

Information Extraction for Event or Activity Monitoring on Social Media



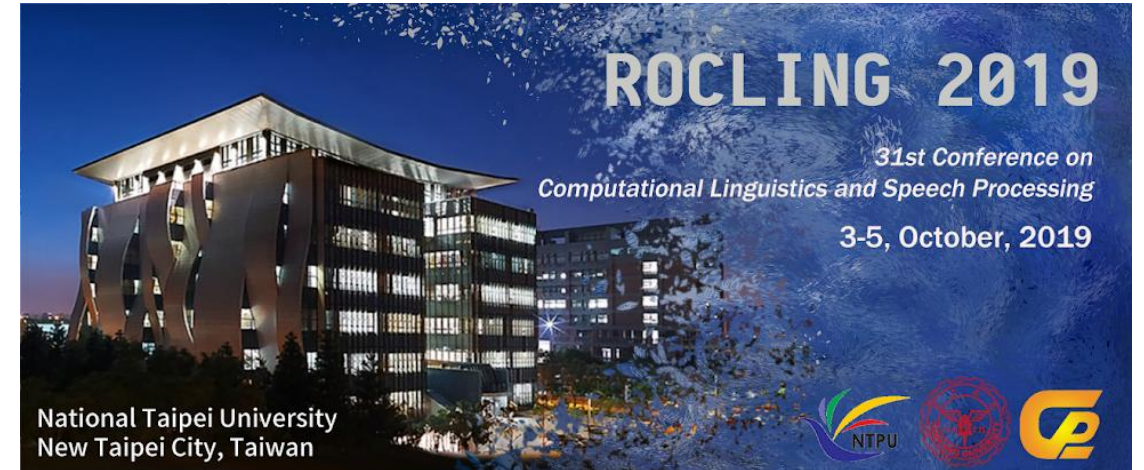
Prof. Chia-Hui Chang (張嘉惠)

WIDM LAB

National Central University, Taiwan

About Me

- [TAAI](#) President (2019-2020)
- [ACLCLP](#) Vice President 2018
 - ACL 2017 Area Co-Chair on IE
 - NAACL 2018 IE Area Co-Chair



Research Work @ WIDM Lab

- Web Named Entity Recognition (NER)
 - From [UGC](#) (user-generated content) on social media
 - Applications
 - [PowerPOI](#): Mining POIs from the Web
 - Damage report monitoring
 - [EventGo!](#) Activity Extraction and Retrieval from Social Network
- Web Data Extraction
 - From [deep web](#) (template pages, semi-structured)
 - Applications
 - [Web ETL API Creator](#)
 - [Mobile Web Creator](#)



Mining POIs from the Web

莊秀敏、張國斌、鄭仲庭、高庭耀、林圓皓



PowerPOI: POI Extraction & Retrieval



[Download PowerPOI \(疾疾店家現身\) APP from Google Play](#)



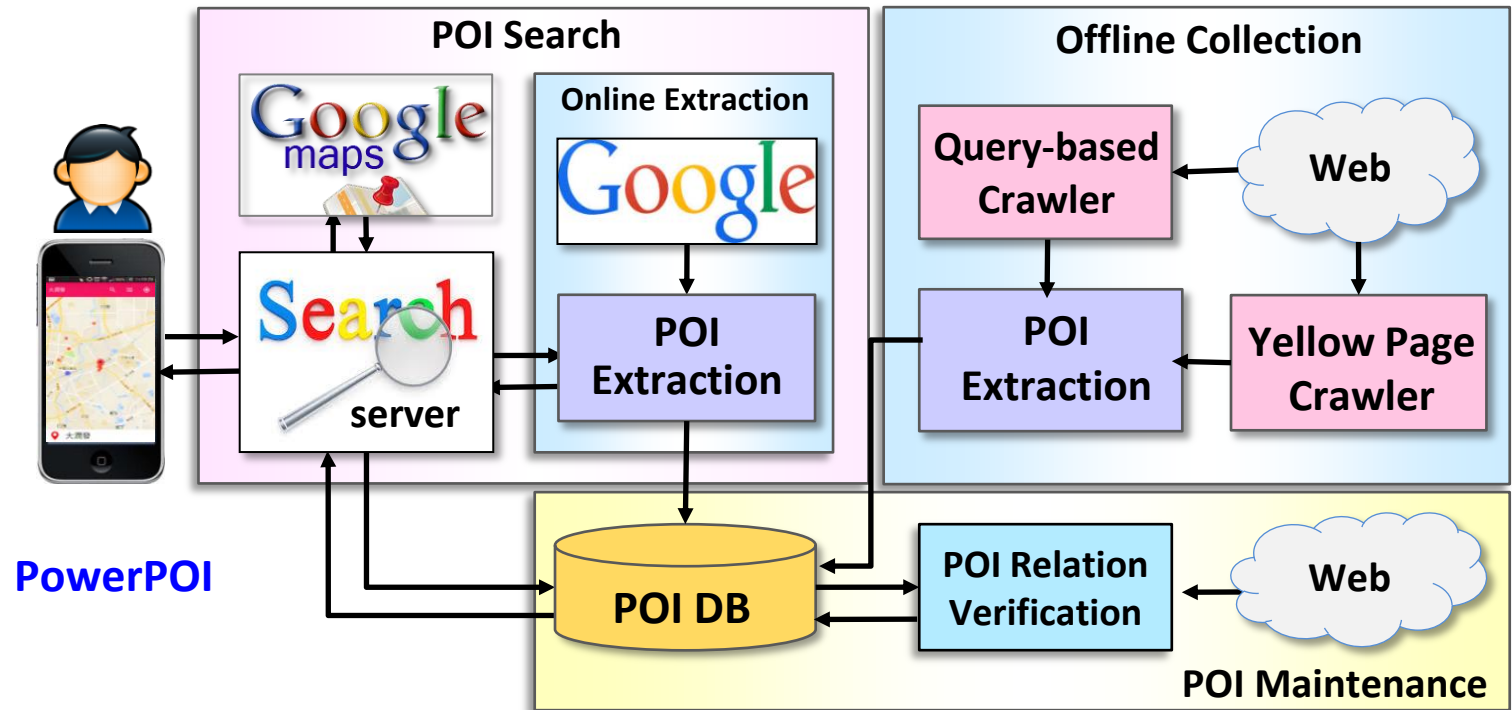
[Download PowerPOI \(疾疾店家現身\) APP from Apple Store](#)

How to Build and Maintain a POI Database

Goal: Automatically extract POIs from Web, construct a POI-DB, and verify POI relations to enable POI search on maps

Task Definition

- POI name
- Address
- Business Hour
- Service description



Enabling Maps/Location Searches on Mobile Devices: Constructing a POI Database via Focused Crawling and Information Extraction, IJGIS, 2016



EventGo Activity Search

林圓皓、程祥恩、莊秀敏



EventGo! -- Activity Extraction and Retrieval



[Download \(Android\)](#)

EventGO! Find Local Activity for In-depth Journey

- Definition of an activity event
 - Activity name
 - Location/Venue
 - Start/End Date
 - Host: organization name
- Sources:
 - 230K FB fanpages in Taiwan
 - Event Post





Damage Report Monitoring

蔣佳峰、胡育維、林圓皓



Damage Report Monitoring

作者r32104565 (JJ)看板BigPeitou標題【轉貼】
梅鄉受難，心如刀割 時間Fri Aug 14 11:11:17
2009

莫拉克颱風狂掃，惡水淹漫、土石流肆虐。

生我、育我的梅仔坑（嘉義縣梅山鄉）竟淪為
嚴重災區。

許多住戶家園全毀，許多鄉親生死不明，連我的
二哥也困在瑞里山上，不能回家。

不只是梅山鄉，整個南臺灣災情慘烈，已超過
九二一大地震，宛如人間煉獄。

Information Extraction Tasks

- Named Entity Recognition
 - Damage
 - Location name
 - Date/Time
- Relation coupling
- Location Positioning

<http://tw.myblog.yahoo.com/jw!s3JxrMyRHADOFXJOEBnfG521/article?mid=2884&prev=-1&next=2867>

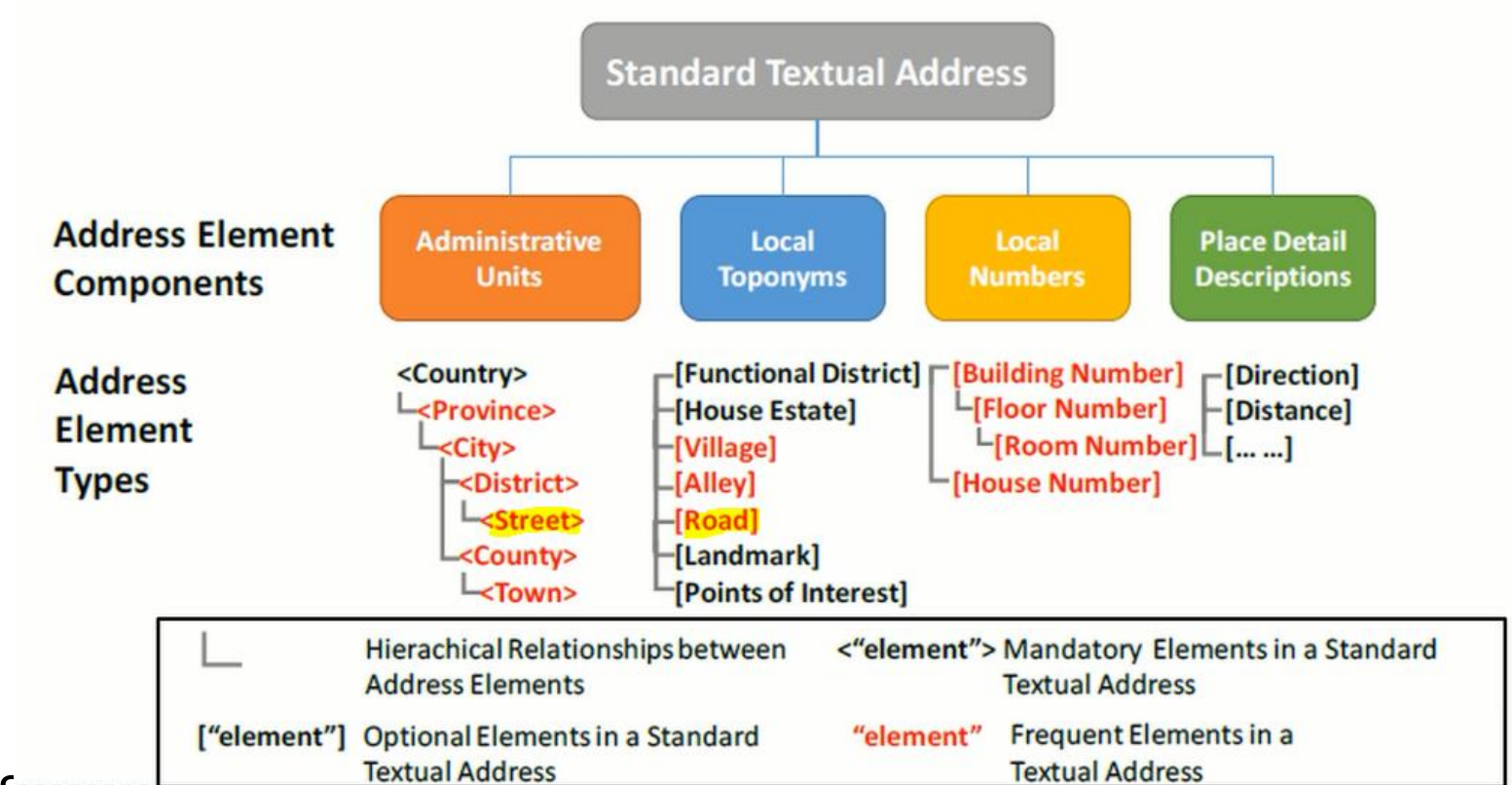
Location Recognition and Geocoding

- Vocabulary

- Location
- Place
- Venue
- Toponym
- Country
- Address

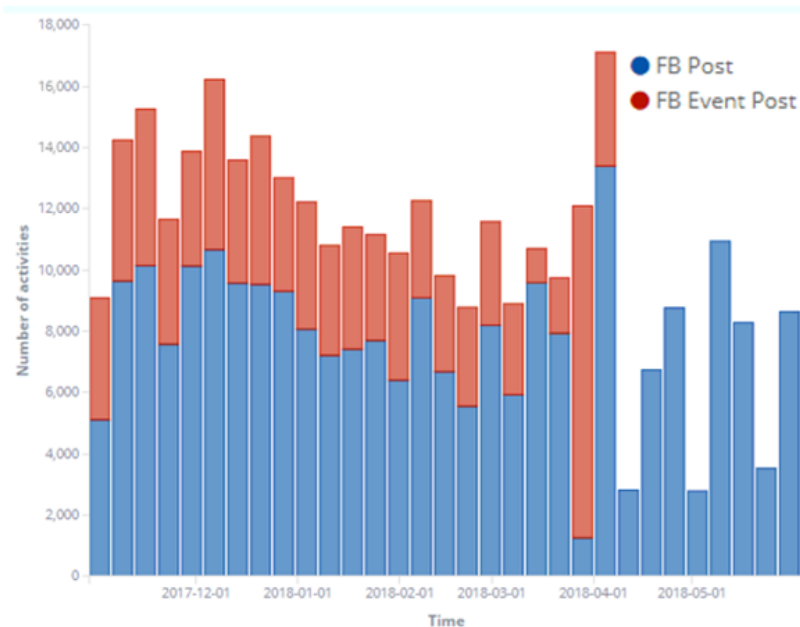
- Geocoding

- May not be a single GPS
- E.g. Central Park



Activity/Event Monitoring & Visualization

- Temporal-Spatial Activity Extraction
- Visualization



Behind the story...

DS4NER: Distant Supervision for
Named Entity Recognition

Named Entity Recognition

- Task: Identifying salient named entities in free text
 - PERSON, PLACE, ORGANIZATION, QUANTITY, DATE/TIME
- EVENT
 - activity name, theorem
- And other **Domain-Specific** fields of interest
 - song, album, drama, movie, book, product names, etc.
- PLACE
 - Location, POI, Toponym, address

DS4NER: Training NER models for New Entities

▣ Issues

- 1) How to collect a lot of good quality training data?
- 2) How to find features for sequence labeling?
- 3) Not enough data? Semi-supervised learning?

▣ Use search engine to collect sentences containing known entities

▣ Automatic Labeling with existing known entities

▣ Mining dictionary based features

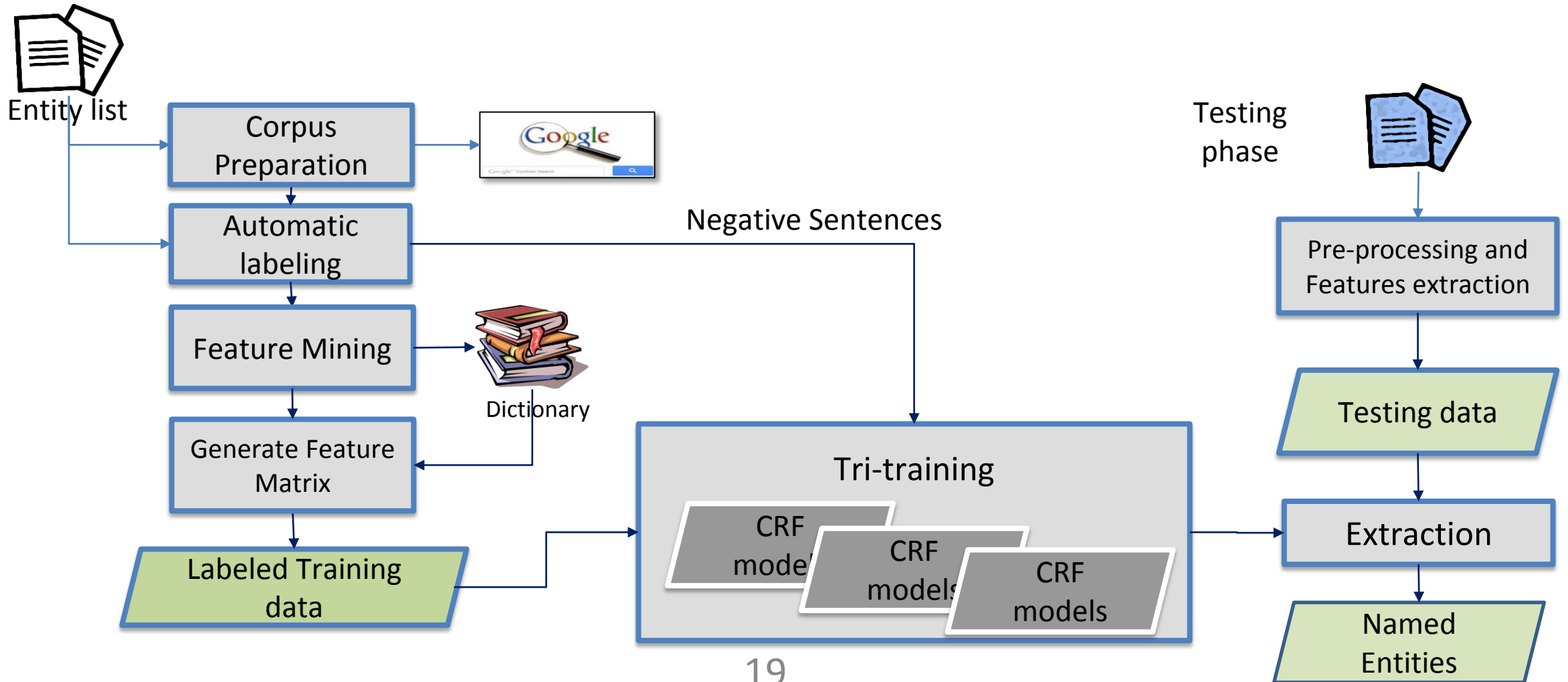
- Common Before, Common After, Entity Prefix, Entity Suffix

▣ Tri-Training with 3 sequence labeling models

- Making use of unlabeled data

<https://sites.google.com/site/nculab/projects/ds4ner>

DS4NER Model Generation



Challenges with Automatic Labeling

- Short seeds
 - Many false positive labeling
- Long seeds
 - False negative labeling with exact match
- Scalability issues
 - 500K seeds against 1M sentences
- Labeling strategy
 - Long seed first?
 - Rank by similarity?
- Avoid nested labeling

Song seeds

希望
背叛
飄

Examples:

【葉婉如／台北報導】李安說：「不<NE>希望</NE>兒子得獎」
今年由黃明志的<NE>飄</NE>向北方拿下冠軍
以<NE>背叛</NE>為主題的電影很多

Activity seed:

國父紀念館館藏展
台灣工藝之家府城聯展

Examples:

今年...五月甫於國父紀念館中山國家畫廊，
國寶級木雕大師吳榮賜曾榮獲「台灣工藝之家」之尊榮，

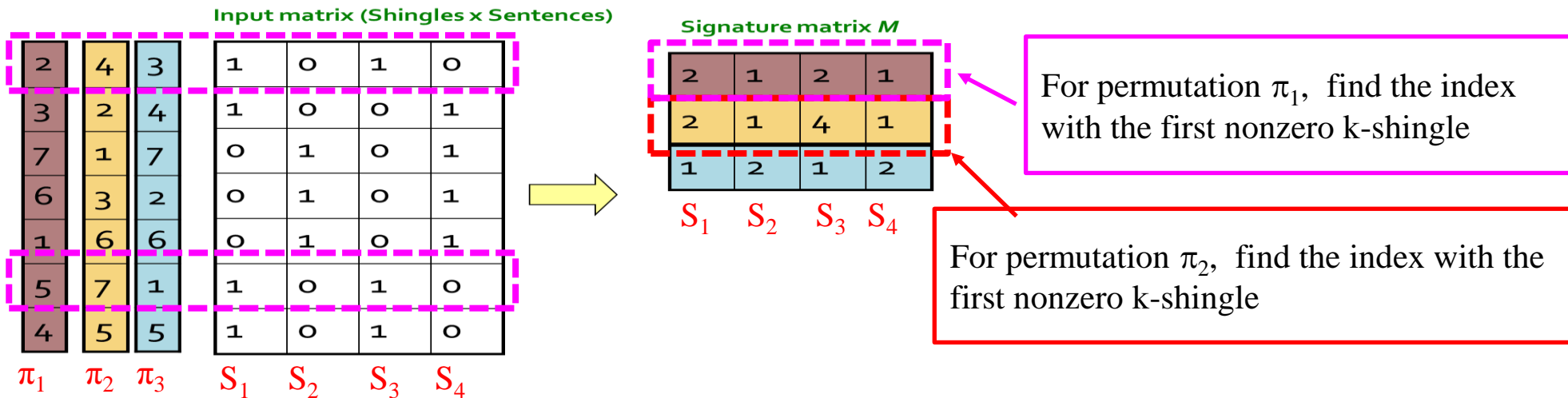
Automatic Labeling: Efficient Labeling with LSH

1) k-shingle

- Sentence $s = \text{abcbab}$, $k\text{-shingle}(s, k=2) = \{\text{ab}, \text{bc}, \text{ca}, \text{ab}\}$
- Transfer s to a long sparse vector format, $v = \{0\ 0\ 1\ 0\ 0\ 1\ 0\ 1\ \dots\}$

2) MinHashing

- Convert k -shingle vectors to short signatures vectors of length L , while preserving similarity.



Automatic Labeling: Efficient Labeling with LSH (Cont.)

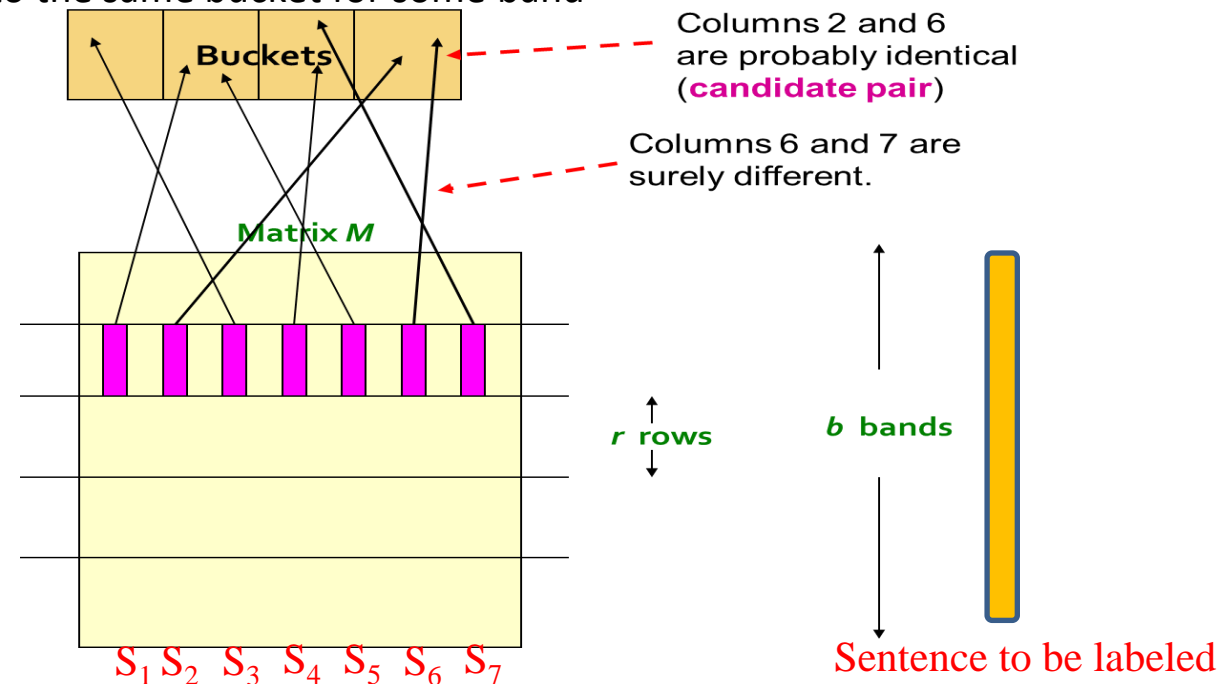
3) Locality-Sensitive Hashing

- Each column of signature matrix M is divided into b bands of r rows.
- For each band, hash the code to a table with k buckets.
- Each pair of seeds hashed into the same bucket represents a candidate pair.
 - If two sentences are duplicates, they will be hashed to the same bucket for some band

4) For each sentence to be labeled

Find candidate seeds from each band using sentence signature

Time 227 days (estimated with 4 threads) → 1d09h (4 threads)



Automatic Labeling Policy

- Build 3 LSH Based on the length of a seed
 - Short: $\text{Length} \leq 2 \rightarrow \text{ignore/exact labeling} + \text{guillemets (option)}$
 - Medium Seed: $3 \leq \text{Length} \leq 5 \rightarrow \text{exact labeling/exact labeling} + \text{guillemets (option)}$
 - Long: $\text{Length} > 5 \rightarrow \text{partial labeling with LCS (alignment)}$
- Labeling Policy 1
 - From long seed labels to short seeds
 - Label the current seed if its similarity $>$ threshold or contain core or last or first then
- Labeling Policy 2
 - Ranking by Similarity
 - Label current seed if the aligned part contain core or first or last of a seed

Automatic Labeling: Policy

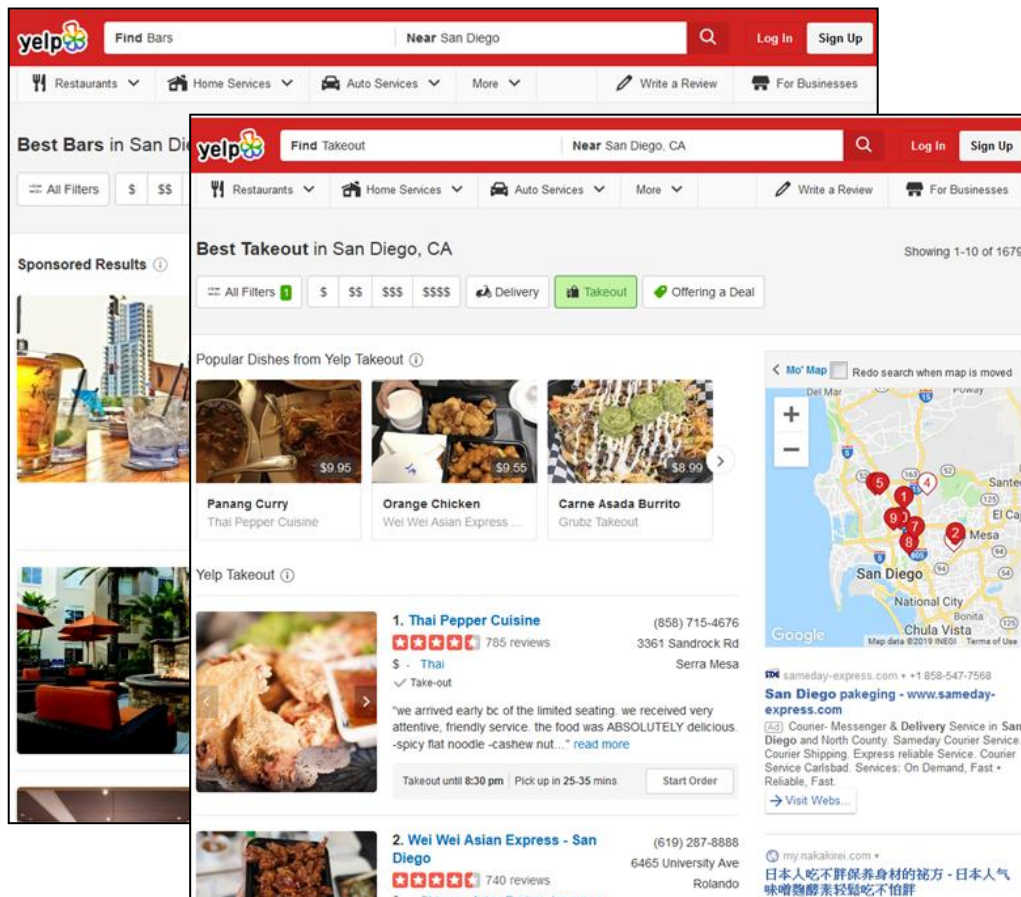
- 1) Replace matched seed with tags
- 1) Early stop the internal for loop
- 2) Filter out negative examples (option)
- 3) Supports multi-threading technology
- 4) Supports LSH to improve performance

How to Obtain Those Seed Entities?

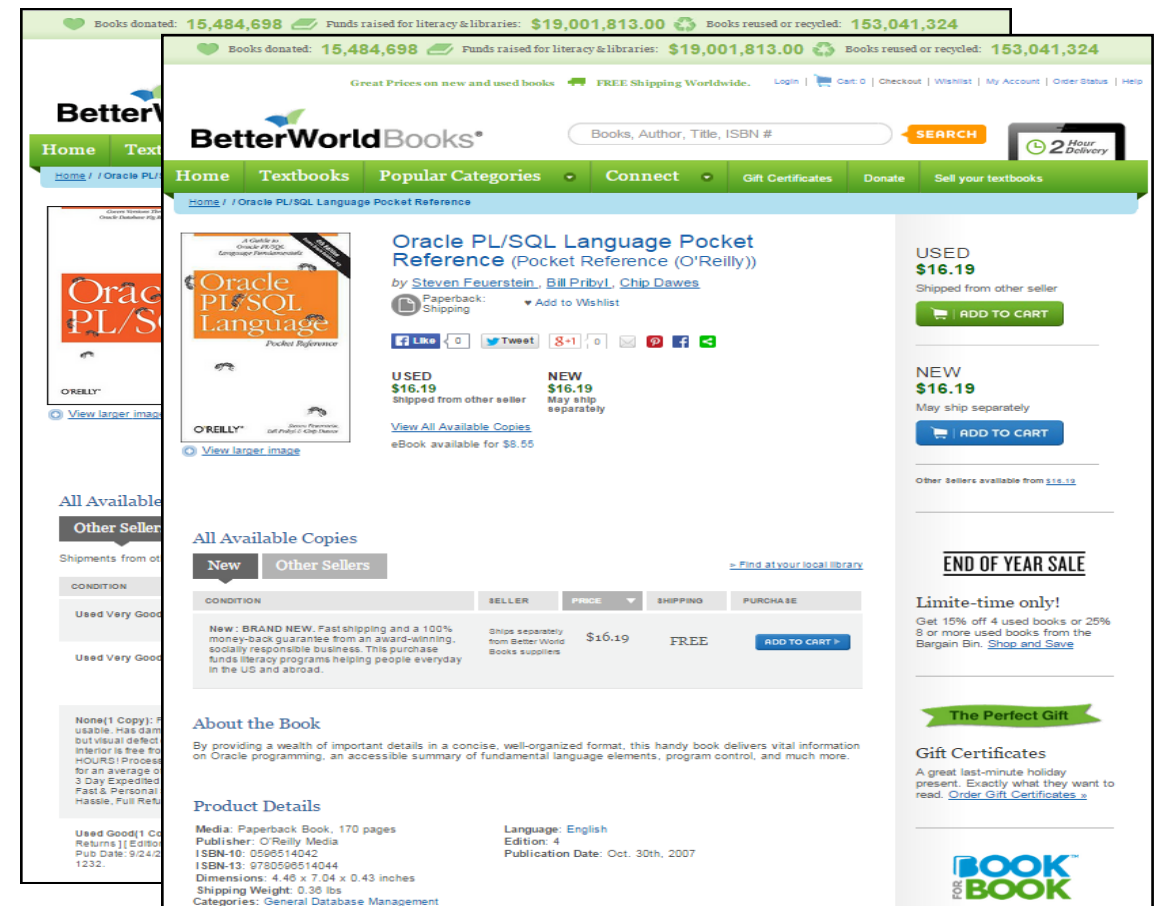
Web Data Extraction

List vs. Singleton Pages

- Most researches focus on data extraction from **list-pages**.
- The performance is evaluated on the **selected data items**.
- Output: Record-level vs. Page-level (full) schema

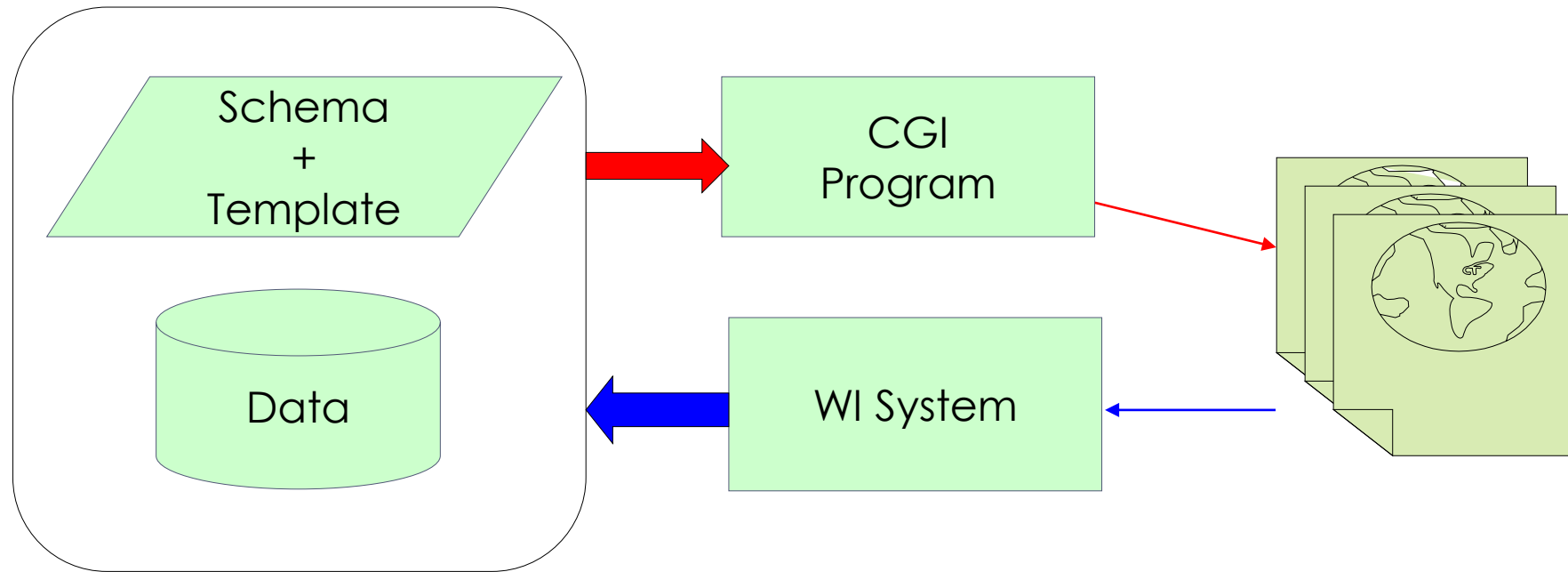


List Pages



SingletonPages

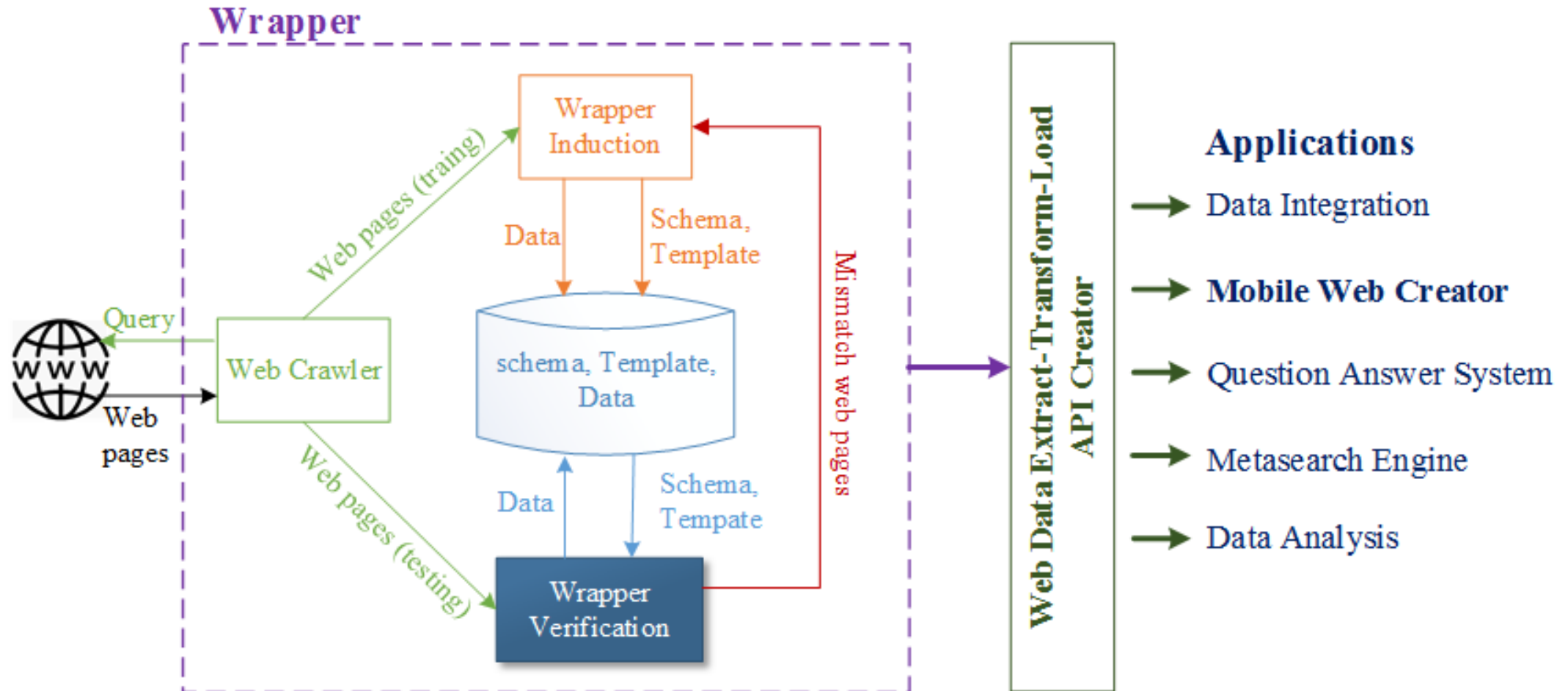
Core Technique: Deep Web Wrapper Induction – Reverse Engineering



- ⇒ CGI programs generate HTML pages based on schema+template and data.
- ⇐ WI systems infer the schema+template and data from Html pages.

DeXaR Project

Data Extraction and Reuse



APPLICATIONS OF DEEP WEB DATA EXTRACTION

- p [Web ETL API Creator](#)
- p [Mobile Web Creator](#)



Extract-Transform-Load Web data from

Create a Mobile Web site from existing web servers (reuse of existing system)

Web Data ETL API Creator

Goal: A tool for creating data API endpoint

Features:

- **Web data crawling and extraction without programming**
- **Two Backend Wrapper Induction Methods :**
 - **Single list page extraction**
 - **Multiple page extraction: DCA [Applied Intelligence 2018]**

How to create a Data ETL API?

1. **Create extractor: Input url pattern and parameters to generate url list.**
2. **Setup extractor: Enable/Disable/Merge columns**
3. **Checkout API endpoint result**

Create Extractor

Extractor name NCU News

Page type
☒ List page
☐ Detail page

Url pattern `https://www.ncu.edu.tw/campus/news/${page}`

Parameters page number range 1 ~ 7

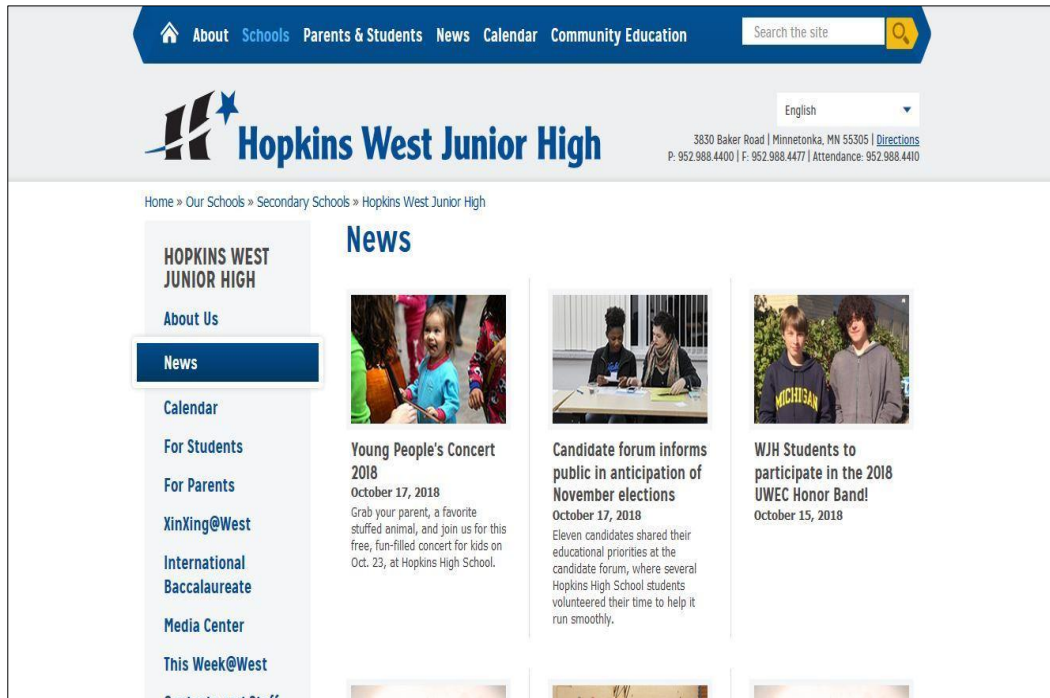
Result urls (max: 30)

1. `https://www.ncu.edu.tw/campus/news/1`
2. `https://www.ncu.edu.tw/campus/news/2`
3. `https://www.ncu.edu.tw/campus/news/3`
4. `https://www.ncu.edu.tw/campus/news/4`
5. `https://www.ncu.edu.tw/campus/news/5`
6. `https://www.ncu.edu.tw/campus/news/6`
7. `https://www.ncu.edu.tw/campus/news/7`

SUBMIT

Web ETL API Creator

<http://140.115.54.44:8001/>



HWJH website
recent news page

Web ETL System

Demo Extractors

My Extractors

Create Extractor

L

Hopkins News

API Endpoint : <http://140.115.54.56:8001/api/output/endpoint/3cvrjw33njnx9gag>

DATA SETS

PAGE SOURCES






OUTPUT & EXPORT

NON-SET

SET 1 ✓

SET 2


☐ Show disabled columns


ID	image	href	title
1	 https://www.hopkinsschools.org/sites/default/fi...	https://www.hopkinsschools.org/district-news/news/y...	Young People's Concert 2018
2	 https://www.hopkinsschools.org/sites/default/fi...	https://www.hopkinsschools.org/district-news/news/c...	Candidate forum informs public in anticipation of Nov...
3	 https://www.hopkinsschools.org/sites/default/fi...	https://www.hopkinsschools.org/schools/hopkins-wes...	WJH Students to participate in the 2018 UWEC Honor ...
4	 https://www.hopkinsschools.org/sites/default/fi...	https://www.hopkinsschools.org/schools/hopkins-wes...	This Week @ West_Episode 5_18/19
5	 https://www.hopkinsschools.org/sites/default/fi...	https://www.hopkinsschools.org/schools/hopkins-wes...	WJH 2018 Fall Parent / Teacher Conferences


Structure (Table-like) output



Web Data ETL API Creator

2. Setup extractor: Enable or disable the selected columns.

 ENABLE

 DISABLE

 MERGE

ID	<input type="checkbox"/> Col1	<input type="checkbox"/> title 	<input type="checkbox"/> link 
1	2018-05-03	107全大運圓滿落幕 5項9人次破全國紀錄	https://www.ncu.edu.tw/campus/article/2122
2	2018-05-03	107全大運柔道 土木系張立坤第四量級摘銀	https://www.ncu.edu.tw/campus/article/2123
3	2018-05-03	楊俊瀚一掃賽前陰霾 刷新200公尺全國紀錄	https://www.ncu.edu.tw/campus/article/2124
4	2018-05-03	臺灣體大王星皓400公尺混合式 破全國紀錄	https://www.ncu.edu.tw/campus/article/2125
5	2018-05-03	向俊賢跳高2.21破大會紀錄 完成六連霸	https://www.ncu.edu.tw/campus/article/2126
6	2018-05-02	撞球9號球個人賽 許睿安、古正晴稱王封后	https://www.ncu.edu.tw/campus/article/2119
7	2018-05-02	網球金牌戰 吳東霖封王 李亞軒傷復奪金	https://www.ncu.edu.tw/campus/article/2120
8	2018-05-02	全台最會超越障礙的女生 8年來摘第5金	https://www.ncu.edu.tw/campus/article/2121
9	2018-05-01	男子110M跨欄 陳奎儒破全國、大會紀錄	https://www.ncu.edu.tw/campus/article/2115
10	2018-05-01	楊俊瀚一日雙金 綿百米、4x100米三連霸	https://www.ncu.edu.tw/campus/article/2116




Edit column name

2. Setup extractor : Rename the column name



<div><div><div><div><div></div><div>Title</div></div><div><div></div><div></div></div></div></div></div>	
中大首創「WATCH全大運」	天氣資訊一把罩
107全大運圓滿落幕	5項9人次破全國紀錄
107全大運柔道	土木系張立坤第四量級摘銀
楊俊瀚一掃賽前陰霾	刷新200公尺全國紀錄

Web Data ETL API Creator

2. Setup extractor : Merge multiple columns into one column.

<input type="checkbox"/> Col5 	<input type="checkbox"/> Col6 	<input type="checkbox"/> Col7 
2019/12/31 23:00(+0800)		
2019/05/18 09:30(+0800)	~16:00	
2019/04/13 09:30(+0800)	~	2019/04/20 16:00(+0800)
2019/03/16 09:30(+0800)	~	2019/03/23 16:00(+0800)
2019/02/16 09:30(+0800)	~	2019/02/23 16:00(+0800)
2019/01/19 09:30(+0800)	~	2019/01/26 16:00(+0800)
2018/12/31 17:40(+0800)		
2018/12/15 09:30(+0800)	~16:00	

Merge

<input type="checkbox"/> Col5+Col6+Col7  
2019/12/31 23:00(+0800)
2019/05/18 09:30(+0800)~16:00
2019/04/13 09:30(+0800)~2019/04/20 16:00(+0800)
2019/03/16 09:30(+0800)~2019/03/23 16:00(+0800)
2019/02/16 09:30(+0800)~2019/02/23 16:00(+0800)
2019/01/19 09:30(+0800)~2019/01/26 16:00(+0800)
2018/12/31 17:40(+0800)
2018/12/15 09:30(+0800)~16:00

Web Data ETL API Creator

3. Checkout API endpoint result

```
→ ↻ 🏠 ⓘ 140.115.54.45/api/endpoint/nyi1rx0jgvgj3hmu

{
  "pagination": {
    "page": 1,
    "itemsPerPage": 36,
    "totalPage": 1
  },
  "data": [
    {
      "date": "2018-05-03",
      "title": "107全大運圓滿落幕 5項9人次破全國紀錄",
      "url": "http://www.ncu.edu.tw/campus/article/2122",
      "image": "http://www.ncu.edu.tw/assets/thumbs/news/9860517b45f561",
      "description": "107年全國大專校院運動會閉幕典禮2日在中央大學依仁堂舉行，才"
    },
    {
      "date": "2018-05-03",
      "title": "107全大運柔道 土木系張立坤第四量級摘銀",
      "url": "http://www.ncu.edu.tw/campus/article/2123",
      "image": "http://www.ncu.edu.tw/assets/thumbs/news/ce5426f45822b1",
      "description": "107全大運柔道競賽，地主隊中央大學柔道隊選手突破創隊以來競"
    },
    {
      "date": "2018-05-03",
      "title": "楊俊瀚一掃賽前陰霾 刷新200公尺全國紀錄",
      "url": "http://www.ncu.edu.tw/campus/article/2124",
      "image": "http://www.ncu.edu.tw/assets/thumbs/news/6090329951664!",
      "description": "擁有「全台最速男」封號的楊俊瀚，於5月2日代表台灣體大征戰男"
    }
  ]
}
```

Web Data ETL API Creator

3. Output API endpoint result

DATA SETSPAGE SOURCESCRAWLER SETUPOUTPUT & EXPORTDANGER ZONE

Live API Endpoint

Endpoint Url :

VIEWCOPY URL

Pagination :

You can use query `page` and `itemsPerPage` in endpoint url for pagination.

Example : `http://etl-api-creator.com/api/output/endpoint/nyi1rr3gjin94jew?page=1&itemsPerPage=3`

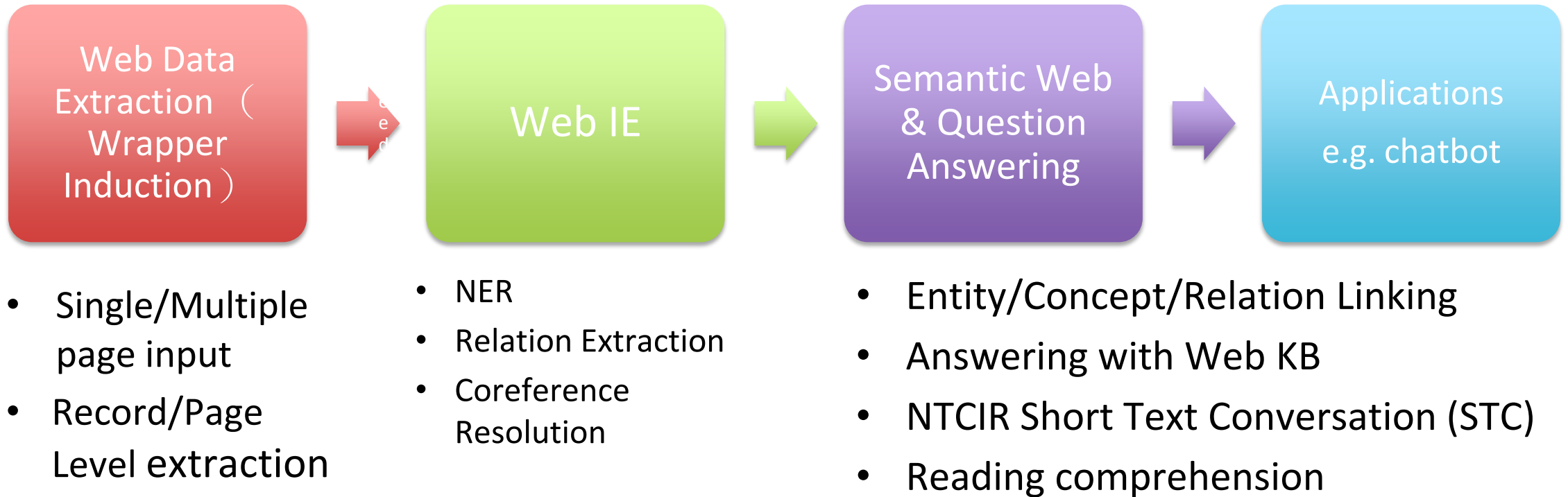
Export Static File

EXPORT CSVEXPORT JSONEXPORT XML

etl-api-creator.com/api/output/endpoint/nyi1rr3gjin94jew

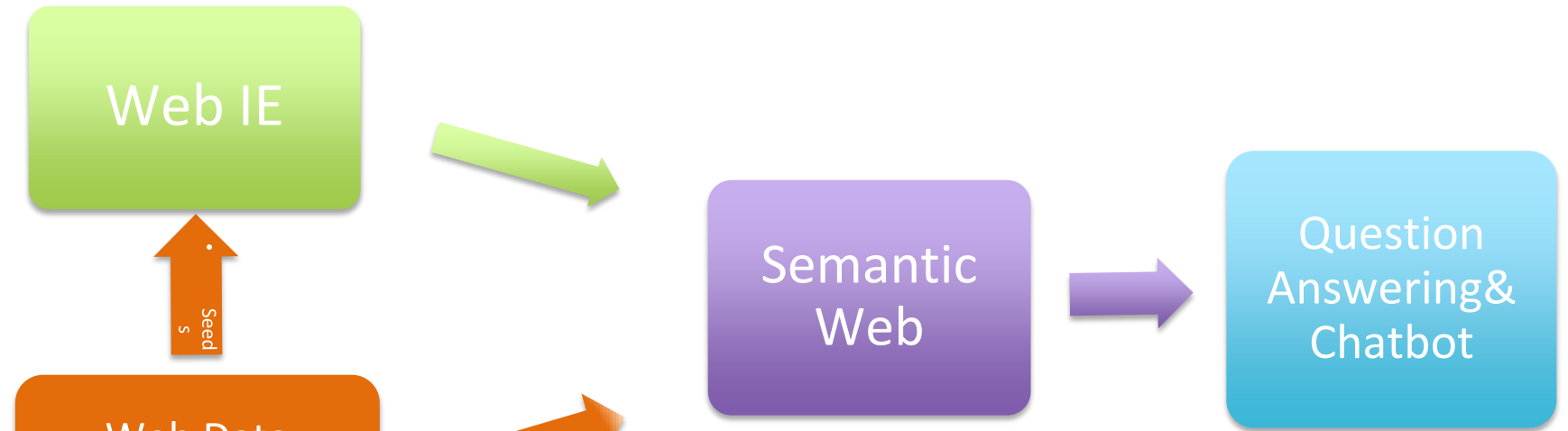
```
{
  "pagination": {
    "page": 1,
    "itemsPerPage": 60,
    "totalPage": 1,
    "prevPageUrl": null,
    "nextPageUrl": null
  },
  "data": [
    {
      "date": "2018-06-20",
      "title": "科技部傑出研究獎公布 中大學術能量豐沛",
      "url": "https://www.ncu.edu.tw/campus/article/2149",
      "image": "https://www.ncu.edu.tw/assets/thumbs/news/ff30d3714f5",
      "description": "科技部106年度「傑出研究獎」為獎勵研究成果傑出的科學技術、質....."
    },
    {
      "date": "2018-06-19",
      "title": "人生公式的解答 統計所江村剛志：擇你所愛",
      "url": "https://www.ncu.edu.tw/campus/article/2147",
      "image": "https://www.ncu.edu.tw/assets/thumbs/news/1e1600b1bec",
      "description": "科技部為獎助及鼓勵青年研究人員之學術發展設立了吳大猷先生過....."
    },
    {
      "date": "2018-06-13",
      "title": "喚起「善」的種子 中大人端午粽香傳愛",
      "url": "https://www.ncu.edu.tw/campus/article/2146",
      "image": "https://www.ncu.edu.tw/assets/thumbs/news/c9694f56b19",
      "description": "端午節即將來到，為關懷弱勢族群，培養學生人文關懷精神，喚到....."
    }
  ]
}
```


Roadmap for Message Understanding and Information Management



Co-Learning with Computers via Distant Supervision

- NER,
- Coreference Resolution
- Relation Extraction
- Single/Multiple page input
- Record/Page Level extraction



- Entity/Concept/Relation Linking
- Answering with Web KB
- NTCIR Short Text Conversation
- Taiwan AI Grand Challenge

Co-Learning with Computers via Distant Supervision

Conclusions

- Deep Web Data Extraction
 - Web data reuse and integration
 - Enhance Web KB for better question answering
- NER & Relation Extraction
 - Question answering
 - Damage report monitoring, Activity search, POI search, etc.
 - Call center client service