# A Case Study on
# Linked Data Generation and Consumption

Jianqiang Li, Yu Zhao
NEC Labs China
{lijianqiang, zhaoyu}@research.nec.com.cn

# Overview

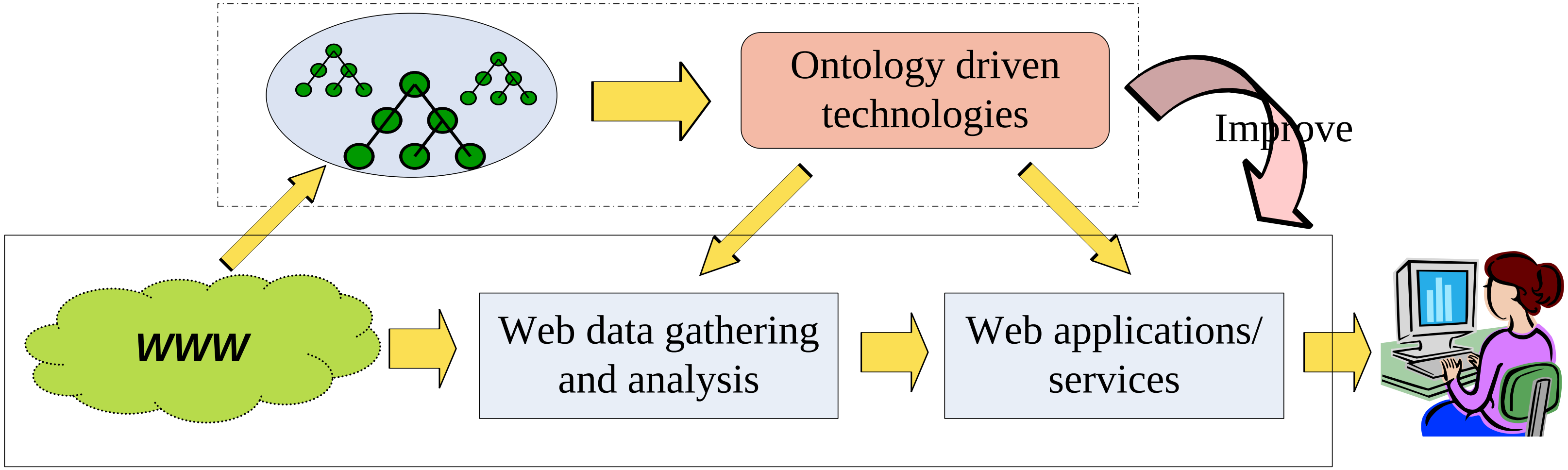
- Motivation and goal
- Our experimental study
  - – Linked data generation
  - – Consuming the linked data for web search improvement
- Conclusion and future work

# Motivation

- The existence of large amounts of interlinked semantic data is a prerequisite for making the Semantic Web come true.
  - Current linked data construction relies heavily on the already existing (structured) data sources and the efforts made by the data publishers.
- The Web provides an unprecedented opportunity and fertile ground for knowledge discovery
  - Our goal is to extract the inherent statements implied in the hyperlinks as a form of semantic data and make the data available to be consumed by various Semantic Web applications

# Our Experimental Work

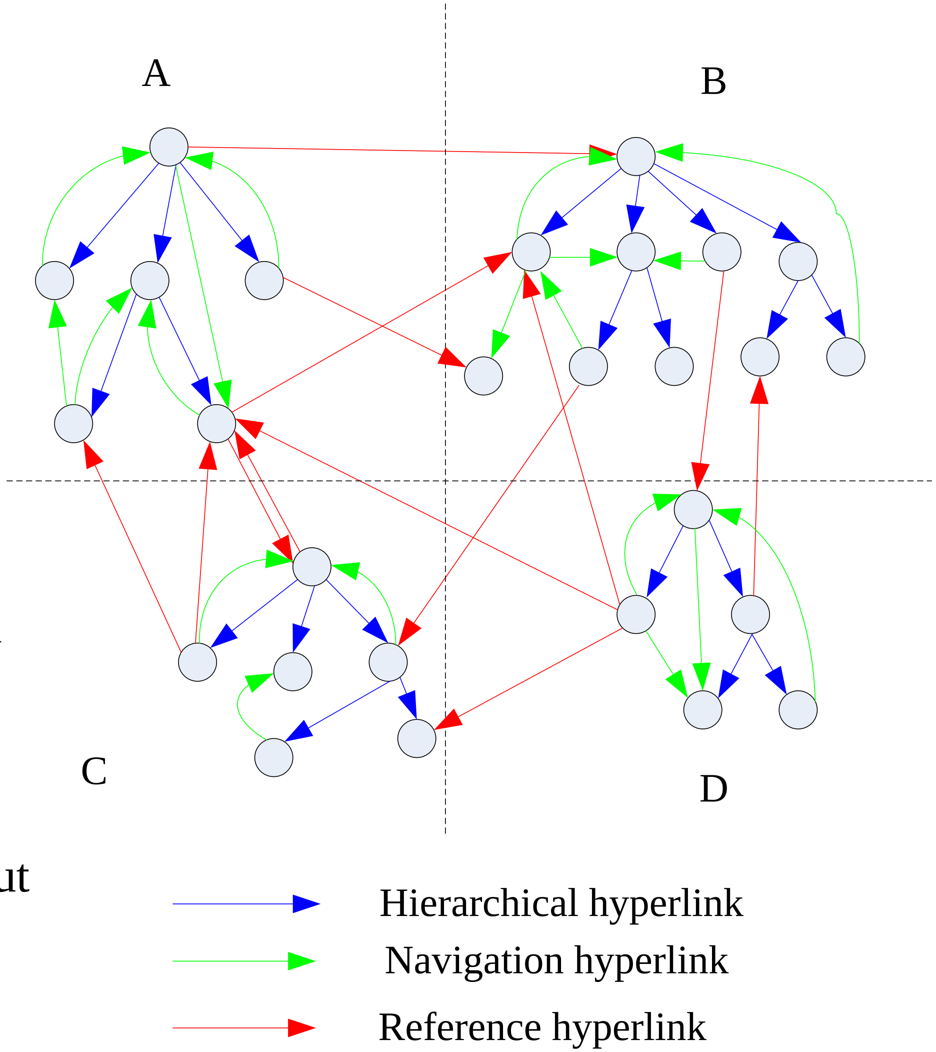
- The case study includes two parts:
  - Semantic data construction
    - Extracting (shallow) semantic data about the interlinked web documents as a new source of linked data
  - Linked data consumption for web search improvement
    - The semantic data provide important indications on the web page content
    - The inference is incorporated implicitly into the web page retrieval process


Ontology driven technologies
Improve
WWW
Web data gathering and analysis
Web applications/ services

# Linked Data Generation (1)

## - Where to find the semantic data

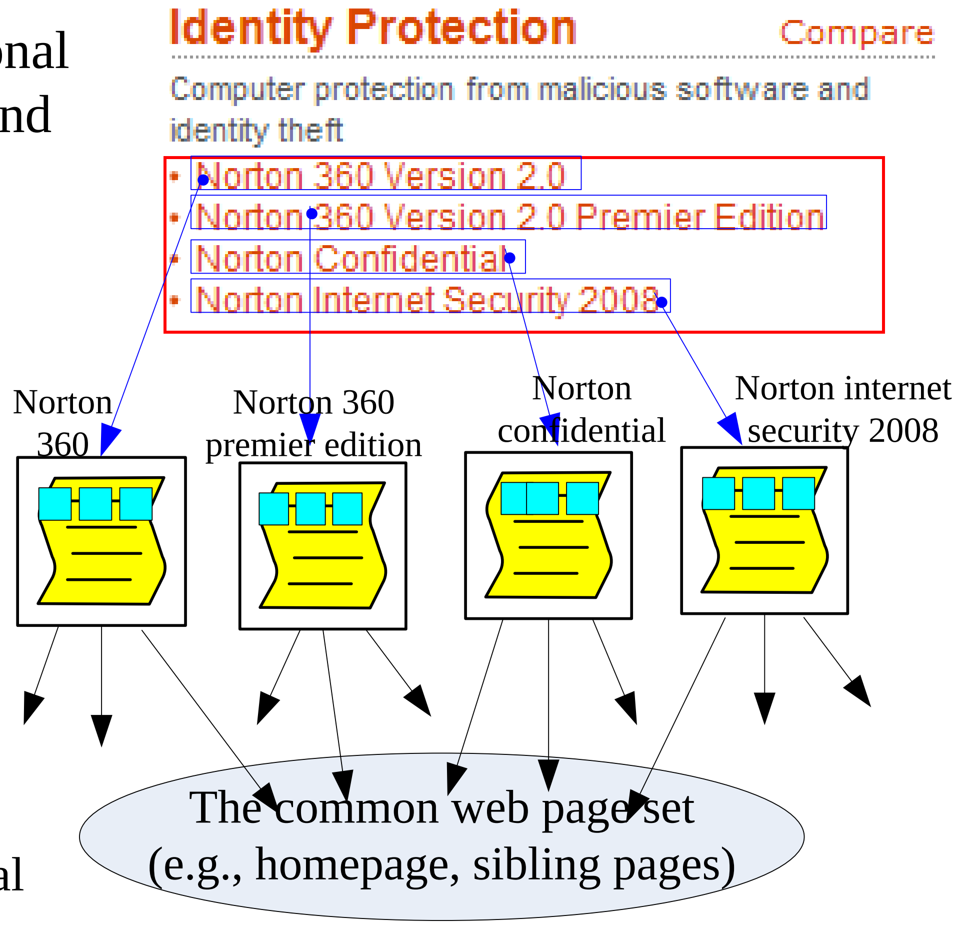
- Hyperlink differentiation
  - Hierarchical hyperlink (intra-site)
    - It exists largely in the local website, are mainly used for organizing the collection of web pages
    - It is used for building the local topic hierarchy
  - Reference hyperlink (inter-site)
    - It represents citations and are implicitly utilized by the web page author for web page recommendation
    - It reflects the inter-linkage relation between multiple topic hierarchy
  - Pure navigation hyperlink (intra-site)
    - Its major role is to provide the shortcut to facilitate the readers to jump from one page to another page.
    - Noise information



# Linked Data Generation (2)

## -How to extract the semantic data

- Hierarchical relation identification
  - Its goal is to remove the pure navigational hyperlinks (the direct/indirect sibling and upward hyperlinks) from the intra-site hyperlink collection
- The method includes two steps :
  - Syntactical URL analysis:
    - Utilizing the information implied in *http://[host]/[path]/[file]#[fragment]*;
  - Semantic hyperlink analysis:
    - Some heuristics are adopted, the core is shown in the schematic diagram: the hyperlinks pointing to the common web page set is identified as pure navigational links (noise information)



Identity Protection Compare

Computer protection from malicious software and identity theft

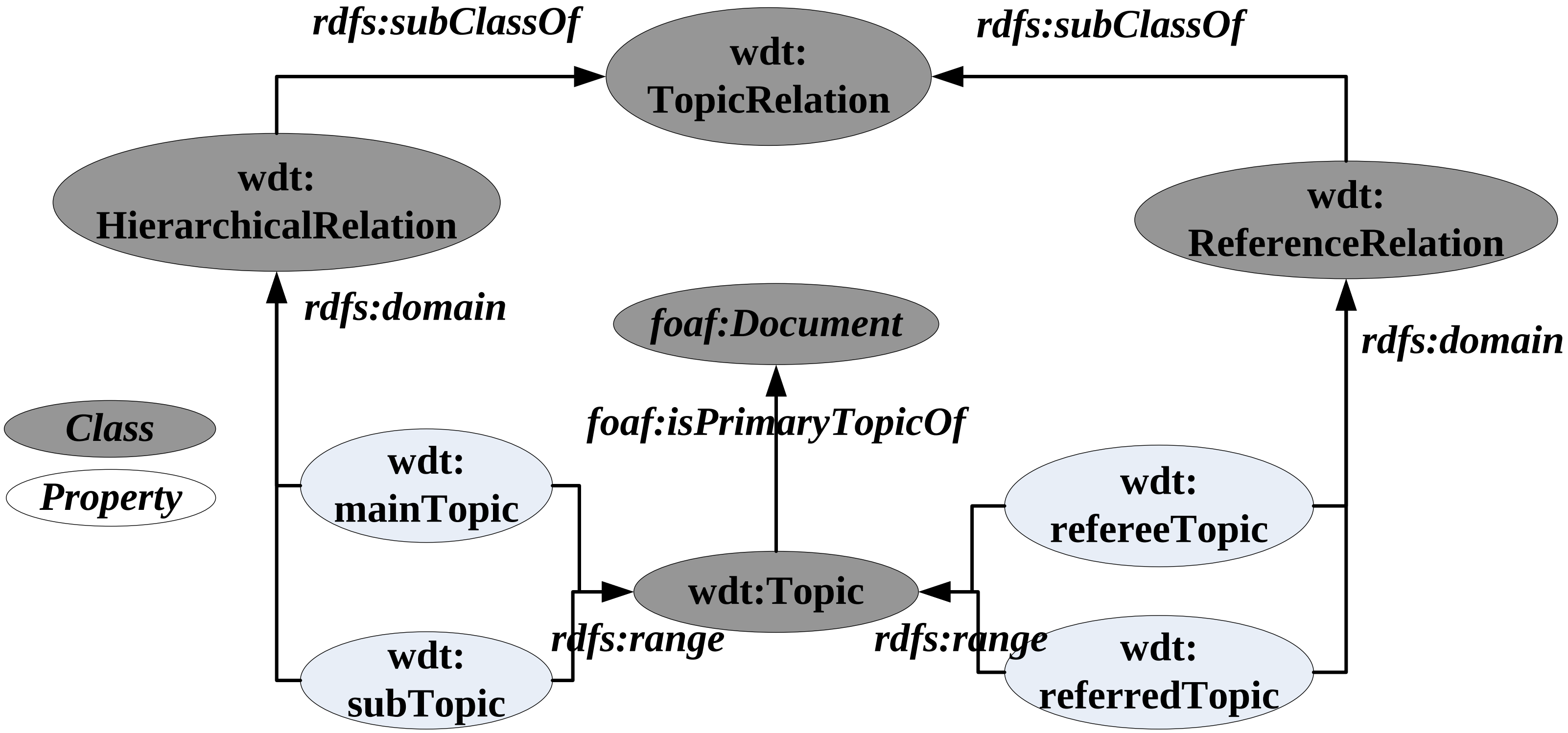
- Norton 360 Version 2.0
- Norton 360 Version 2.0 Premier Edition
- Norton Confidential
- Norton Internet Security 2008

Norton 360
Norton 360 premier edition
Norton confidential
Norton internet security 2008
The common web page set
(e.g., homepage, sibling pages)

# Linked Data Generation (3)

## - How to publish the linked data

- The WDT vocabularies for the semantic data representation



- The semantic data (**hierarchical relation** between web pages **)** regarding to **the website** is specified by the WDT framework, and the various datasets are **inter-linked** with **reference relations**. Such data is also connected to document web.

# Linked Data Generation (4)

## - Example of the resultant linked data

- A segment of the topic hierarchy of *stanford.edu*

```
# Topic "Protégé"
<http://www.nec.com.cn/lab/WDT/data/stanford.edu#34211>
   rdf:label "The Protégé Ontology Editor and Knowledge Acquisition System" ;
rdf:type wdt:Topic ;
foaf:isPrimaryTopicOf <http://protege.stanford.edu> .
# Topic "Overview of Protégé"
<http://www.nec.com.cn/lab/WDT/data/stanford.edu#34212>
   rdf:label "What is Protégé?" ;
rdf:type wdt:Topic ;
foaf:isPrimaryTopicOf <http://protege.stanford.edu/overview/> .
# Hierarchical relation between above two topics
<http://www.nec.com.cn/lab/WDT/data/stanford.edu#34302>
   rdf:label "OVERVIEW" ;
rdf:type wdt:HierarchicalRelation ;
wdt:mainTopic < http://www.nec.com.cn/lab/WDT/data/stanford.edu#34211> ;
wdt:subTopic < http://www.nec.com.cn/lab/WDT/data/stanford.edu#34212> .
```

# Linked Data Generation (5)

## - Example of the resultant linked data

- **An example of a reference relation:**

```
# Reference relation between protégé and OWL
<http://www.nec.com.cn/lab/WDT/data/stanford.edu#34311>
   rdf:label "OWL Ontology Web Language Guide" ;
rdf:type wdt:ReferenceRelation ;
wdt:refereeTopic < http://www.nec.com.cn/lab/WDT/data/stanford.edu#34212> ;
wdt:referredTopic < http://www.nec.com.cn/lab/WDT/data/w3.org#1421> .
```

- **Link from data to document:**

```
<rdfs:isDefinedBy rdf:resource=”http://www.w3.org/TR/2004/REC-owl-semantics-20040210/” />
```

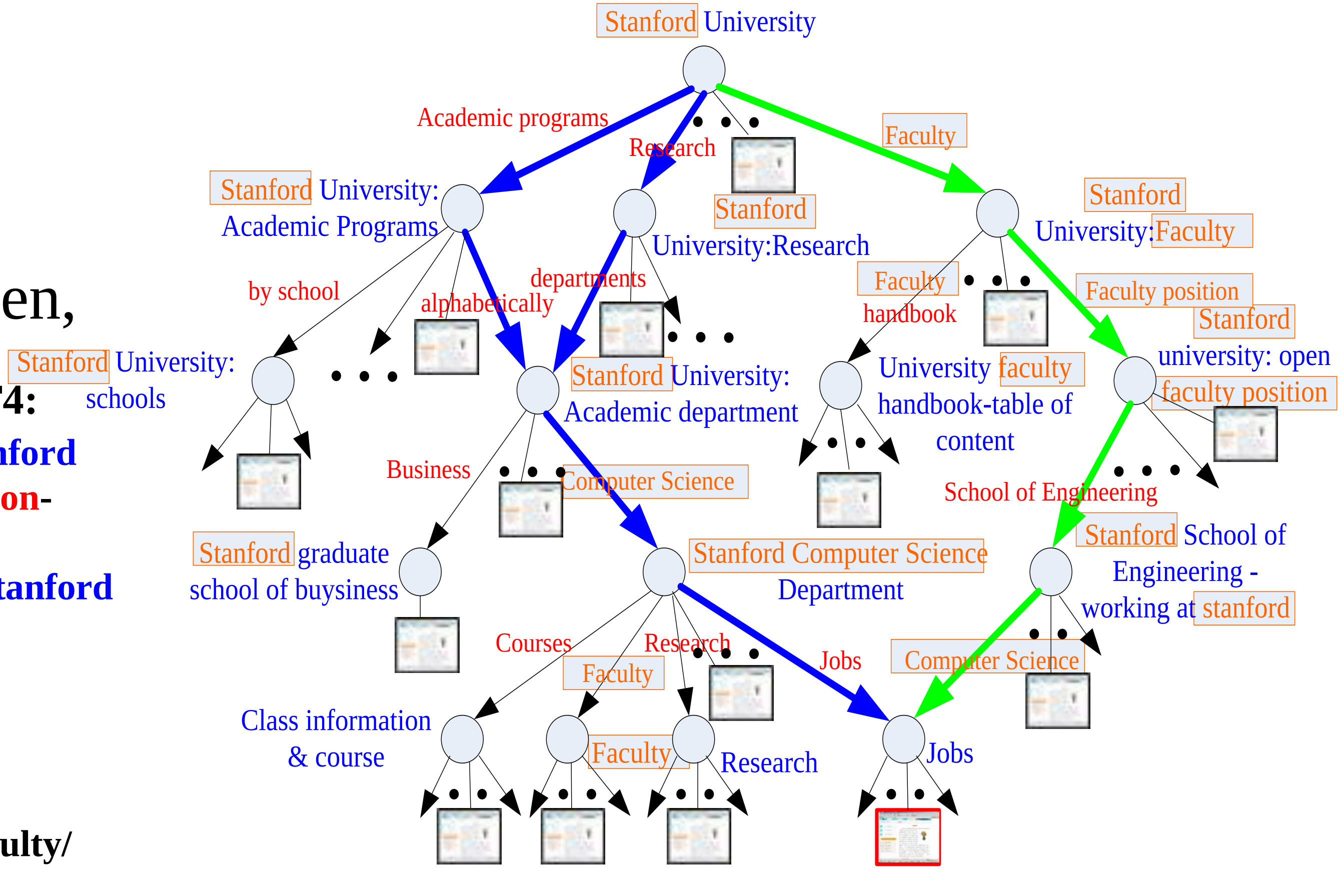
# Linked Data Consumption (1)

## -Building a new resource from the generated linked data

- Hierarchical Navigation Path (HNP): HNP=<TL, UL, C>

- An example:

navigation path in green,


Stanford University
Academic programs
Research
Faculty
Stanford University: Academic Programs
Stanford University:Research
Stanford University:Faculty
by school
alphabetically
departments
Faculty handbook
Faculty position
Stanford university: open faculty position
Stanford University: schools
Stanford University: Academic department
University faculty handbook-table of content
Business
Computer Science
School of Engineering
Stanford graduate school of buysiness
Stanford Computer Science Department
Stanford School of Engineering - working at stanford
Courses
Research
Faculty
Jobs
Computer Science
Class information & course
Faculty
Research
Jobs

**TL=T1+A1+T2+A2+T3+A3+T4:**

**Stanford University->faculty->Stanford University: Faculty->Faculty position->Stanford University: open faculty position->school of engineering->Stanford School of Engineering: working at stanford->computer science->Jobs**

**UL=U1+U2+U3+U4:**

**http://www.stanford.edu/**

**-http://www.stanford.edu/home/faculty/**

**-http://www.stanford.edu/home/faculty/positions.html**

**-http://soe.stanford.edu/about/jobs.html-http://cs.stanford.edu/Info/jobs.php**

**C=Domain/host_Name: Stanford**

# Linked Data Consumption (2)

## - Exploiting the HNP for web page ranking

- A three-step-procedure to realize the query-path match for Web page ranking:
  - Using link structure analysis of the Web to estimate the rank value *RW* for each website *W* at global level, i.e., the relative importance of *W*;
  - Computing the rank value *Rpath* for each HNP *path* according to its located web site and the query;
  - The pathrank value *Rpage* of a web page *page* is determined by all its corresponding HNPs (or together with the page’s content-based score).

# Linked Data Consumption (3)

## - Evaluation

- The experiments are conducted on 30+ company websites and *stanford.edu*
- For hierarchical relation identification, roughly 80%+ is correct; For the HNP, the recall rate is 90%+ and the precision is 70-80%.
- For webpage retrieval (the website search engine in *stanford.edu* as the baseline ):

| | **S@5** | **S@50** | **P@10** | **P@20** | **SP** |
|---|---|---|---|---|---|
| ***stanford.edu* search** | 64% | 74% | 82% | 79% | 73% |
| **PathRank1** | 78% | 86% | 75% | 69% | 77% |
| **PathRank1+content** | 76% | 90% | 81% | 72% | 78% |
| **PathRank2** | 85% | 89% | 88% | 71% | 81% |
| **PathRank2+content** | 88% | 92% | 86% | 77% | 87% |

- The results show that through exploiting the (shallow) semantic data, our path-based approach can improve the accuracy of web page retrieval significantly

# Conclusion and Future Work

- A method for constructing the (shallow) semantic data from the Web is proposed
  - An alternative view to make a contribution to the vision of Web of Data
- The experiment on consuming the resulting linked data to enhance web page retrieval is studied
  - Since the inference is incorporated inside implicitly, the results is improved promisingly.
- Future work will focus more on refining the (shallow) semantic data and their consumption, e.g.,:
  - Search result organization
  - Object mining from the Web
  - Hierarchy learning from the Web
  - …

Empowered by Innovation

NEC