Knowledge Extraction from the Web and Trust-oriented Web Search

Hiroaki Ohshima
Assistant Professor
Kyoto University, Japan

2009/10/31@浙江大学
Application to discover rivals of a given term

- DEMO (UC Berkeley)
- DEMO (Google)
- DEMO (Hillary Clinton)
Knowledge Extraction

- **Ontological Knowledge**
  - Hypernyms, Hyponyms, Synonyms, Antonyms, …
  - Coordinate terms: terms which have the same hypernym
  - Topic terms, Description terms, Nickname, … (vaguely defined)

![Diagram](attachment:image.png)
Where is Knowledge Extracted from?

- **Large Text Corpora**
  - 10 years news archive
  - 1,000,000 HTML documents
  
    - good precision, many related terms, but very slow
    - static

- **Web Search Results**
  - Possible to collect only required sentences
  
    - fast, some related terms, good precision
    - dynamic, large scale, updated
Knowledge Extraction Using Web Searches

- Using Web searches NOT for page search, but for discovering knowledge

queries using lexico-syntactic patterns

“Hillary or” “or Hillary”

Knowledge from the Web

obtain

Knowledge Extraction from Web Search Results
Coordinate Terms Discovery

A term is given

2 Queries for Web Search made by connecting “OR”

Get 100 Search Results for each query

Find terms connected to the given term with “OR”

Count the appearance before and after “OR”

Candidate terms scored

Google Search result
Page Title
snippet ........................................
........................................
Nutcracker or Swan Lake
8 times

Swan Lake

Swan Lake or

Google Search result
Page Title
snippet ........................................
........................................
Swan Lake or Nutcracker
4 times

Geometrical average
5.7

Nutcracker is a coordinate term of Swan Lake

about 6 coordinate terms
70~80% precisions
3~5 seconds
A term is given

2 lexico-syntactic patterns

Make 2 Web queries

Get Web search results
100 titles and snippets

Find terms that match the patterns

Count the appearance for each pattern

Found terms scored

Pre pattern

such as Toyota

Post pattern

Toyota are

2.4

Geometrical average

3 times

2 times

car manufactures

is a hypernym of Toyota

Toyota

Get Web search results
100 titles and snippets

Find terms that match the patterns

Count the appearance for each pattern

Found terms scored

Pre pattern

such as Toyota

Post pattern

Toyota are

2.4

Geometrical average

3 times

2 times

car manufactures

is a hypernym of Toyota

Toyota
Web Information Credibility Analysis

The Cloud (XaaS)

- To help users to judge the credibility of Web text information
- To show some additional information with original contents
  - Trustworthiness
    - Goodness or morality of the source
  - Expertise
    - Knowledge, skill, and experience of the source (B. J. Fogg: Persuasive Technology)
- Target contents
  - Web search results
  - Q&A contents (Yahoo! Answers)
Analysis of Content itself and Analysis of Support for Content

Additional information to evaluate text credibility is obtained in real time

**Fairness of the Content**
Features of the content itself.
- **Topic Coverage of the Content**
  - How much the content contains typical topics.
- **Topic Majority of the Content**
  - How many other contents are similar to the content.
- **Speciality of the Content**
  - How much the content contains specialized information.

**Social Support for the Content**
Evaluation from the outside of the content.
- **Social Bookmark for the Content**
  - How many the content is bookmarked.
- **Geographical Social Support for the Content**
  - How widely people support (link to) the content locate.
- **The Number of Links to the Content**
### System Overview

<table>
<thead>
<tr>
<th>Trust Text</th>
<th>Search</th>
<th>WebSearchResult</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Amazon.com: Online shopping for ...</strong></td>
<td>Online shopping from the earth's biggest selection of books, ...</td>
<td>1 - 4,380,000</td>
</tr>
<tr>
<td><strong>Amazon Web Services</strong></td>
<td>Amazon Web Services (AWS) delivers a set of services that together form a reliable, scalable, and inexpensive computing platform</td>
<td>2 - 18,700</td>
</tr>
<tr>
<td><strong>Amazon.co.uk: low prices in ...</strong></td>
<td>Browse our selection of Ugg footwear from featured sellers at Amazon.co.uk</td>
<td>3 - 4,360,000</td>
</tr>
<tr>
<td><strong>Amazon.ca: Online shopping for ...</strong></td>
<td>Amazon.ca: Online shopping for Canadians - books, electronics, music, DVDs, ...</td>
<td>4 - 2,840,000</td>
</tr>
<tr>
<td><strong>Amazon Elastic Compute Cloud (Amazon EC2)</strong></td>
<td>Amazon EC2's simple web service interface allows you to obtain and configure capacity with minimal friction</td>
<td>5 - 850</td>
</tr>
<tr>
<td>**Amazon.co.uk: **</td>
<td>Hello. Sign in to get personalised recommendations. New Customer?</td>
<td>6 - 10</td>
</tr>
<tr>
<td><strong>Amazon.com Books: New &amp; used books ...</strong></td>
<td>... of Amazon Prime™. Just add any eligible textbook to your cart to qualify. ...</td>
<td>7 - 477</td>
</tr>
</tbody>
</table>
Geographical Social Support

- A Web page is linked from many other pages
- A link can be regarded as a social support
An Example of Geographical Social Support

Blue Ridge
- Highly biased from the east coast

North Cascades
- Uniformly distributed
Uniformity of Support (US)

- **Uniformity**
  - If the page is supported by pages locating in many different places

→ Better!
High Credibility?
Proximity of Support (PS)

- Proximity
  - If the page is supported by proximate pages.

→ Better!
High Credibility?
Summary

- Knowledge Extraction from the Web
  - Bi-directional Lexico-syntactic Patterns
    - Pre Pattern: Target related terms will appear just before this.
    - Post Pattern: Target related terms will appear just after this.
  - Advantages
    - Fast: 200 search results are enough to obtain several related terms.
    - Finding proper cut point for a compound word

- Trust-oriented Web Search
  - To show additional information obtained by light-weight analysis of text
  - Fairness of the Content
    - Topic Coverage
    - Topic Majority
    - Speciality
  - Social Support for the Content
    - Social Bookmark
    - Geographical Social Support