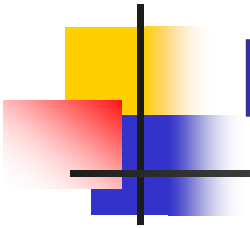
Personal Name Classification in Web queries

Dou Shen*, Toby Walker*, Zijian Zheng*,
Qiang Yang**, Ying Li*

*Microsoft Corporation
** Hong Kong University of Science and Technology

Introduction

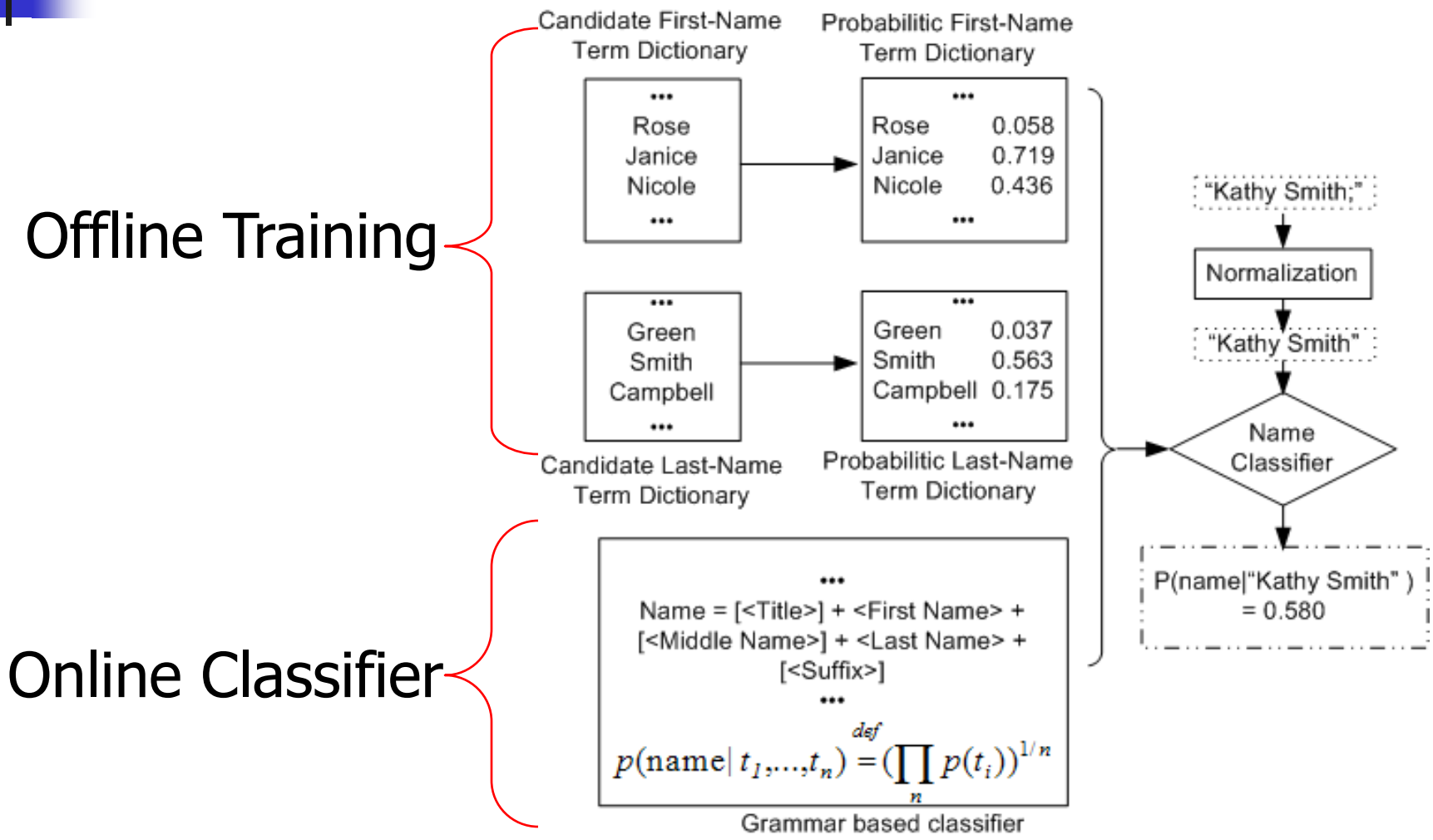
- Goal:
  - To detect whether a Web query is a personal name, without referring to any other context information;
- Motivation
  - 2~4% of daily Web queries are personal names
  - ~6% of jumping queries are personal names
  - Users tend to test a search engine by their names
- Applications
  - Paid Search
    - "toby walker" → "Wheeled Walkers Sale $89."
  - Query Suggestion
    - Show the profile-related information once a query is determined as a personal name



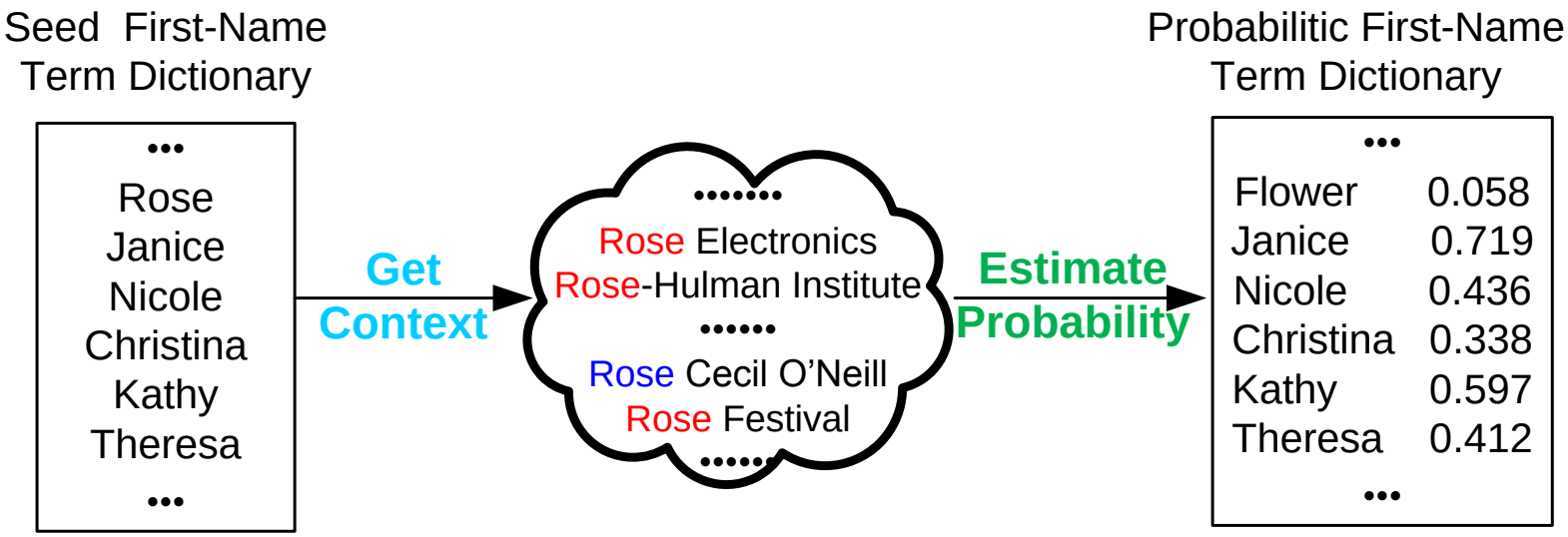
Related Work

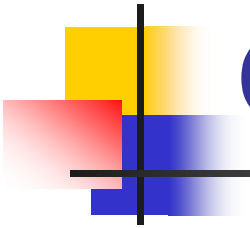
- Improve Personal Name Search
  - Dozier studied some specific strategies for personal name search [1]
  - Wan et al. studied a person resolution system to improve people search performance [2]
- Personal Name Extraction/Recognition
  - As a special case of named entities, personal name extraction from document/webpage/emails has been widely studied recently [3, 4, 5]
- Web Query Enrichment
  - Use click-trough data [6]
  - User Web Search [7]

Overview of Our Solution



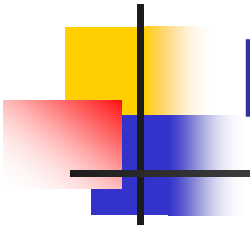
Offline Training





Offline Training

- Terminology
  - Candidate Dictionaries
  - Term Context: "…toby walker…"; "…city of walker…"
  - Name Term Context: "…toby walker…"
  - Name Context: "…Dr. Qiang Yang's student…"
- Probability Estimation Methods
  - Relative Frequency
  - Context Probability
  - Co-Occurrence in Search Snippets
  - Co-Occurrence in Bigrams



Probability Estimation Methods (1)

- Relative Frequency
  - Get a set of names
  - Get the relative frequency of each term
- Context Probability
  - Assumption: if a term is name term, its term contexts should be name contexts
  - Train a unigram model over some name contexts
  - Given a term, get its term contexts through search engines and calculate the probability of term contexts using the trained model

Probability Estimation Methods (2)

- Co-Occurrence in Search Snippets (S-CoOcc)
  - Get term contexts through search engines
  - Identify name term contexts using some rules
    - followed by a last name term, such as "john smith";
    - followed by a first name term and then a last name term, such as "John Maynard Smith";
    - followed by a special kind of verbs such as "did, said, announced, claimed…" as in "John said";
    - ……
  - Estimate the term probability as the ratio between name term contexts and term contexts
  - Golden Dictionaries

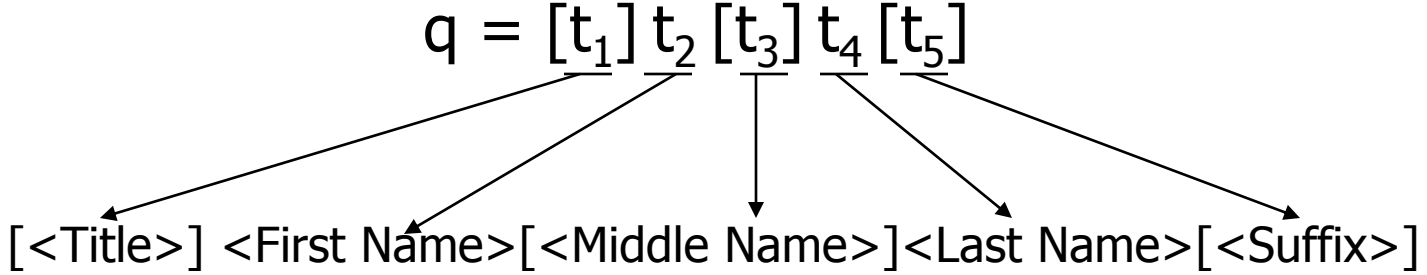
Probability Estimation Methods (3)

- Co-Occurrence in Bigrams (B-coOcc)
  - Get term contexts from a bigram file
  - Estimate the term probability as in S-coOcc

.....			
2227	2173	toby	up
1250	758	toby	wachter
2307	1304	toby	walker
1013	358	toby	walking
1178	689	toby	walsh
18782	12951	toby	was
2244	1876	toby	we
.....			

Online Classifier (1)

- Grammar matching



- Geometric average

$$p(name \mid q) = \left(\prod_{i=1}^{n} (p(t_i) \right)^{1/n}$$

- Tricks
  - p(title) = 1; p(suffix)=1

Online Classifier (2)

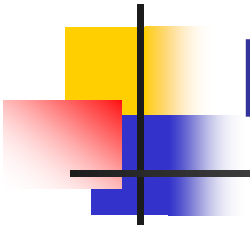
- Personal Name Grammars
  - <Personal Name> ::= [<Title>] <First Name>[<Middle Name>]<Last Name>[<Suffix>]
  - <Title> ::= dr | doctor | ms, …
  - <Suffix> ::= sr.| jr.| III …
  - <First Name> ::= <first name term>|
    <first name term>-<first name term>;
  - <Middle Name>::= <First Name> | <Last Name>;
  - <Last Name> ::= <last name term>|
    <last name term>-<last name term>;

Experiments

- Data Sets
  - Dictionaries

	DBLP	CENSUS	WP
First Name	64,187	5,494	762,905
Last Name	167,965	88,799	2,045,637

  - Name Context:
    - Top 200, 000 personal names from WP
    - 10,000,000 name contexts
  - Term Context
    - 2,000,000 candidate terms
    - 80,000,000 term contexts
  - Testing Data Sets
    - Validation dataset: 2,000 queries, 81 names
    - Test dataset: 10,000 queries, 232 names



Baselines

- Dictionary Look-up
- Supervised Methods
  - Classifiers
    - SVM, Logistic Regression
  - Features
    - f_1: the length of the query;
    - f_2: whether the query contains a title term;
    - f_3: whether the query contains a suffix term;
    - f_4: the probability of a term being a first name term
    - f_5: the probability of a term being generated by a character level bigram model trained on first-name terms;

Experiment Results (1)

- Comparison among our method and the baselines

		Pre	Rec	F1
Our Methods	Prob.DBLP	0.891	0.457	0.604
	Prob.CENSUS	**0.942**	0.487	0.642
	Prob.WP	0.779	0.819	**0.798**
Dictionary Look-up	Bool.DBLP	0.651	0.466	0.543
	Bool.CENSUS	0.803	0.491	0.610
	Bool.WP	0.127	**0.884**	0.222
Supervised Methods	SVM	0.798	0.595	0.681
	LR	0.785	0.659	0.717

Experiment Results (2)

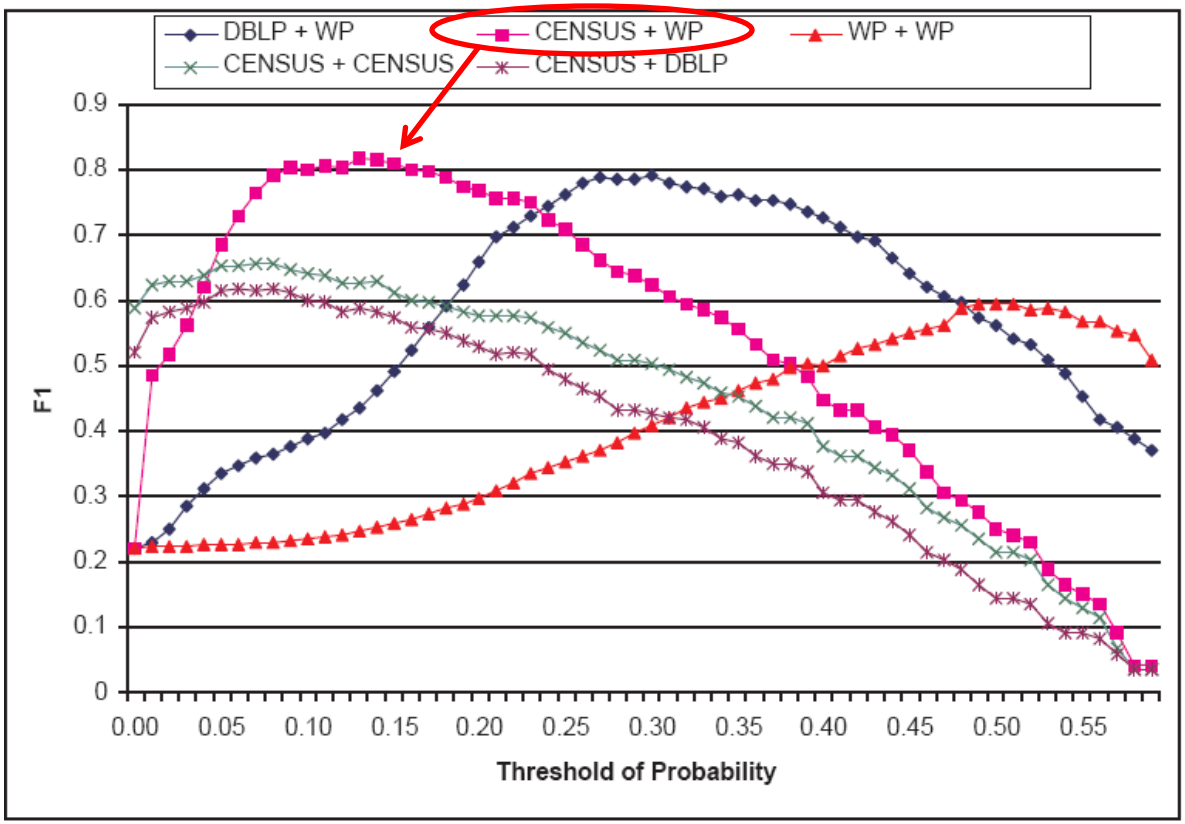
- Comparison of different ways of constructing probabilistic dictionaries.

	Pre	Rec	F1
S-coOcc	**0.779**	**0.819**	**0.798**
B-coOcc	0.765	0.672	0.716
RF	0.628	0.349	0.449
CP	0.237	0.569	0.335

  - Candidate Dictionaries: WP
  - Golden Dictionaries: CENSUS

Experiment Results (3)

- Effect of Golden Dictionaries and Candidate Dictionaries



Experiment Results (4)

- Effect of Enlarging Golden Dictionaries

	0	1	2	3	4	5
Pre	0.779	0.797	0.844	0.948	0.914	0.910
Rec	0.819	0.810	0.772	0.707	0.737	0.724
F1	0.798	0.803	0.806	0.810	**0.816**	0.806

  - Golden dictionaries: CENSUS and its expansions
  - Candidate dictionaries: WP
  - S-coOcc

Experiment Results (5)

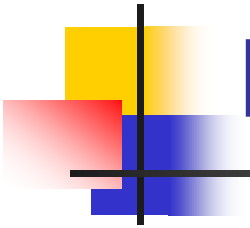
- Effect of Enlarging Candidate Dictionaries

	0	1	2	3	4	5
Pre	0.944	0.802	0.789	0.737	0.737	0.737
Rec	0.366	0.595	0.595	0.603	0.603	0.603
F1	0.528	**0.683**	0.678	0.664	0.664	0.664

  - Gold dictionaries: CENSUS
  - Candidate dictionaries: CENSUS and its expansions
  - B-coOcc

Conclusion and Future Work

- Conclusion
  - Put forward an easy but effective method for personal name classification in Web queries
  - Exploit the methods of enlarging golden dictionaries and candidate dictionaries.
- Future work
  - Instead of using rules, try to define name term contexts using existing named entity recognition algorithms
  - Validate the contribution of personal name classification to Web Search and Advertising
  - Validate the classifier in non-US names



References

[1] C. Dozier. Assigning belief scores to names in queries. *HLT '01.*

[2] X. Wan, J. Gao, M. Li, and B. Ding. Person resolution in person search results: Webhawk. C*IKM '05*

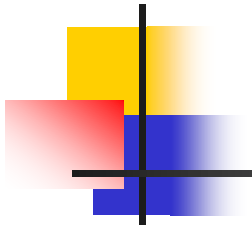
[3] Z. Chen, L. Wenyin, and F. Zhang. A new statistical approach to personal name extraction. *ICML '02:*

[4] H. L. Chieu and H. T. Ng. Named entity recognition with a maximum entropy approach. HLT-NAACL '03

[5] V. Krishnan and C. D. Manning. An effective two-stage model for exploiting non-local dependencies in named entity recognition. ACL-COLLING '06

[6] J.-R. Wen, J.-Y. Nie, and H.-J. Zhang. Clustering user queries of a search engine. WWW '01

[7] D. Shen, R. Pan, J.-T. Sun, J. J. Pan, K. Wu, J. Yin, and Q. Yang. Query enrichment for web-query classification. TOIS '06.



Thanks!