Personal Name Resolution of Web People Search

Krisztian Balog
University of Amsterdam

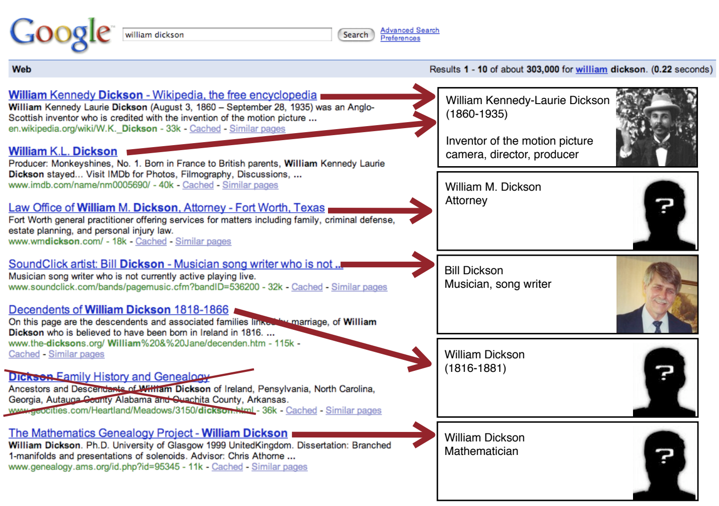
Leif Azzopardi
University of Glasgow

Maarten de Rijke
University of Amsterdam

WWW2008 Workshop
NLP Challenges in the Information Explosion Era (NLPIX 2008)
April, 22, 2008, Beijing China

Motivation, task

- People-related search tasks
  - E.g., building profiles, creating biographies, finding experts, etc.
  - 5-10% of web searches contain person names (Schein et al., SIGIR 2002)
- Task of personal name resolution
  - Given a set of documents, all of which refer to a particular person name
  - Identify which documents are associated with each single individual (referent)
  - Generally approached as a clustering problem



The person clustering hypothesis

- Cluster hypothesis (Jardine and van Rijsbergen, 2001)
  - *Similar documents tend to be relevant to the same request*
- Re-stated in the context of personal name resolution: "person clustering hypothesis"
  - *Similar documents tend to represent the same person (referent)*

In this paper...

- Examine to which extent the person clustering hypothesis holds under the most general conditions
  - Only feature: distribution of terms in documents
- Two forms of clustering, identifying relationships between documents
  - Term level
  - Latent space

Outline

- **Clustering approaches**
  - Assumptions
  - Single Pass Clustering
  - Probabilistic Latent Semantic Analysis
- Evaluation platform
- Experiments and results
- Conclusions

Assumptions

1. One document is associated with one referent
2. The distribution of documents assigned to referents follows a power law
3. Every document refers to a distinct person sense, unless there is evidence to the contrary
4. The number of person senses is not known a priori (but is limited by the number of documents available)
5. Documents are unstuctured (no guarantees about the format or structure within documents)

Single Pass Clustering (SPC)

- Mimic user behavior
- For each document
  - If a cluster representing that person already exists, then assign document to that cluster
  - Otherwise assign it to a new cluster
- Capitalize on the fact that most popular (dominant) senses of the person name are highly ranked
- Very efficient, can be computed online

SPC (2)

- Document is assigned to the most similar cluster as long as

  (1) similarity is higher than a threshold

  $$SIM(D,C) > \gamma$$

  (2) maximum number of clusters has not been reached

  - if reached, assign document to the last cluster ("left overs")

SPC (3)

Measuring document and cluster similarity

- (SPC-NB) Naive Bayes

  $$sim(D,C) = O(D,C)$$

  $$O(D,C) = \frac{p(D|\theta_C)}{p(D|\theta_{\bar{C}})} = \frac{\prod_{t \in D} p(t|\theta_C)^{n(t,D)}}{\prod_{t \in D} p(t|\theta_{\bar{C}})^{n(t,D)}}$$

- (SPC-COS) Cosine using TF.IDF weighting

  $$\mathrm{sim}(D,C) = \cos(\vec{t}(D), \vec{t}(C)) = \frac{\vec{t}(D) \cdot \vec{t}(C)}{||\vec{t}(D)|| \cdot ||\vec{t}(C)||}$$

Probabilistic Latent Semantic Analysis (PLSA)

- Decomposition of the term-document matrix into a lower dimensional latent space

  $$p(t,d) = p(d) \sum_z p(t|z) p(z|d)$$

- Obtained using the EM algorithm
- Each latent topic z represents one of the different senses of the person name

PLSA (2)

- A document *d* is assigned to one of the person-topics z, if

  (1) *p*(z|*d*) is the maximum argument

  (2) odds of the document given z is greater than a threshold: $O(z,d) > \gamma$

  $$O(z,d) = \frac{p(z|d)}{p(\bar{z}|d)} = \frac{p(z|d)}{\sum_{z',z' \neq z} p(z'|d)}$$

PLSA (3)

- Automatically finding the number of person senses (i.e., |z|)

(1) set z=2, compute the log-likelihood of the decomposition

(2) increment z and compute the log-likelihood again

  - if log-likelihood increased (>0.001), then repeat (2)
  - else goto (3)

(3) STOP

Outline

- Clustering approaches
- **Evaluation platform**
  - Data set
  - Performance measures
  - Document representation
- Experiments and results
- Conclusions

Data set

- WePS 2007 platform
(Web People Search track at the Semantic Evaluation Workshop 2007)
- Web pages obtained from the top (up to) 100 results for a person name query to a web search engine
- Each page from the result list is stored
  - URL, title, position in the ranking, snippet

Data set (2)

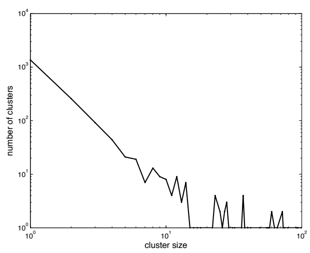
- Annotators manually classified each web page
  - Original task statement allows a document to be assigned to multiple clusters
  - Some documents were discarded (e.g. out-of-date)
- Training (49 names) and test (30 names) sets
- Names from 4 different sources
  - US Census, Wikipedia, ECDL06, ACL06

Data set - sources

Data set / source	#names	avg(docs)	discarded	referents
Training set	**49**	**71.02**	**26.00**	**10.76**
US Census	32	47.20	18.00	5.90
Wikipedia	7	99.00	8.29	23.14
ECDL06	10	99.20	30.30	15.30
Test set	**30**	**98.93**	**15.07**	**45.93**
US Census	10	99.10	14.90	50.30
Wikipedia	10	99.30	17.50	56.50
ACL06	10	98.40	12.80	31.00

- Ambiguity in the test data is much higher than in the training data
- To measure performance as reliably as possible, we use *all names*

Distribution of documents to person senses



- Size of the clusters follows a power law
  - Exponent of approx. 1.31
- Confirms our assumption (2) about the data

Performance measures

- Standard clustering measures
  - Purity — “precision”
    - Rewards methods that introduce less noise in each cluster
  - Inverse purity — “recall”
    - Rewards methods that gathers more elements of each class into a corresponding single cluster
- F-measure (weighted average of purity and inv. purity)
  - $F_{0.5}$ harmonic mean
  - $F_{0.2}$ user’s point of view (more importance to inv. purity)
  - $F_{0.8}$ machine’s point of view (more importance to purity)

Document representation

- Separate index for each person
- Document is represented using
  - Title and snippet from the search engine’s output
  - Body text extracted from HTML
    - Segments of the page, separated by block-level HTML tags, that contain 10 or more words

Outline

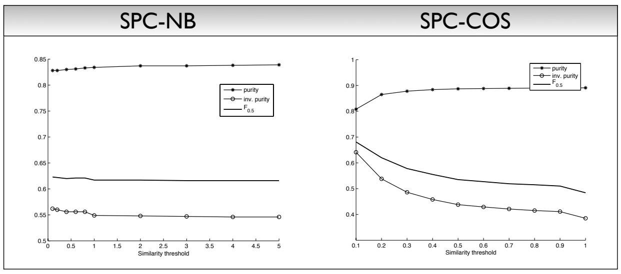
- Clustering approaches
- Evaluation platform
- **Experiments and results**
  - SPC, PLSA
  - Comparing methods
  - Group-level analysis
  - Comparison to other approaches
- Conclusions

Research questions

- What factors affect performance?
  - Similarity threshold
  - Limiting the number of clusters
- How stable is performance?
- What is the best number of clusters to use? Can we determine this automatically?

SPC

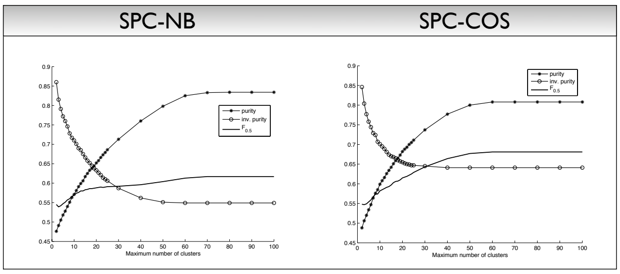
Similarity threshold



- Performance is stable w.r.t. the threshold
- Best performance is obtained with low threshold

SPC

Limiting the number of clusters



- Enforcing a limit on the number of clusters hurts (independent of the similarity threshold)

PLSA
Experimental conditions

- Manual
  - Assuming that each latent topic is representative of each person-sense
  - Set the number of latent topics to the actual number of person senses (based on the ground truth)
  - Should provide a theoretical upper bound
- Auto
  - Realistic experimental setting
  - Unsupervised learning

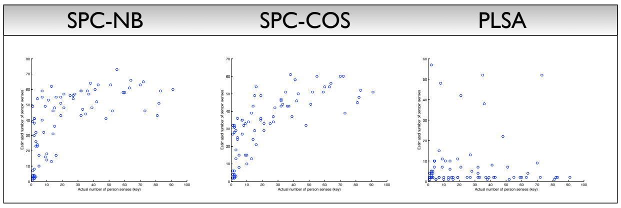
PLSA
Results

Exp. cond.	pur.	invp.	$F_{0.5}$	$F_{0.2}$	$F_{0.8}$
Manual	0.530	0.647	0.547	0.591	0.530
Auto (0.5)	0.495	0.800	0.536	0.624	0.501
Auto (1.0)	0.517	0.782	0.543	0.622	0.515
Auto (5.0)	0.662	0.647	0.561	0.583	0.584

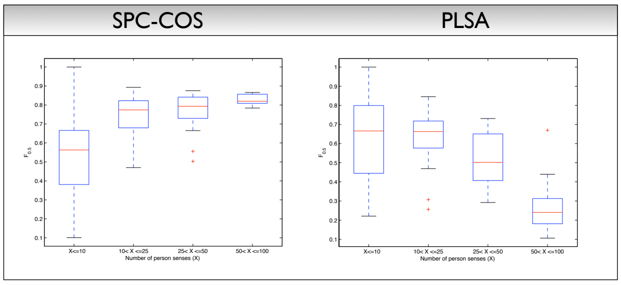
- Manual setting does not perform very well
  - Latent topics are not really that representative of the individual person senses
- The automatic method identifies a relatively small number of clusters
  - Latent topics are dominated by a few "principal" components

Comparing methods

Method	All names				
	pur.	invp.	$F_{0.5}$	$F_{0.2}$	$F_{0.8}$
SPC-NB	0.828	0.562	0.623	0.579	0.705
SPC-COS	0.808	0.641	0.681	0.651	0.736
PLSA	0.517	0.782	0.543	0.622	0.515



Performance against different cluster sizes



Findings

- SPC
  - Good estimate of person senses
  - High purity scores
- PLSA
  - Underestimates the number of person senses
    - Identifies the prominent person senses, but fails when only limited examples (1-2 docs) of the other referents are available
  - Very high inverse purity
    - referents are usually not dispersed among clusters

Comparison to other approaches

Method	Test set			
	pur.	invp.	$F_{0.5}$	$F_{0.2}$
"Naive baselines"				
ONE-IN-ONE	1.000	0.470	0.610	0.520
ALL-IN-ONE	0.290	1.000	0.400	0.580
This paper				
SPC-NB	0.884	0.688	0.747	0.707
SPC-COS	0.850	0.777	0.791	0.780
PLSA	0.370	0.885	0.442	0.581
SemEval 2007 Top 3				
CU_COMSTEM	0.720	0.880	0.780	0.830
IRST-BP	0.750	0.800	0.750	0.770
PSNUS	0.730	0.820	0.750	0.780

Wrap up

- Task of person name resolution in web search
- Two approaches
  - SPC (term based)
  - PLSA (semantic based)
- SPC outperforms PLSA and delivers excellent performance
- The "person clustering hypothesis" holds to a large extent

Future work

- Combine advantages of both methods
- Richer feature set (e.g., named entities)
- Pre-processing documents (removing irrelevant content)

Questions?

Krisztian Balog

kbalog@science.uva.nl
http://www.science.uva.nl/~kbalog