Contact searching for business partners

PV115 Laboratory of Knowledge Discovery

Juraj Jurčo

Faculty of Informatics, Masaryk University

October 4, 2011

Outline

1. Motivation
2. My work
3. WePS
   - Definition
   - Problems
   - Business Partners
   - Example
   - Resources
4. Documents processing
   - Approaches
   - Main content
   - Named entities
5. Disambiguation
   - Methods
6. What is done

Why we want to search?

- Business reasons:
  - Sponsors
  - Partners
  - Looking for similar area of work
- Personal reasons:
  - Our contacts are not actual
  - We need contact for specific person

My master thesis

- In my master thesis I'm trying to find business partners for Faculty of Informatics of Masaryk University
- The scope of searching is for persons owning company or managers of company
- I will get list of graduated (but also students) of our faculty and program will try to find information about them
- If they are successful program suggests these persons as possible business partners for our faculty
- Output of the program will be list of found contact information about person

Informal definition of problem

We need to find information about concrete person on the Internet and these information are in different pages or documents.

- Usually we search throw search engine which returns us documents sorted by some criteria (e.g. PageRank)
- Search engine usually pre-process our query

More formal definition of WePS

Let q be a query for search engine with the name of the searched person. Let $S = \{S_1, S_2, ...S_m\}$ be a set of documents returned by search engine E. Let $PD = \{D_1, D_2, ..., D_n\}$ be a set of documents containing the name of searched person where $PD \subseteq S$ and $m \geq n$. We suppose every document D_i contains only information about one real person. Set PD contains k persons. Input for algorithm is query q and output is set of k clusters of documents D_i belonging to one real person.[1, Vu2007] [2, Yoshida2010]

Problems connected with WePS

- One name can refer to more people in the same time
- One page can contain more people
- Information on page can be out of date
- Pages are using JavaScript or Flash to display their content

Search for business partners

- Problem for searching for business partners is similar to problem for searching people on the web (WePS)
- From all found persons with given name we need to choose the most likely person
- We need to take into account also scope of the work
- Persons or companies which sponsored some event are more relevant

Problems connected with searching

- Page can contains only information about company which sponsored event and no information about searched person
  - We have to find information about company too
- Search engines does not index whole content of the Internet (`robots.txt`)
- Searched name can be mentioned on the other page of the company
  - cooperation of companies

Human vs. Computer

Example:
Ing. Zdeněk ŠTĚPÁNEK

Chief executive of ALIMA a.s.
Birth date: 22.9.1961
Place: Caslava
Birth number: 6109220040
Address: Zdiměřická 852,

25242 Jesenice

Project manager
Address: Jugoslavska 25,
Ostrava-Jih, 700 30
Mobil: 777 077 709
Email: zstepanek@zeu.cz

IC: 27846181

Chairman of Czech office
for guns and munition testing
Address: Jilmová 759/12,
130 00 Praha 3
Phone: 271773064
Web: www.cuzzs.cz

Email: stepanek@cuzzs.cz

Useful information resources

- Trade Register - Information about companies and business persons
- Managers moving between companies
- Social networks
- Job portals
- Companies homepage

Used approaches

- Preprocessing of documents is very important
  - Has impact for accuracy
  - Documents have to be pre-processed "on-the-fly"
- At the beginning the number of persons is not known
- Clustering algorithms with fixed number of classes are unusable

Main content extraction

- Web pages contain also irrelevant information
  - Header and footer contain other contact information
  - Menus and advertisements
- Methods used for main content extraction
  - Not extract, transform to plain text [2, Yoshida2010]
    - then we need to use other methods for determine main content of document
  - Extract some few characters (200, 500, 1000) around found keyword[4, Jiang2010]
    - Some named entities occur near the searched name.
  - Extraction based on similar pages on the server
    - Navigation elements referring to pages on the server
    - Time and data transfer consuming job, but maybe more accurate

Named Entities I

- 8 important Named Entities (NE)
  - Name, Company, Address, Profession, Telephone number, Date of birth, E-mail, URL
- Name
  - Name can occurs in different forms: John Smith, John Kennedy Smith, John K. Smith, and J. Smith [3, Lefever2010][4, Jiang2010]
- Name, Company, Address
  - Can be determined from text by: [4, Jiang2010]
    - Character Language Model – previously learned model from examples[5, Alias2010]
    - Hidden Markovov Model – dynamic Bayes network[6, HMM2010]

Named Entities II

- Profession, Telephone number, Date of birth
  - It was found, that these entities occurs more often near searched name than other entities [4, Jiang2010]
- E-mail
  - Addresses often contains part of searched name [4, Jiang2010]
- URL
  - URL referring to same page helps in person disambiguation
- Compound Key-Words (CKW)
  - Does not occurs often, but when they occur it is strong identifier of person [2, Yoshida2010]
  - Example: "chief software architect" in connection with "Bill Gates"

Disambiguation model

- The goal of algorithms is from the set PD make k clusters, where every cluster represents one real person
- The same person occurs in the different contexts, usually with few and often with noisy items [7, Han2009]
- More same Named Entities occurred on the same page are more likely true [7, Han2009]

Disambiguation methods

- TruthFinder [7, Han2009]
  - Different resources can contain dispute informations
  - Truth analysis should find true information about every object
- Distinct [7, Han2009]
  - Distinguish objects with same name
  - Uses Agglomerative Hierarchical Clustering and repeatedly groups most similar clusters
- Graph
  - Creating graph based on Named Entities
  - NE are nodes and edges between nodes represent occurrence of NE in one document
  - After running clustering algorithm on graph is graph split to clusters, where every cluster means one person

Disambiguation methods

- Fuzzy Ants [3, Lefever2010]
  - Inspired by ants which collect dead corpus of other ants to heaps
  - Number of heaps is not specified
  - Ants can pick up one item or heap of items
  - Ants have their own intelligence - whether pick up or release item/heap
- Two-stage Agglomerative Hierarchical Clustering
  - Words are split as *strong* and *weak*
  - Strong (NE and CKW) can split document between clusters, weak (*Algorithm* for programmer) cannot
  - In the first stage algorithm splits clusters based on strong words and assigns weights for weak words. In the second step take into account also weights of weak words.

What is done

Modules

- Search through web search engines: Google, Yahoo! (Bing uses Yahoo! results)
- ARES - Access to Registers of Economic Subjects
  - Search person by name
  - Basic – Basic info about person
  - Address – address standardization

- Výpis identifikačních údajů (Standard)
- Obchodní rejstřík (Commercial Register)
- Registr živnostenskho podnikání (Trade Register)
- Statistický registr RES (Register Economic Entities Statistical Office)
- Registr církví a náboženských společností (Churches and Religious Societies)
- Registr pojišťovacích zprostředkovatelů a likvidátorů pojistných událostí (Insurance Intermediaries Loss Adjusters)
- Seznam devizových míst a licencí (Foreign Exchange Spots and Licences)
- Seznam občanských sdružení a spolků (Civic Associations Guilds Clubs)
- Přehled ekonomických subjektů (Economic Entities)
- Registr zdravotnických zařízení (Register Of Health)
- Zemědělský registr (Common Agricultural Register)
- Registr politických stran a hnutí (Political Parties Movements)
- Rejstřík škol (School Register)
- Insolventní registr (Insolvent Register)

What is done

- Detection of the document encoding (cpDetector)
  - http://cpdetector.sourceforge.net/
- Processing of JavaScript (HtmlUnit project)
  - inline JavaScript, event handlers
  - http://htmlunit.sourceforge.net/javascript-howto.html
- Multi thread download manager
  - Download documents from web servers
  - Every server can have connection restrictions
    - IP address restrictions (proxy server)
    - Maximum m simultaneous downloads (in the one time server can process only m requests from one IP address)
    - Request delay (requests cannot be so often)
    - Maximum n requests in t time units (seconds, minutes, hours, days)
  - Violate these restrictions can lead to blocking up of the server

I'm working on

Download web pages from different sites

- Why?
  - Use searching on these web pages
  - Not every page is indexed by search engines (`robots.txt`)
  - Pages use JavaScript (events)
- Automatic transformation of the page to RDF/FOAF format)in the case result is XML or well structured
- When it is not easy to recognize Named Entities and document require linguistic pre-processing, document is just downloaded
- Sites involved: facebook.com, foaf.sk,

Future

- Extract Named Entities from documents
- Cluster documents
- Suggest the most likely person
- Visualize (time, time, time)
- Any suggestions?


[Vu2007] Vu, Quang Minh, Tomonari Masada, Atsuhiro Takasu, and Jun Adachi
Disambiguation of People in Web Search Using a Knowledge Base.
2007 IEEE International Conference on Research, Innovation and Vision for the Future (March 2007): 185-191


[Yoshida2010] Yoshida, Minoru
Person Name Disambiguation by Bootstrapping.
Search (2010): 10-17


[Lefever2010] Lefever, E., T. Fayruzov, V. Hoste, and M. De Cock
Clustering web people search results using fuzzy ants.
Information Sciences 180, no. 17 (September 2010): 3192-3209


[Jiang2010] Jiang, Lili, Wei Shen, Jianyong Wang, and Ning An.

GRAPE : A System for Disambiguating and Tagging People Names in Web Search.
Analysis (2010): 1257-1260

[Alias2010] Alias-i
LingPipe: Character Language Model Tutorial
URL: http://alias-i.com/lingpipe/demos/tutorial/lm/read-me.html

[HMM2010] Hidden Markov model.
In Wikipedia, The Free Encyclopedia.
URL: http://en.wikipedia.org/w/index.php?title=Hidden_Markov_model&oldid=398697711

[Han2009] Han, Jiawei
Mining Heterogeneous Information Networks by Exploring the Power of Links (2009): 13-30