**
  - --provides effectiv
  - --launching donati



☞Current Situation and Problems

- **RESCUE Project:** funded by the National Science Foundation(NSF). **Its goal is to dramatically improve the ability of emergency responders to gather, process, and disseminate information with each other and the general public.**

—For example, PSAP (Personalized Situation Awareness Portal), an information portal, a research sponsored by RESCUE Project

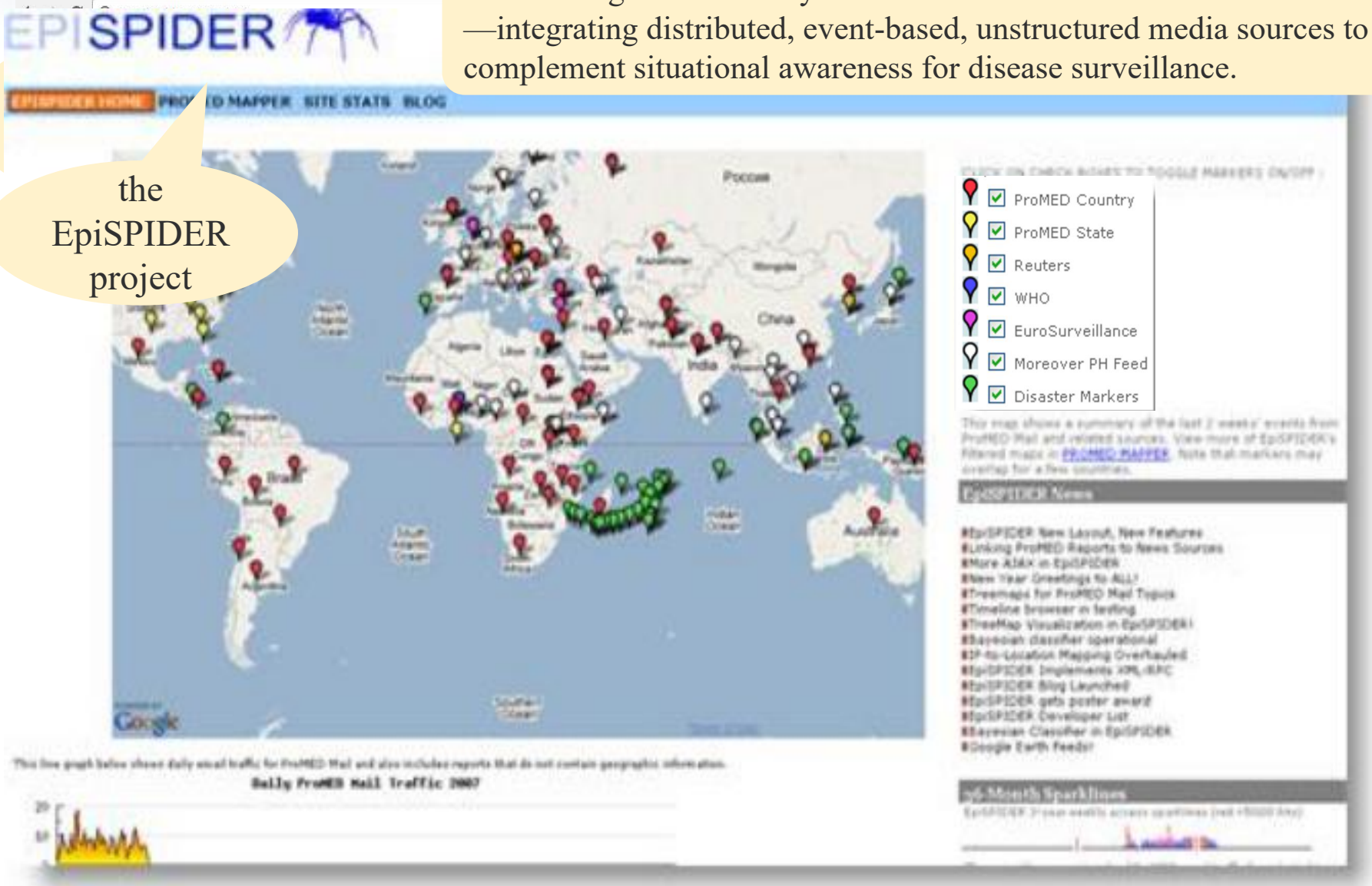


System Architecture of PSAP, illustrates the key technical components

☞Current Situation and Problems

—was designed in January 2006 .
—integrating distributed, event-based, unstructured media sources to complement situational awareness for disease surveillance.

the EpiSPIDER project



☞Current Situation and Problems

Problems

The massive, heterogeneous and dynamic characteristics of Internet Information brings difficulties in information retrieval technology.

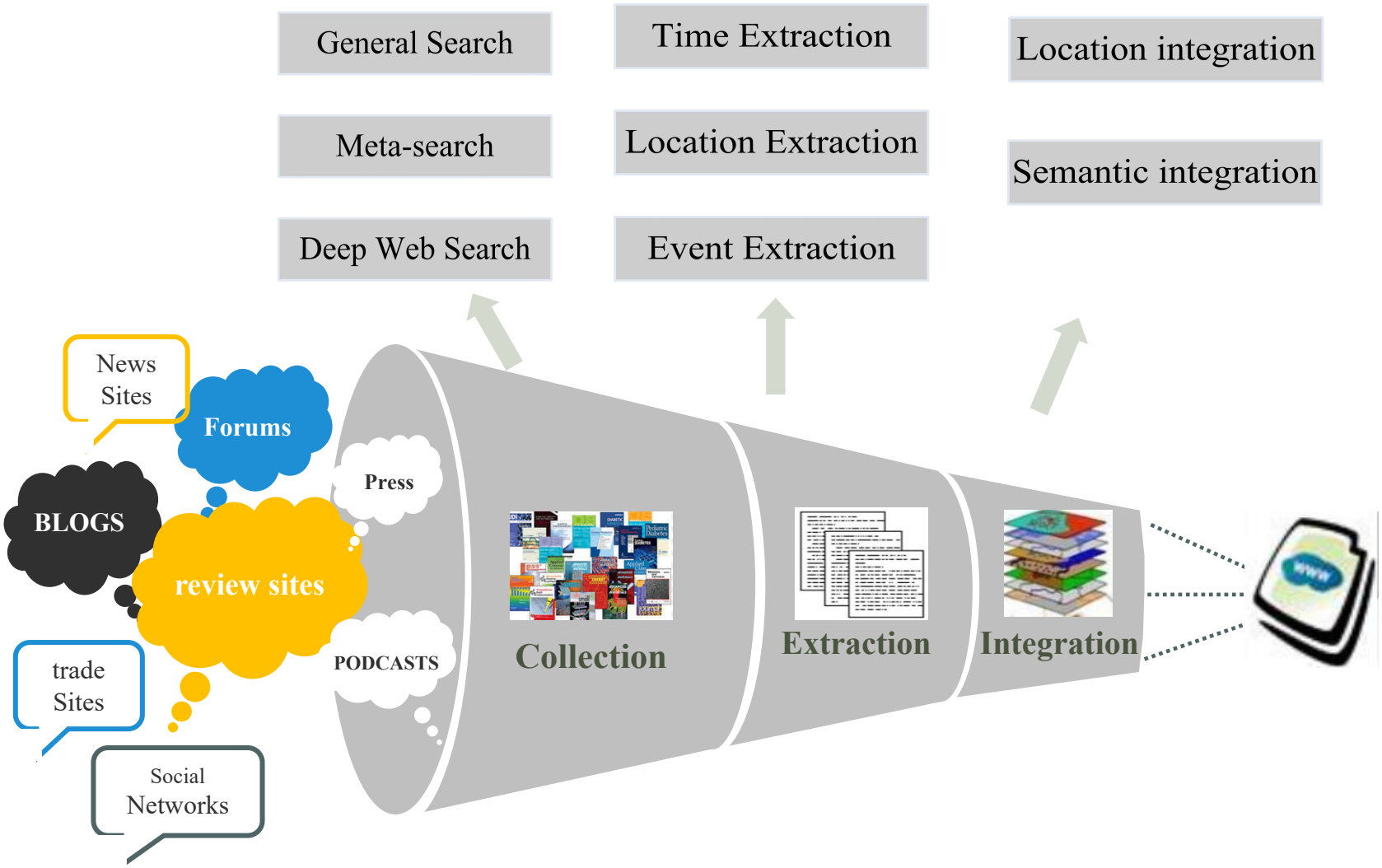
—Spatial information search: how to choose the driven crawler to get the ideal set of target sites.

—Information Extraction: how to distinguish and extract the polymorphic geographic information in web pages.

—Information Integration: how to deal with the Web data and apply them to practical application more appropriately.

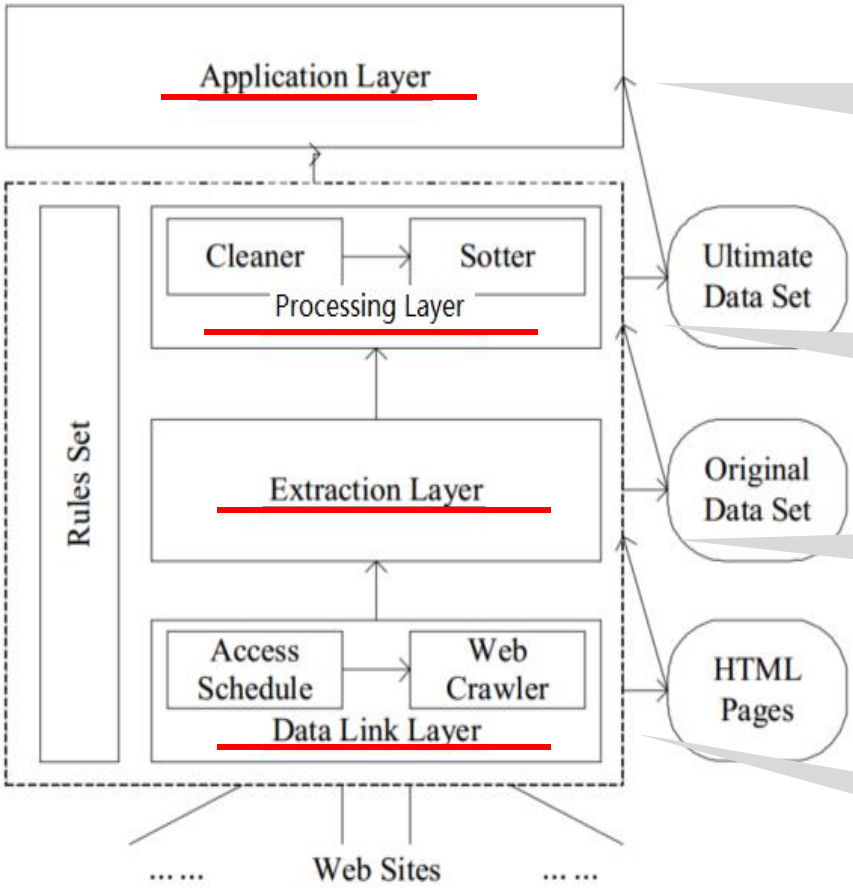
☞Research methods

Framework



☞Research methods

- The structure contains four main levels from bottom to top:



- the communication interface between model and users.
- receiving query requests from users
- illustrating the results

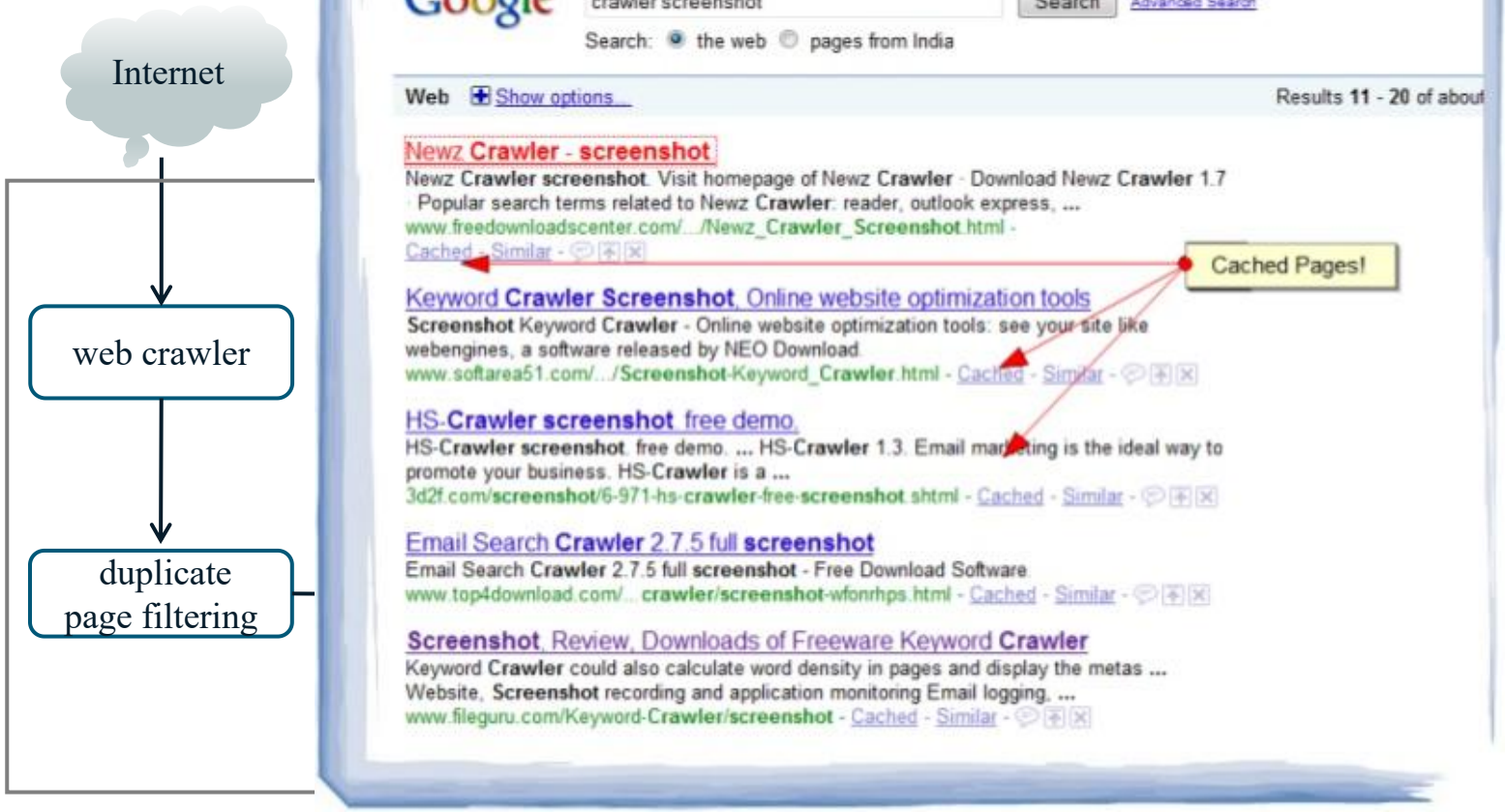
- cleaning the raw data sets
- getting the final data sets
- sorting the results

- analyzing the structural and logical characteristics of HTML pages
- identifying the meaning of data unit

- the communication interface between model and Web sites

☞Research methods——Collection

- General search
  - an Internet portal through which a person can search numerous compiled resources for topic-related information.
  - include **Google, Yahoo!, Alta Vista and Bing**
  - to perform a se



☞Research methods——Collection

- Meta-search Engines
  - a search tool that uses another search engine's data to produce their own results.
  - take input from a user and simultaneously send out queries to third party search

☞Research methods——Collection

- Deep web search

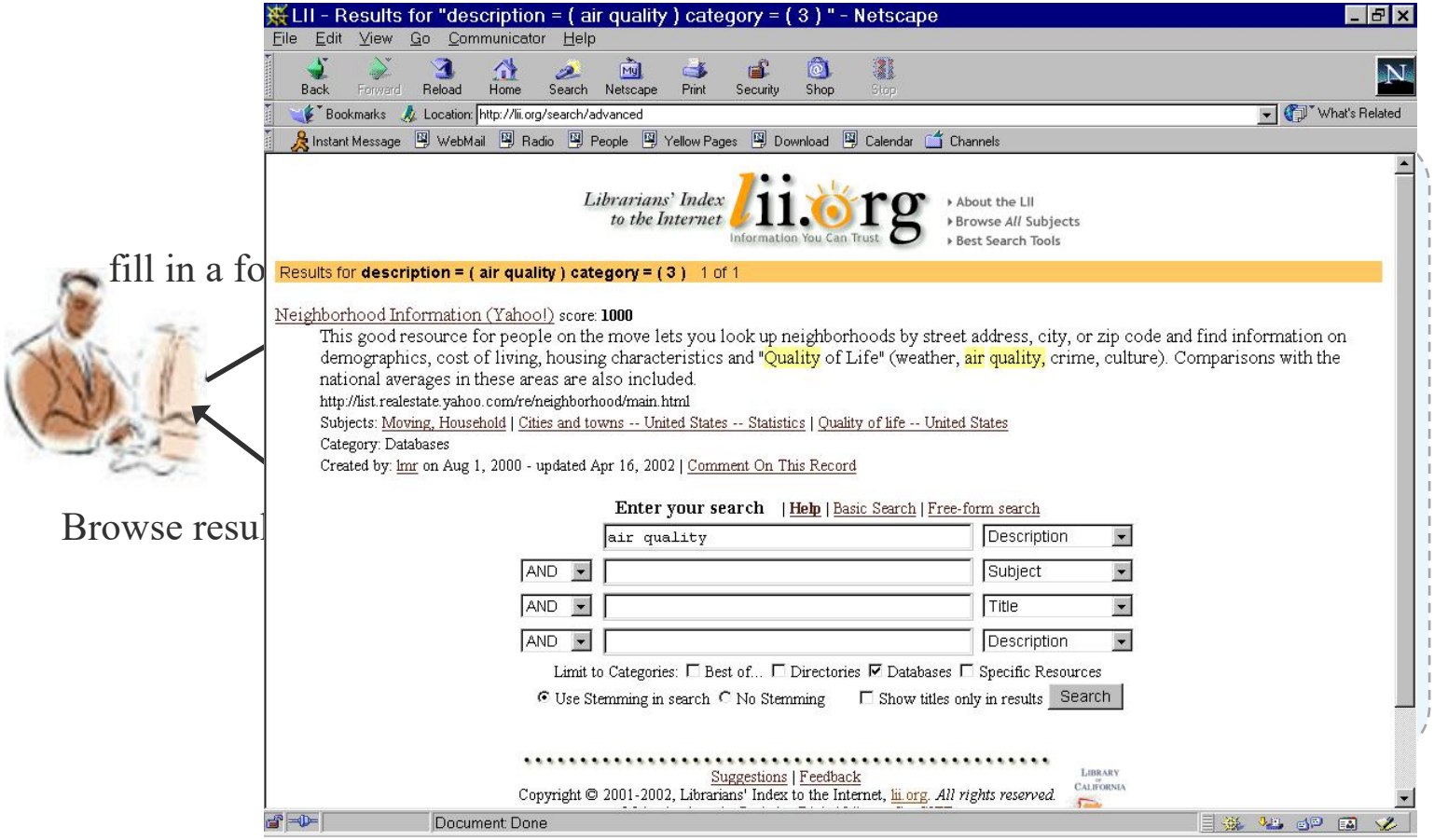
Not Found by Standard Search Engines	Examples
Database content (dynamically generated for a particular inquiry)	Databases such as ERIC, PubMed, Library catalogs
Formats	pdf, audio, video, any new format
Real-time information	Current weather, stock quotations, airline flight checkers
Subscription or fee based services	EBSOChost, INFOTRAC, LexisNexis Academic
Sites requiring password access or registration for use	Blackboard, membership sites
Sites that require that forms be filled out	Travel direction sites, job hunting sites
Sites with a no-index protocol	Private sites

- Deep Web—refers to all web pages that search engines cannot find,
- —such as user databases, registration-required web forums, webmail pages, and pages behind paywalls.
- —content behind web search forms on publicly available pages

☞Research methods——Collection

- Deep web search

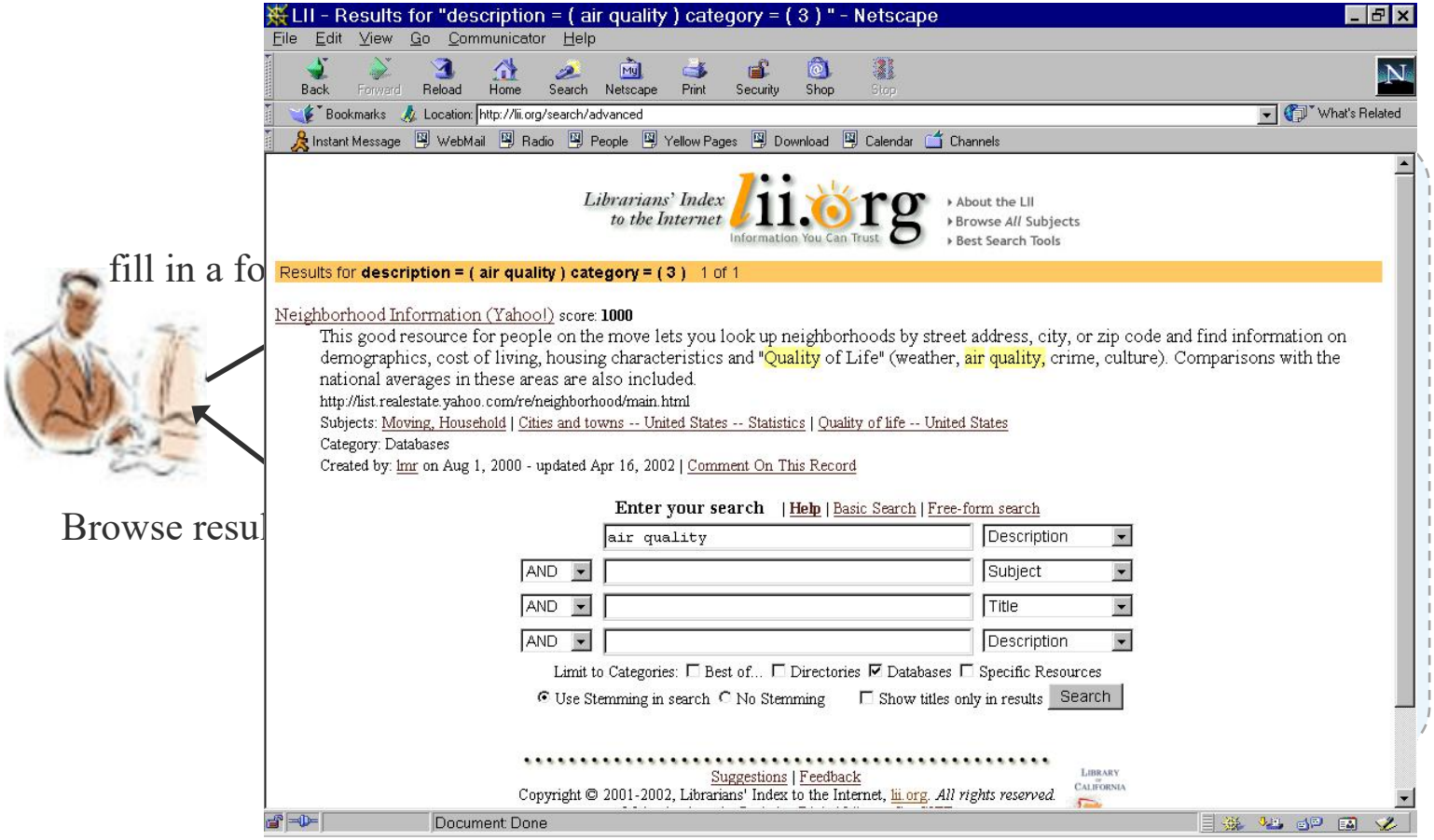
Deep web technologies: aims to achieve web data integration, by automating the process of analyzing database structures and cross referencing the results.



☞Research methods——Collection

- Deep web search

Deep web technologies: aims to achieve web data integration, by automating the process of analyzing database structures and cross referencing the results.

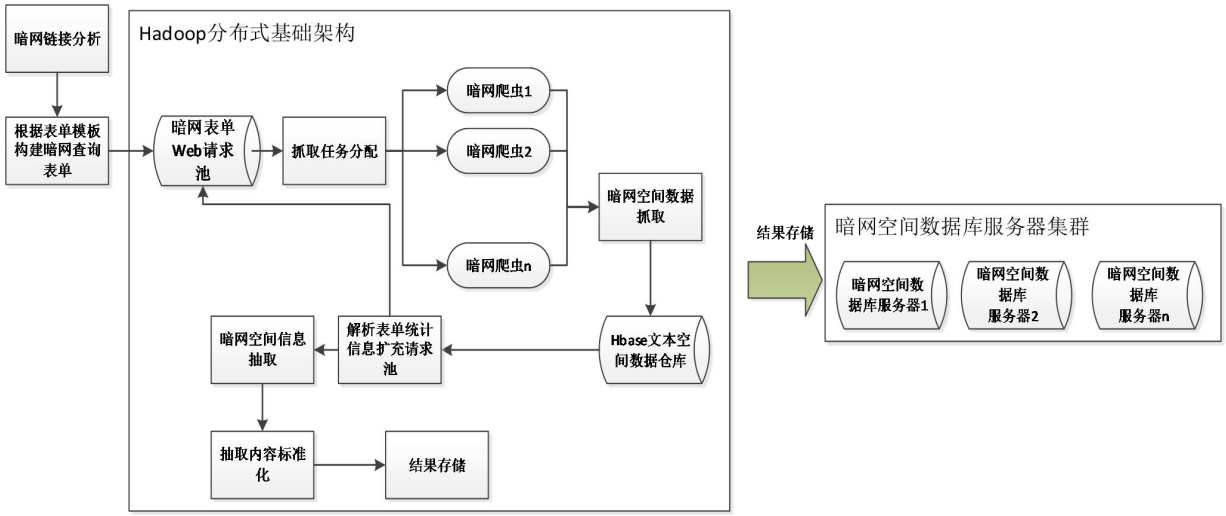


☞Research methods——Collection

- Deep web search

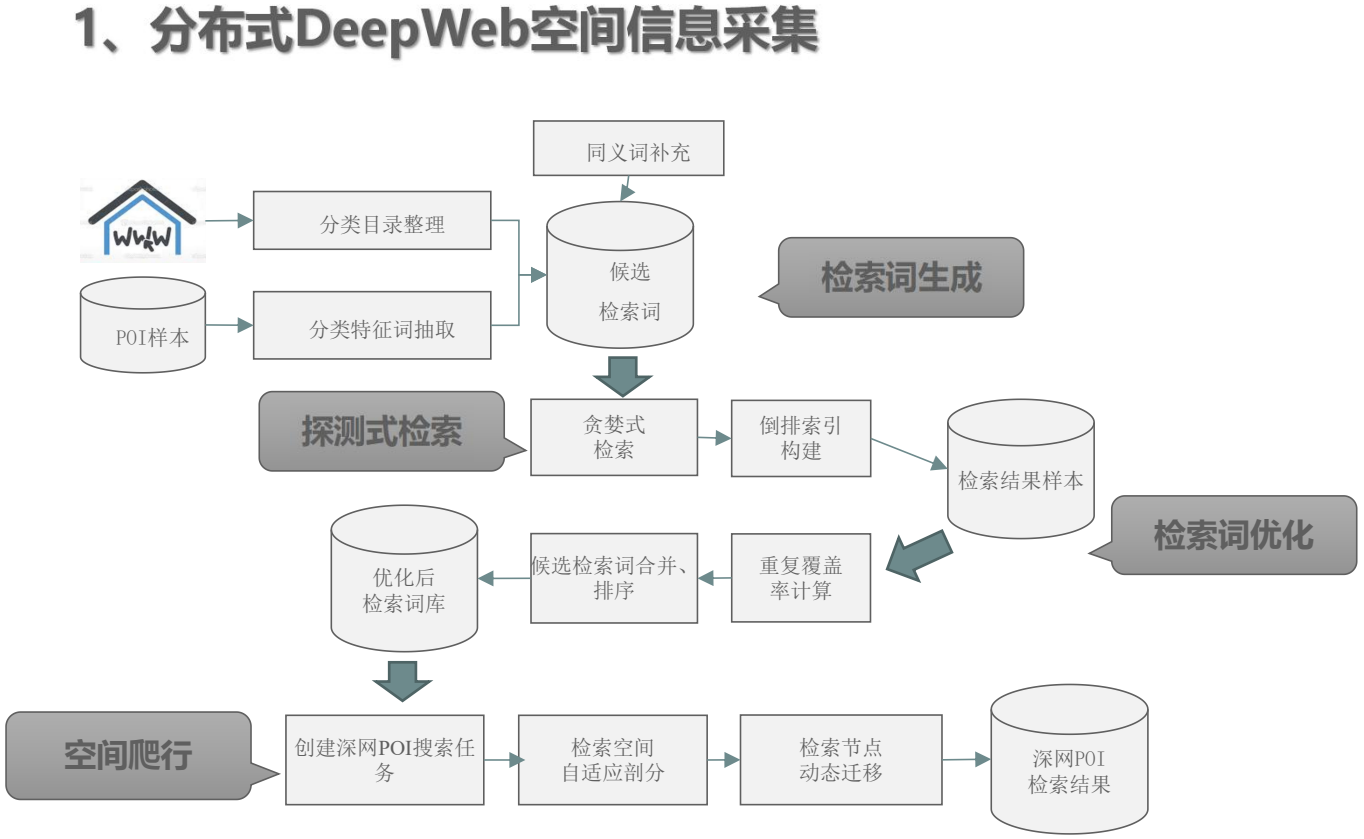
1、分布式DeepWeb空间信息采集

针对DeepWeb的抓取特点，利用Map/Reduce分布式系统架构、Web请求池构建、抓取任务弹性分配技术、异步I/O模型等关键技术，提出并实现暗网空间数据采集。



☞Research methods——Collection

- Deep web search

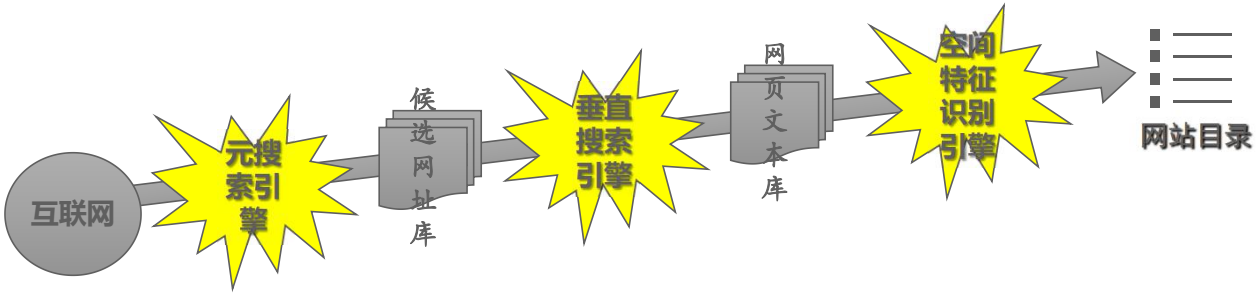


☞Research methods——Collection

- Deep web search

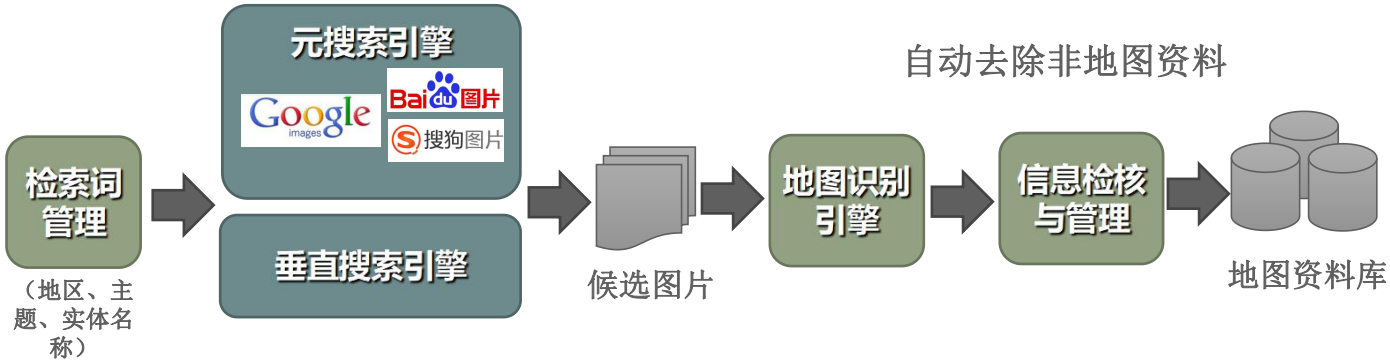
2、地理信息网站特征识别与自动发现

利用垂直搜索、元搜索等关键技术，分析和提取地理信息网站语义特征，构建地图网站发现与判别模型，实现地图网站的自动搜集与提取。

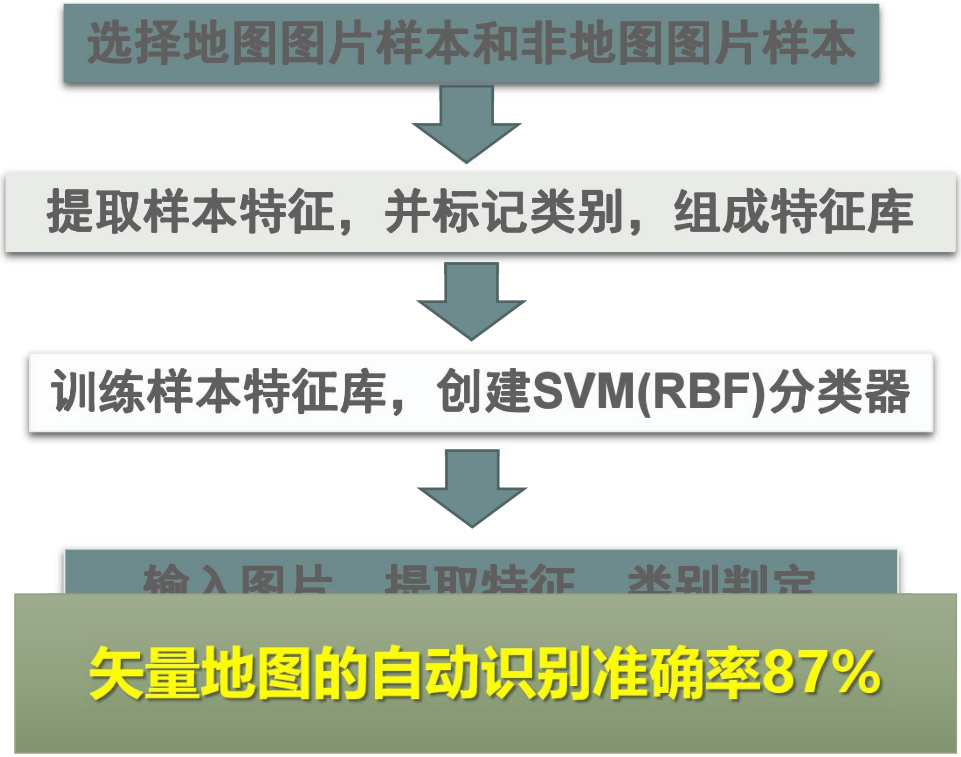


自动发现地理信息网站

3、基于像素统计与纹理特征的地图图像搜索

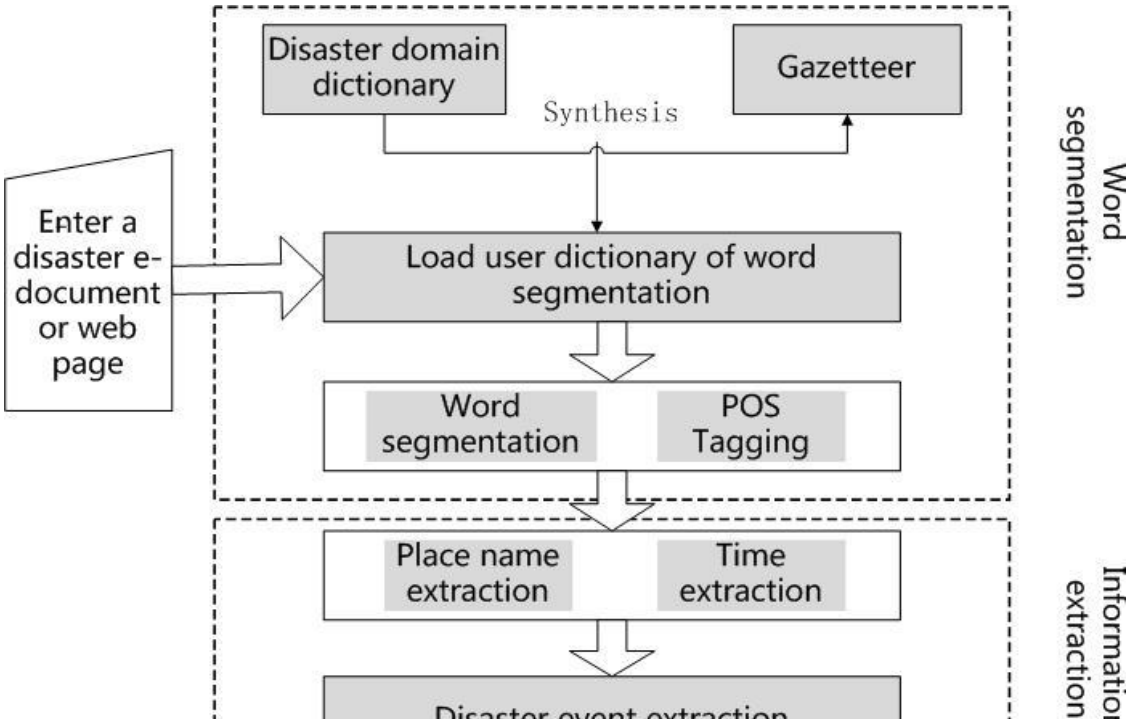


3、基于像素统计与纹理特征的地图图像搜索



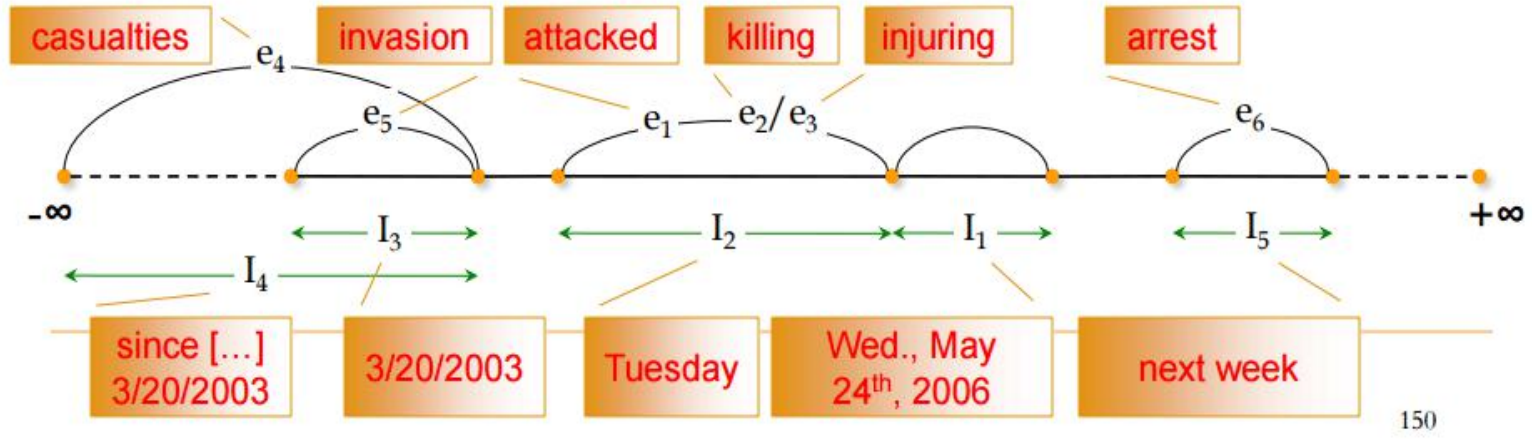
☞Research methods——Information Extraction

- Information Extraction:
  - Information Extraction (IE) is the process of finding structured text from unstructured or semi-structured text by annotating semantic information.
  - **Knowledge base:** Lexical database, Syntactic pattern, Semantic relation
  - **Information extraction model:** Rule models, Machine learning models
  - **Semantic parsing model:** Spatial matching, space-time matching, space-time-attribute matching, space-time-attribute-event matching,



☞Research methods——Information Extraction

- Time Extraction :
  - based on the combination of trigger word and rule model
  - information extraction and filtering process (three levels):
    1. time semantic role tagging
    2. phrase recognition
    3. syntactic pattern matching



Interval Based Event Timeline Construction

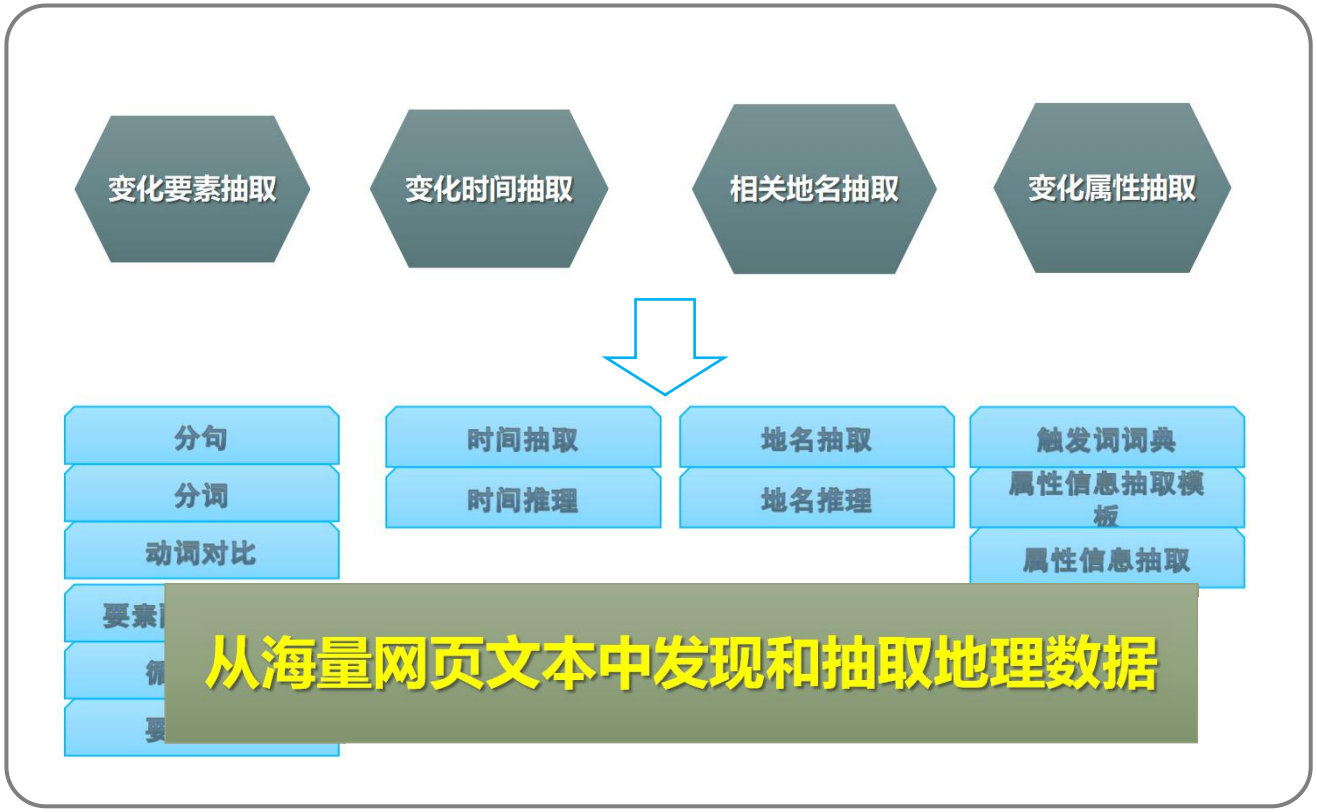
☞Research methods——Information Extraction

- Location Extraction
  - Place name extraction:
    word segmentation and POS tagging
    place name recognition
  - Place name reasoning:
    principle: the nearest distance

- Event Extraction
  - Rule Learning based Approach:
    dictionary based method, rule based method, wrapper induction.
  - Classification Model based Approach:
    Support Vector Machines (SVMs)
  - Sequence Labeling based Approach:
    Hidden Markov Models (HMMs) , Maximum Entropy Markov Models (MEMMs) , Conditional Random Fields (CRFs)

☞Research methods——Information Extraction

- Geographical Entity Extraction:

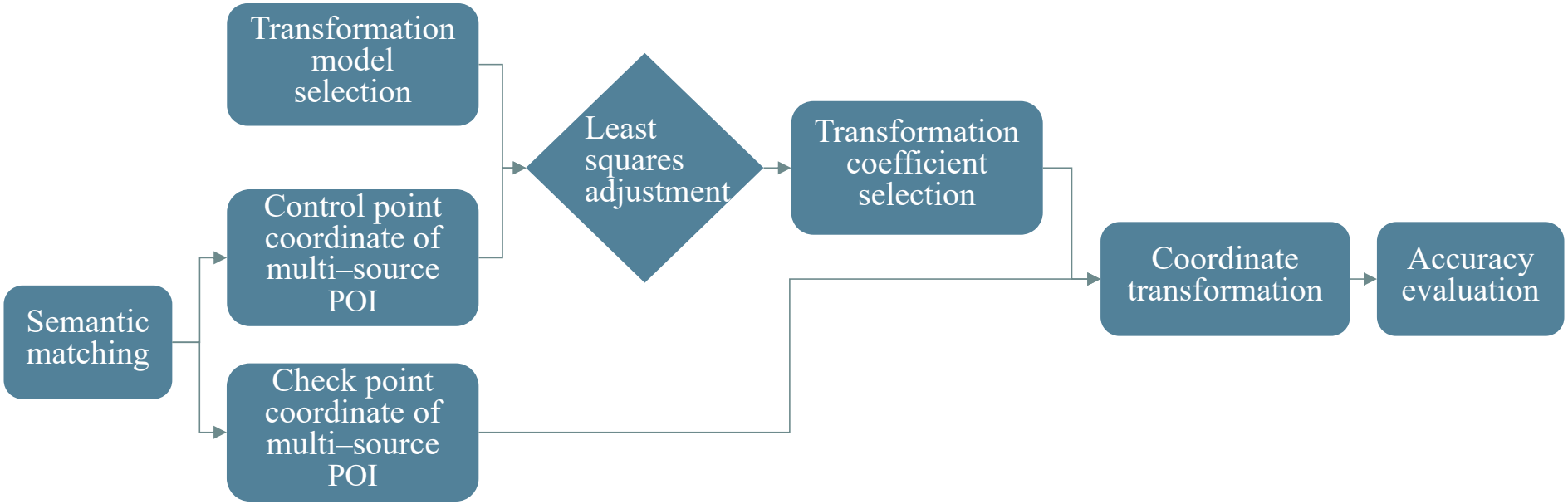


☞Research methods——Information integration

➢ Location Integration—Spatial position correction of multi–source POI

There exists systematic deviation in the spatial position of different POI data.

- Obtaining the coordinate information of the control point
- Calculating the parameters of the coordinate transformation model
- Evaluating the spatial position correction accuracy.



Semantic-aided spatial location correction scheme for multi-source POI

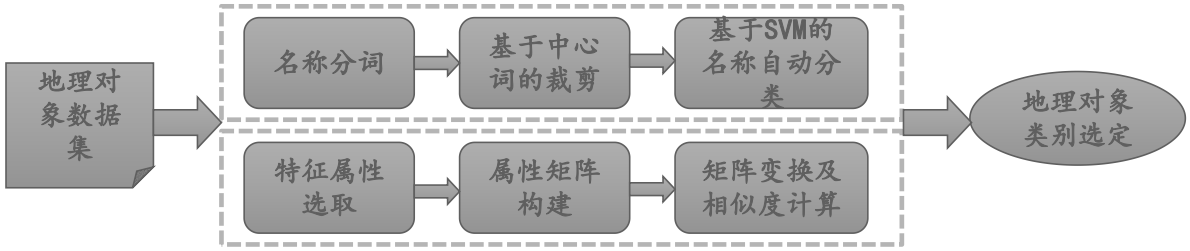
☞Research methods——Information integration

- Semantic Integration
  - bring together information from diverse sources
  - link them by using semantic information
  - **three dimensions of semantic-integration research:**

Mapping discovery	• Using a Shared Ontology • Using Heuristics and Machine-learning
Representation of mappings	• representing mappings as instances in an ontology of mappings • using views to describe mappings
Reasoning with mappings	• reasoning over the source ontologies and the mappings

4、地理对象语义分类

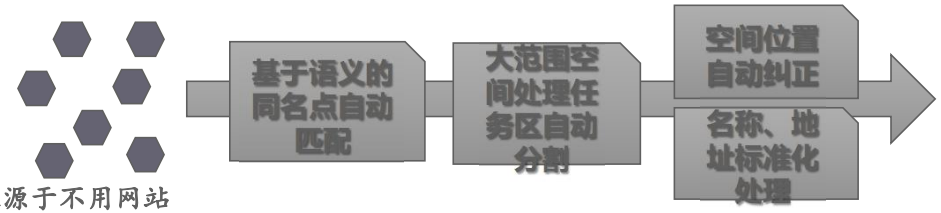
开展了兼顾文本与空间特征属性的地理对象语义分类技术研究，利用中文文本处理、语义相似度计算、空间分析等关键技术，实现基于角色标注的地理对象语义分类。



自动分类+跨网站类别转换

5、多源地理信息数据融合处理

根据要素名称、描述信息、地理坐标等特征，建立统一的资源分类体系，实现多源信息的**名称消歧**和**地址信息标准化**；建立非线性坐标映射转换模型，实现**多源POI数据位置纠正**。



- 来源于不用网站
- 命名模式不用
- 地址表达不同
- 空间位置存在差异

- 融合为统一数据库
- 同名目标合并
- 地址差异消除
- 空间位置纠正

持续整合多网站地理目标

Typical experiments

Experimental research

- ✓ Geographical Information Acquisition and Processing from Deep Web
- ✓ Background Map Search System
- ✓ Internet-based Information Extraction
- ✓ Integration of the Fundamental Geographical Information on the Internet

Typical experiments

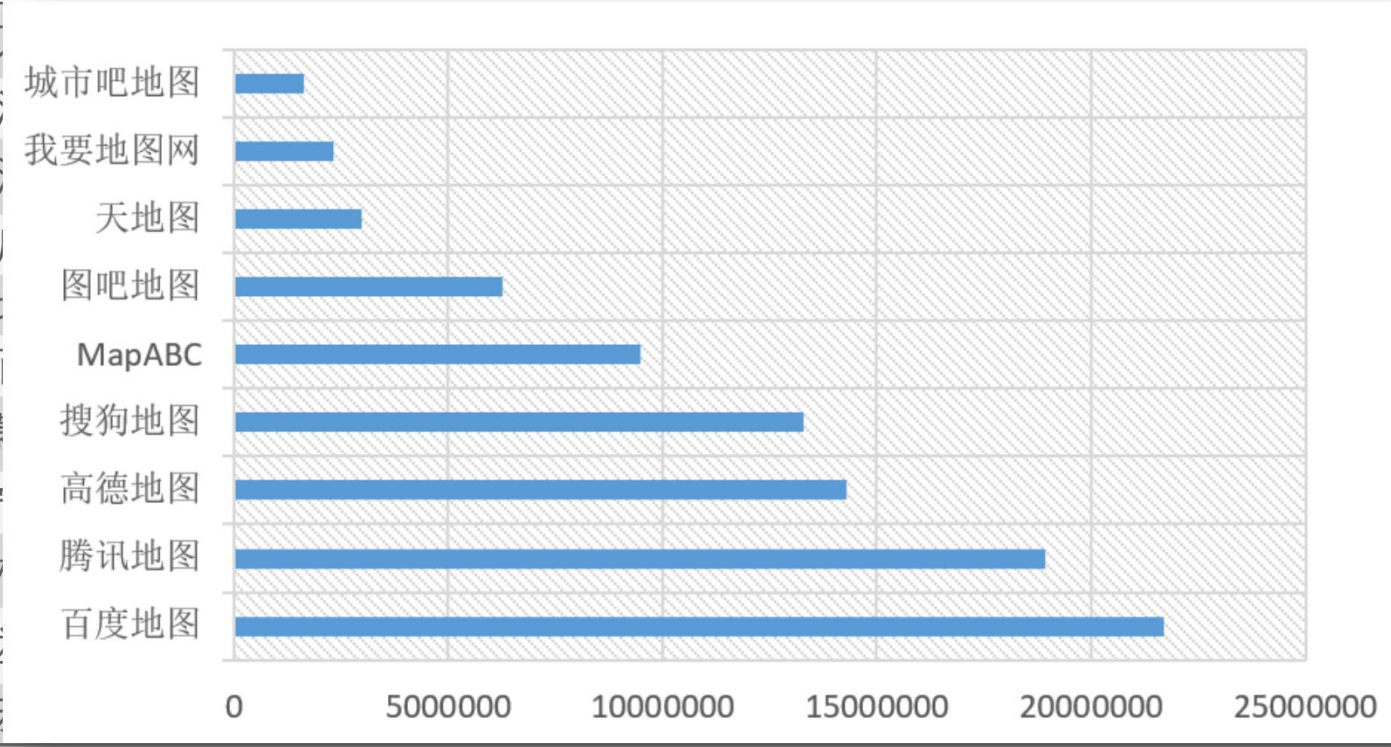
- Geographic Information Acquisition and processing from deep web pages
  - POI search engine:

Typical experiments

- Geographic Information Acquisition and processing from deep web pages

Statistics on the number of POI in different provinces:

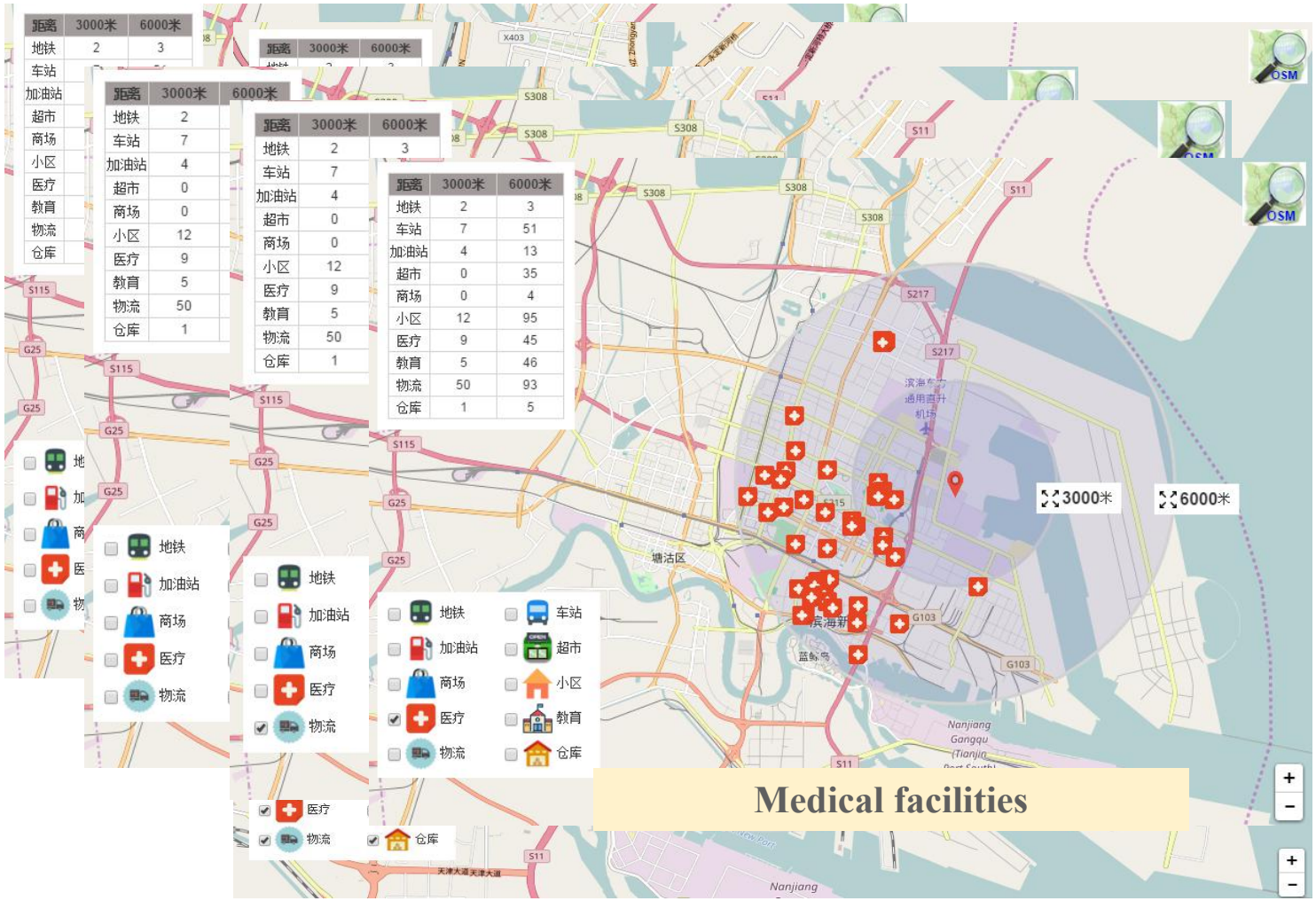
序号	地区	数量	序号	地区	数量	序号	地区	数量
1	北							
2	天							
3	上							
4	重							
5	河							
6	山							
7	内蒙							
8	辽							
9	吉							
10	黑龙							
11	江							



- **Crawling search test was conducted on the 9 kinds of target website.**
- **Obtained a total of about 90.87 million POI (except Taiwan) within three months.**
- **POI number of Baidu, Gaode, Tencent and Sogou is respectively more than 10 million.**

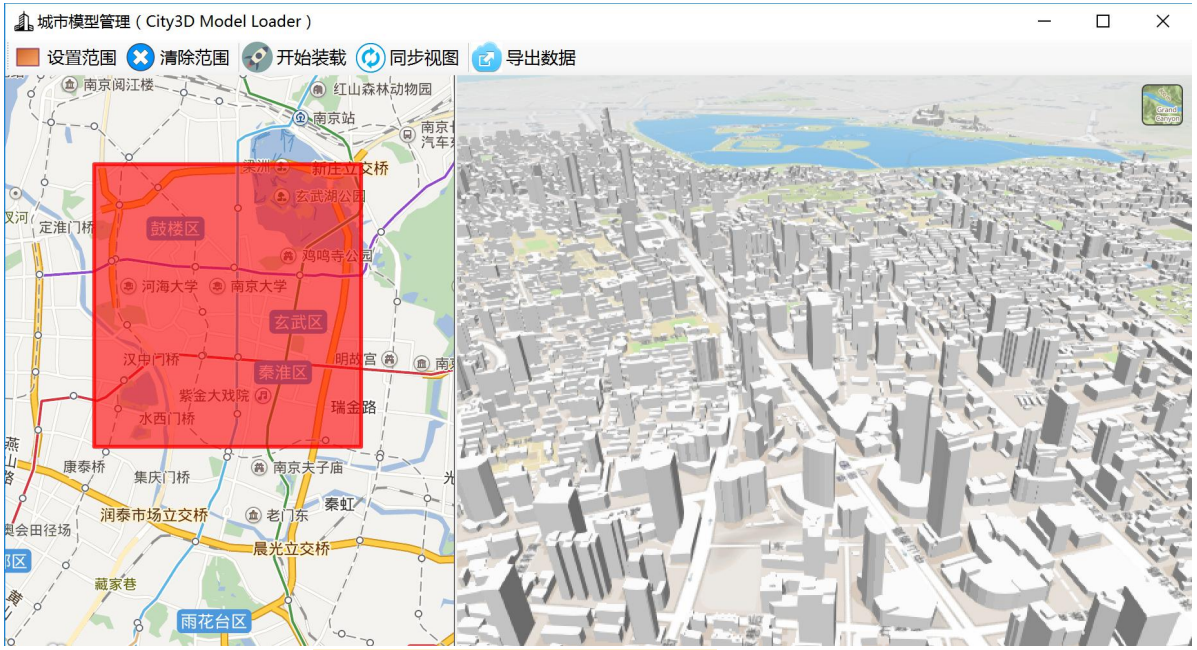
Typical experiments

The distribution of POI in 3km and 6km around the site of explosion in Binhai New Area, Tianjin



Typical experiments

- Geographic Information Acquisition and processing from deep web pages
  - Urban construction information acquisition:



Nanjing

Shanghai

Beijing

Typical experiments

- Disaster events search:

event name | content | time | URL | Area | Location

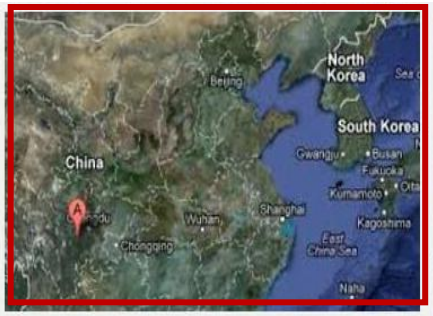
kTi	kC nt	Tin Ori	Time_Dig	RL	Ad s_Ori	X	7
海淀区发生一起1人死亡事故	12月22日15时30分，在海淀区温泉镇双坡路路北北京科弘投资管	12月22日15时30分	12/22/2015 15:30:00	http://www.bjsafety.gov.cn/accidentin	北京市海淀区	116.3218	39.95606977
海淀区发生一起1人死亡事故	12月18日10时许，在海淀区清华大学化学系实验室发生一起爆炸	12月18日10时	12/18/2015 10:00:00	http://www.bjsafety.gov.cn/accidentin	北京市海淀区	116.3218	39.95606977
顺义区发生一起1人死亡事故	12月18日8时50分,顺义区北京现代汽车有限公司一厂废料处理厂	12月18日8时50分	12/18/2015 08:50:00	http://www.bjsafety.gov.cn/accidentin	北京市顺义区	116.6679	40.128491057
通州区六环一甲醇罐车遭拉煤车追	12月17日9时15分，在通州区六环徐辛庄内环路，一拉煤货车追	12月17日9时15分	12/17/2015 09:15:00	http://www.bjsafety.gov.cn/accidentin	北京市通州区	116.6912	39.896040779
西城区发生不明气体情况	12月3日中午，北京四中初中部周边有疑似汽油味道,西城区安监	12月3日	12/3/2015 00:00:00	http://www.bjsafety.gov.cn/accidentin	北京市西城区	116.3534	39.8928799
海淀区发生一起1人死亡事故	11月16日8时许，在海淀区西三环花园桥东约200米路北，北京	11月16日8时	11/16/2015 08:00:00	http://www.bjsafety.gov.cn/accidentin	北京市海淀区	116.3218	39.95606977
海淀西北旺发生一起坍塌事故	北京龙泉机动车驾驶员培训中心承租的西北旺甲1号空地南侧山	10月30日16时30分	10/30/2015 16:30:00	http://www.bjsafety.gov.cn/accidentin	北京市	116.3946	39.894104004
门头沟区发生一起高处坠落事故	10月11日13时许，在北京中铁西城工程项目工地,工人聂某某（	10月11日13时	10/11/2015 13:00:00	http://www.bjsafety.gov.cn/accidentin	北京市门头沟	116.0980	39.932906923
海淀区发生一起高处坠落事故	10月11日14时50分，在中关村大街59号人大逸夫会堂室外设施	10月11日14时50分	10/11/2015 14:50:00	http://www.bjsafety.gov.cn/accidentin	北京市海淀区	116.3218	39.95606977
海淀区发生一起工地砸伤1人死亡	10月11日11时40分，在海淀区清河毛纺路阳光南里32号楼南侧	10月11日11时40分	10/11/2015 11:40:00	http://www.bjsafety.gov.cn/accidentin	北京市海淀区	116.3218	39.95606977
通州区发生一起一人死亡事故	10月6日16时许，北京岱摩斯变速器有限公司委托现代重工对本	10月6日16时7时	10/6/2015 16:00:00	http://www.bjsafety.gov.cn/accidentin	北京市通州区	116.6912	39.896040779
平谷区发生一起一人死亡事故	10月3日16时许，在平谷区山东庄镇桥头营村西北京浩天荣华商	10月3日16时17时	10/3/2015 16:00:00	http://www.bjsafety.gov.cn/accidentin	北京市平谷区	117.1279	40.137940577
大兴区发生一起2死2伤事故	10月1日10时20分,北京康家兴旺物业管理有限公司工人肖自叶、	10月1日10时20分	10/1/2015 10:20:00	http://www.bjsafety.gov.cn/accidentin	北京市大兴区	116.3520	39.747321955
朝阳区发生一起1人死亡事故	10月1日13时30分许,在朝阳区建外金地中心A座,北京中建华宇	10月1日13时30分	10/1/2015 13:30:00	http://www.bjsafety.gov.cn/accidentin	北京市朝阳区	116.4976	39.919742721
朝阳区发生一起燃气泄漏事故	9月30日8时40分，在朝阳区东坝乡东坝南二街二期3号标段市政	9月30日8时40分	9/30/2015 08:40:00	http://www.bjsafety.gov.cn/accidentin	北京市朝阳区	116.4976	39.919742721
朝阳区发生一起1人死亡事故	9月12日13时许，北京市朝阳安贞地区的808工程工地,山东省肥	9月12日13时16时	9/12/2015 13:00:00	http://www.bjsafety.gov.cn/accidentin	北京市朝阳区	116.4976	39.919742721
房山区发生一起一死一伤事故	9月5日18时10分许，北京中铁房山桥梁有限公司生产区轨枕车间	9月5日18时10分	9/5/2015 18:10:00	http://www.bjsafety.gov.cn/accidentin	北京市房山区	116.1735	39.732537984
怀柔区发生一起1人死亡事故	8月18日1时30分，在怀柔区雁栖经济开发区北京威亚高性能纤维	8月18日1时30分	8/18/2015 01:30:00	http://www.bjsafety.gov.cn/accidentin	北京市怀柔区	116.6693	40.313043209
顺义发生一起1人死亡事故	15日14时30分许，在顺义区南法信镇焦各庄村南30米，钩机司机	15日14时30分	8/15/2015 14:30:00	http://www.bjsafety.gov.cn/accidentin	北京市顺义区	116.6679	40.128491057
朝阳区发生一起1人死亡事故	【事故信息】8月4日21时04分，接市应急办通报：8月3日16时	8月4日21时04分	8/4/2015 21:04:00	http://www.bjsafety.gov.cn/accidentin	北京市朝阳区	116.4976	39.919742721
昌平区阳坊镇一饭馆煤气罐爆燃	7月29日7时38分，阳坊镇政府对面一驴肉火烧店（北京昌阳幸福	7月29日7时38分	7/29/2015 07:38:00	http://www.bjsafety.gov.cn/accidentin	北京市昌平区	116.2628	40.216635475
海淀区发生一起触电死亡事故	7月28日14时50分左右，在海淀区中关村东路苏宁联想桥店，北	7月28日14时50分	7/28/2015 14:50:00	http://www.bjsafety.gov.cn/accidentin	北京市海淀区	116.3218	39.95606977
大兴区发生一起火灾事故	7月19日14时10分许，在通州区梨园镇大稿村北京市大稿京苑加	7月19日14时10分	7/19/2015 14:10:00	http://www.bjsafety.gov.cn/accidentin	北京市通州区	116.6912	39.896040779
通州区发生一起一人受伤事故	7月18日16时左右，在通州区台湖镇，东城区安置房工地三层彩	7月18日16时	7/18/2015 16:00:00	http://www.bjsafety.gov.cn/accidentin	北京市东城区	116.4056	39.887611445
大兴区发生一起火灾事故	13时20分，大兴区狼垡北京铁路局动车段院内，中铁七局下属	13时20分	1/1/0001 00:00:00	http://www.bjsafety.gov.cn/accidentin	北京市大兴区	116.3520	39.747321955
昌平区发生一起一人死亡事故	7月17日19时许，在昌平区阳坊阅兵训练基地北门外约1000米处	7月17日19时	7/17/2015 19:00:00	http://www.bjsafety.gov.cn/accidentin	北京市昌平区	116.2628	40.216635475
房山区发生一起氨水泄漏事故	7月15日20时50分，位于房山区燕化公司炼油三厂石蜡成型装置	7月15日20时50分	7/15/2015 20:50:00	http://www.bjsafety.gov.cn/accidentin	北京市房山区	116.1735	39.732537984
丰台区发生一起火灾事故	7月12日16时13分，在丰台区长辛店南岗洼村宏达经典有限公司	7月12日16时13分	7/12/2015 16:13:00	http://www.bjsafety.gov.cn/accidentin	北京市丰台区	116.3184	39.852302354
房山区发生一起交通事故	7月7日14时许，北京燕飞蓝天润滑油厂从			.bjsafety.gov.cn/accidentin	北京市房山区	116.1735	39.732537984
西城区发生一起触电事故	7月6日15时左右，中国联通公司的施工单			.bjsafety.gov.cn/accidentin	北京市西城区	116.3534	39.8928799
朝阳区发生一起火灾事故	7月7日11时30分许，朝阳区建外街道永安			.bjsafety.gov.cn/accidentin	北京市朝阳区	116.4976	39.919742721
密云县发生一起死亡事故	6月22日15时30分许,在密云县溪翁庄镇京			.bjsafety.gov.cn/accidentin	北京市密云县	116.8615	40.365381076
朝阳区发生一起高坠事故	11时30分许，朝阳区黑庄户乡双楠村280号在建库房,北京兴龙伟	11时30分20时	1/1/0001 00:00:00	http://www.bjsafety.gov.cn/accidentin	北京市朝阳区	116.4976	39.919742721
丰台区发生一起煤气罐泄漏事故	17时15分，丰台区卢沟桥乡晓月园三里一海鲜饺子馆后厨液化石	17时15分	1/1/0001 00:00:00	http://www.bjsafety.gov.cn/accidentin	北京市丰台区	116.3184	39.852302354
丰台区发生一起罐车泄漏事故	6月18日15时30分左右,在丰台区王佐镇南宫459路公交场站内，	6月18日15时30分	6/18/2015 15:30:00	http://www.bjsafety.gov.cn/accidentin	北京市丰台区	116.3184	39.852302354
延庆县发生一起1人死亡事故	6月19日20时许，延庆县城中路广场配电室内一人触电身亡，该	6月19日20时20时	6/19/2015 20:00:00	http://www.bjsafety.gov.cn/accidentin	北京市延庆县	116.0087	40.458442001
房山区发生一起爆燃事故	6月14日9时30分许 ，房山区良乡拱辰大街9号“首品一家	6月14日9时30分	6/14/2015 09:30:00	http://www.bjsafety.gov.cn/accidentin	北京市房山区	116.1735	39.732537984
石景山区发生一起3人受伤事故	31日19时40分，中铁建设集团工程资源分公司在石景山苹果园	31日19时40分	1/1/0001 00:00:00	http://www.bjsafety.gov.cn/accidentin	北京市石景山	116.1890	39.923428953

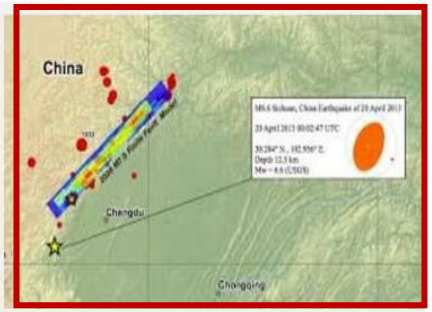
disaster data crawling result

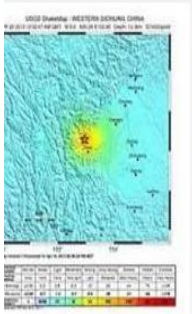
Typical experiments

- Background map images search














How to automatically and accurately identify the map images about disaster?

Typical experiments

- Background map images search

Residential area map images aroud the site of explosion in Tianjin

外，还包括增设

Typical experiments

- Internet-based Disaster Information Extraction

Conditional Random Fields(CRF) was used to identify the event:

- Preprocessing: Word Segmentation
- Training for Information Extraction Model : used CRF++ tool

Uniform framework of disaster events

time
place
losses
casualties
rescue
cause
compensation

```
灾难性事件
{
  时间;地点;
  损失 keywords:倒塌 损坏 受损 毁坏 破坏
  伤亡 keywords:死亡 受伤 蒙难 丧生 重伤 下落不明 失踪 遇难 遇难者 罹难 丧命
  援助 keywords:救援 抢救 搜救 抢救 援助 营救 施救 疏散 救治 赈灾 援救
  原因 keywords:分析 责任 排除 原因 调查 因素 查清 查找 分析 结果 报告
  赔偿 keywords: 赔偿 理赔 保险 获赔 保险金 投保 购买 赔付
  pattern:
  p1: {*}{袭击|侵袭} 2{sp} {造成} 4{mbar} {人/n|名/q} {蒙难|无家可归|失踪|死亡|
      遇难|丧生|受伤|重伤|轻伤}
  p2: 2{sp} {发生/v} {*}{造成}4{mbar}{人/n}{蒙难|死亡/|失踪/|无家可归|遇难|丧
      生|受伤|重伤|轻伤}
```

Typical experiments

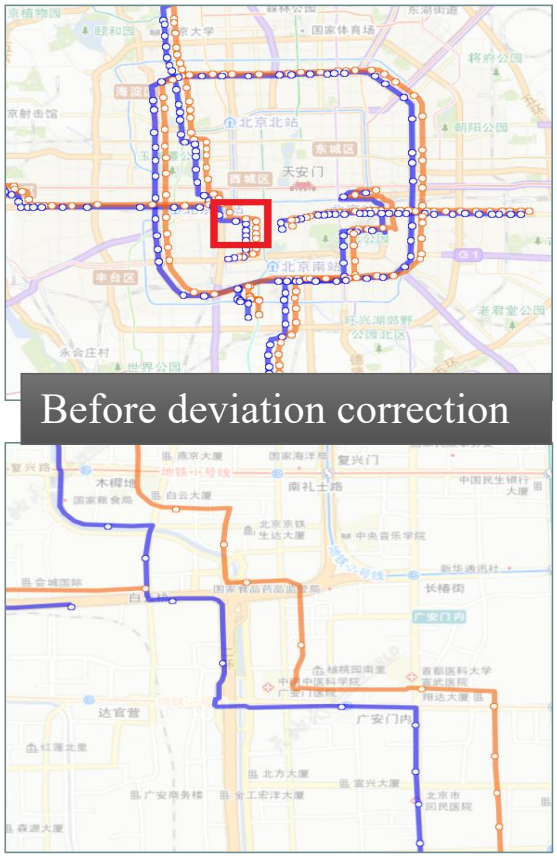
Typical experiments

➢ Integration of the Fundamental Geographical Information on the Internet

Deviation Correction:

Bus lines and POI data from different sources:

- serious spatial position deviation amomg the same points
- after correction, the deviation can be controlled within 10 meters



Typical experiments

- Integration of the Fundamental Geographical Information on the Internet

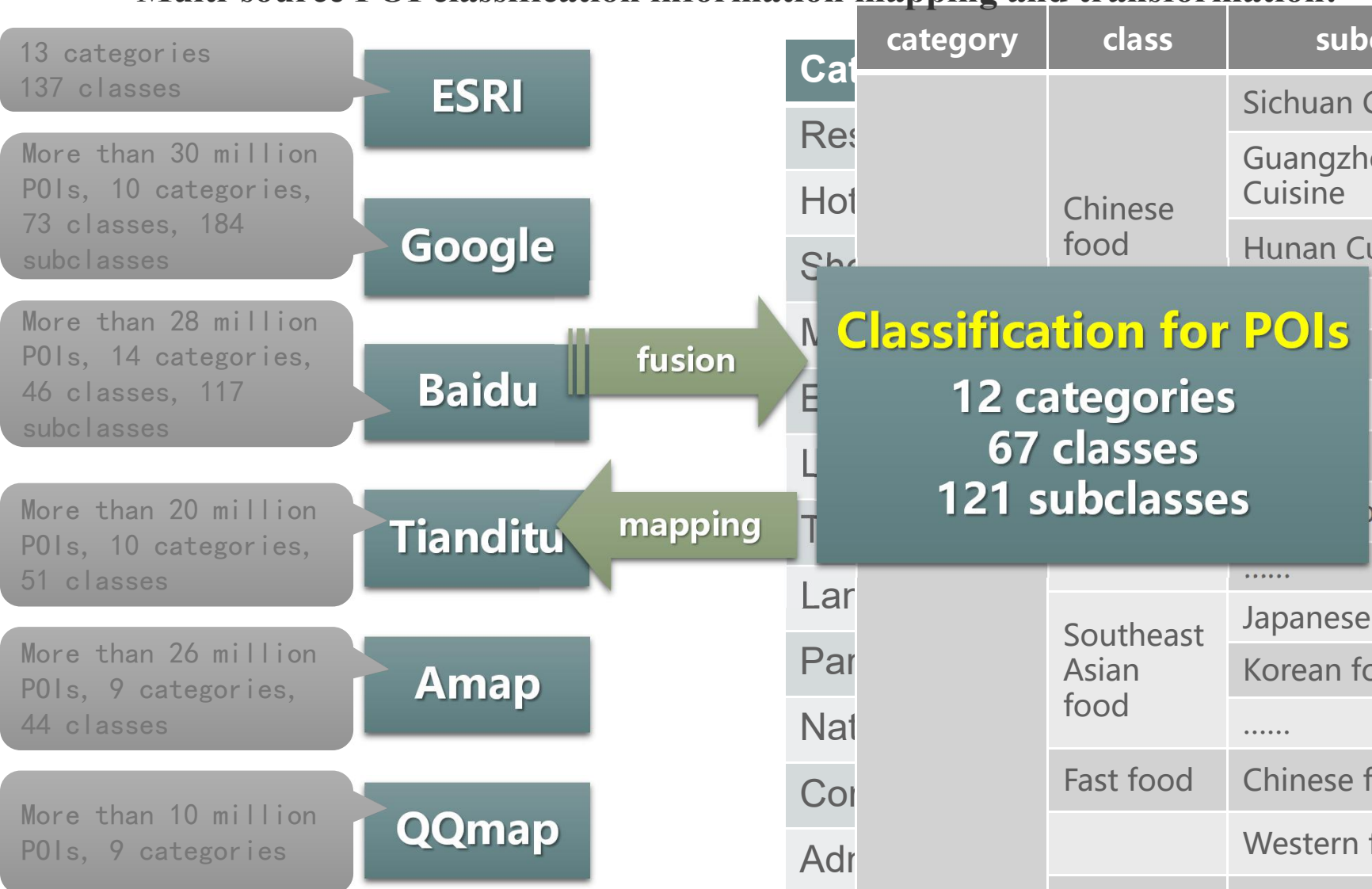
Multi-source POI classification information mapping and transformation:

- Multi-source POI classification information mapping
- Category attribute transformation processing
- Cross-site POI information extraction

Typical experiments

- Integration of the Fundamental Geographical Information on the Internet

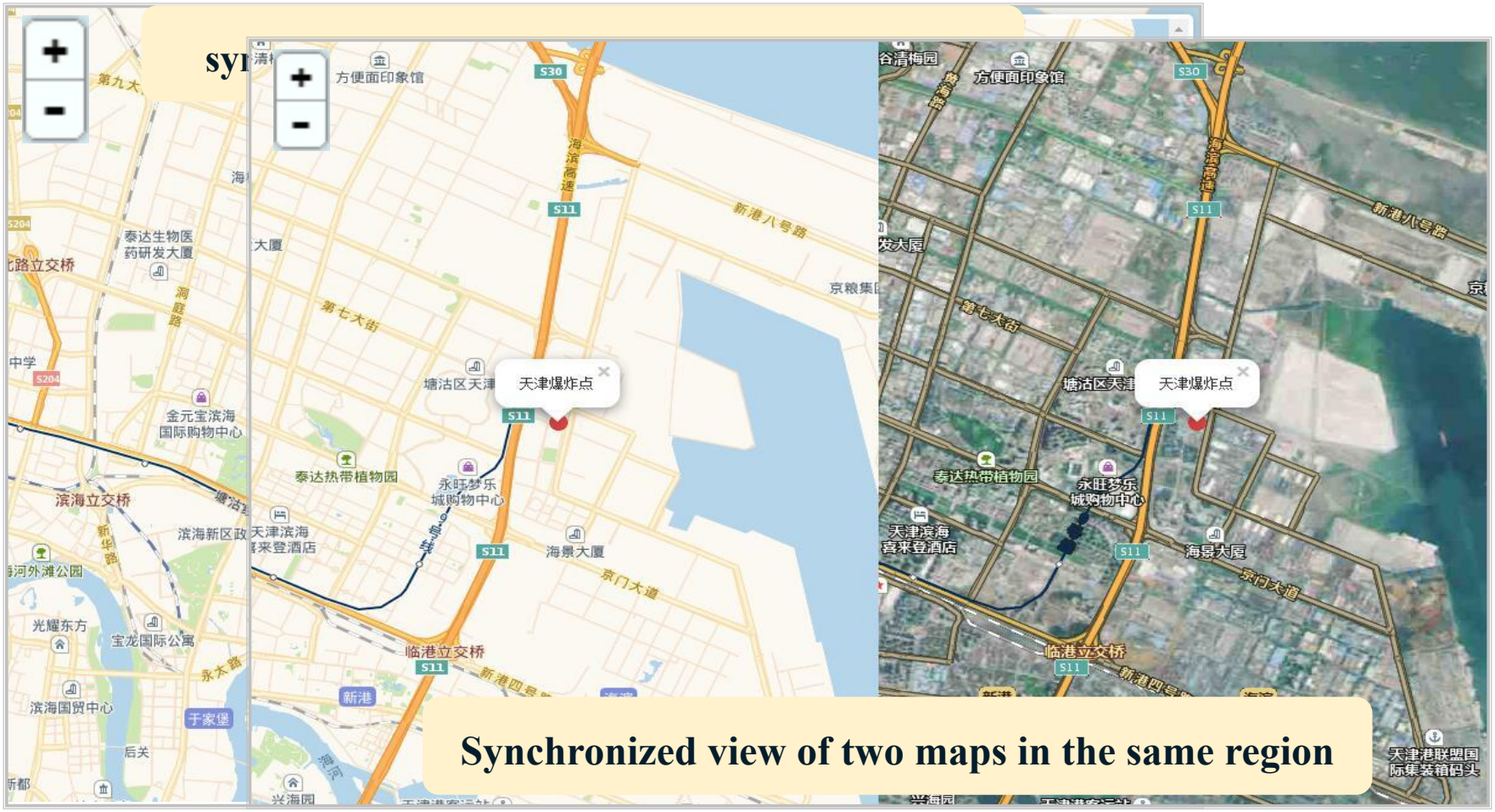
Multi-source POI classification information mapping and transformation:



Typical experiments

- Integration of the Fundamental Geographical Information on the Internet

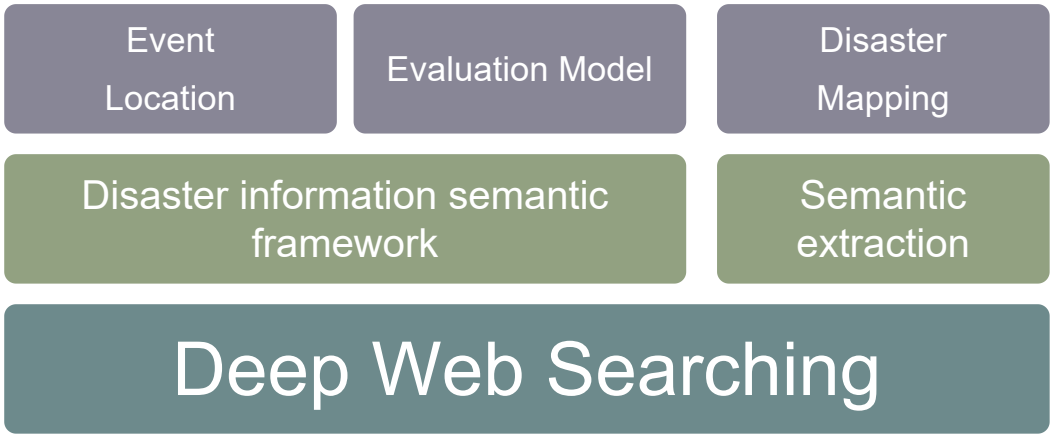
multi-source map data fusion:

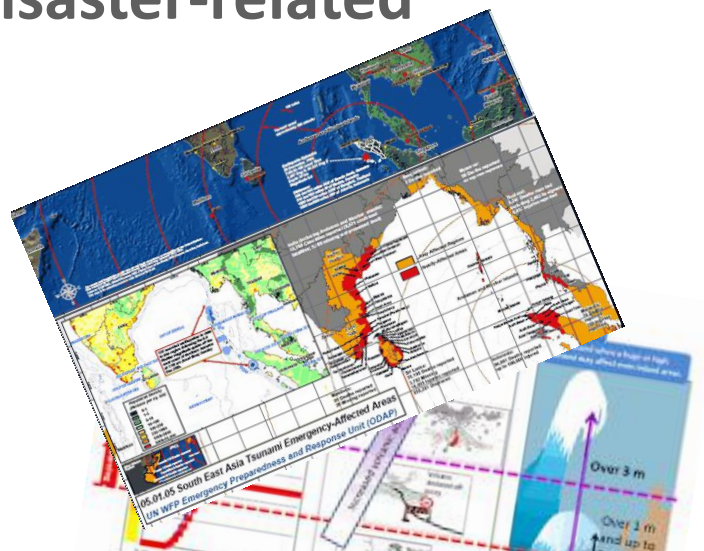


Synchronized view of two maps in the same region

☞Conclusion

1. **Internet geographic information contains huge values for emergency response, damage evaluation and reconstruction.**
2. **For the massive, heterogeneous, unstructured internet geographical information, a radically new technical architecture would be adopted from the data acquiring, cleanning to data fusion and application.**
3. **Through several typical experiments, it is proved that it is feasible to discover and integrate disaster-related geospatial data from Web.**





☞Ongoing work

1.Deep web searching technology

- It is necessary to develop high effective Internet search techniques to improve the speed of content traversal and information extraction.
- intellectualized, personalized, specialized and visualized search

2.Disaster information semantic framework

- A standard mapping framework for understanding and extracting the disaster information

3.Temporal and spatial position description and matching

- transform text description to geo-knowledge

4.Evaluation model

- Evaluate the reliability and availability of information sources, in order to develop the full value of Internet geographical information

5.Effective information integration framework

Thanks