Interactive Classification of Keyword Search Queries

Master's Thesis Presentation

Sebastian Arnold <sarnold@mailbox.tu-berlin.de>

Fachgebiet Datenbanksysteme und Informationsmanagement
Technische Universität Berlin

http://www.dima.tu-berlin.de/

Outline

1) **Model of Search Intentions**
2) Process of Search
3) Feature Extraction and Classification
4) Evaluation
5) Summary and Future Work

Observation of Keyword Queries (1)

- What do these queries have in common?
  - barack obama
  - paris hilton model
  - michael and janet jackson
  - white house employees
  - IOS vs android
  - age of bill gates
  - travel by bike

Observation of Keyword Queries (2)

- What do these queries have in common?
  - barack obama
  - paris hilton model
  - michael and janet jackson
  - white house employees
  - IOS vs android
  - age of bill gates
  - travel by bike
- They have **informational search goals**. [Bro02]

family relations of Michael Jackson — showing 1 to 10 of 150 entries — prev | next

	person_relative compare ▾	familyrelationtype	person compare ▾	proofs
1	Joe Jackson	father	Michael Jackson	13 found
2	Katherine Jackson	mother	Michael Jackson	13 found
3	Jermaine	brother	Michael Jackson	9 found
4	Katherine	mother	Michael Jackson	9 found
5	Debbie Rowe	ex-wife	Michael Jackson	9 found
6	Lisa Marie Presley	wife	Michael Jackson	7 found
7	Michael Jackson	brother	Janet Jackson	6 found
8	Jermaine Jackson	brother	Michael Jackson	6 found
9	La Toya	sister	Michael Jackson	5 found
10	Debbie Rowe	wife	Michael Jackson	4 found

Observation of Keyword Queries (3)

- But is there a **difference** between these queries?
  - `barack obama`
  - `paris hilton model`
  - `michael and janet jackson`
  - `white house employees`
  - `IOS vs android`
  - `age of bill gates`
  - `travel by bike`

Observation of Keyword Queries (3)

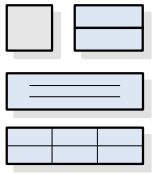
- But is there a **difference** between these queries?
  - `barack obama`
  - `paris hilton model` — undirected
  - `michael and janet jackson`

  - `white house employees`
  - `IOS vs android` — directed
  - `age of bill gates`

  - `travel by bike` — advice
- Yes! They express different types of **goals**. [Ros04]
  - scope
  - expected result
  - complexity

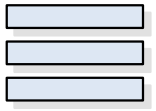
Classification of Informational Intentions

- **Undirected** [Ros04]
  - EXPLORE — `barack obama`
  - RESOLVE — `paris hilton model`
  - RELATE — `michael and janet jackson`

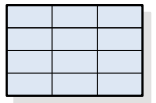


over-view

- **Directed** [Ros04]
  - LIST — `white house employees`
  - COMPARE — `IOS vs android`
  - ANSWER — `age of bill gates`



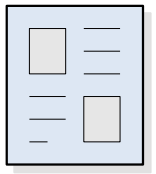
list



table

fact

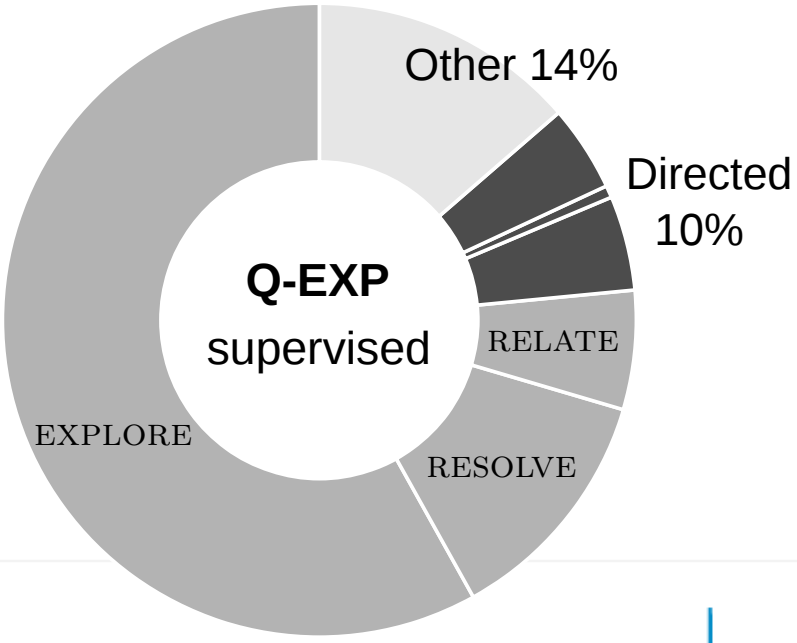
- **Advice** [Ros04] `travel by bike`



article

Observation of Intent Classes

- 102,360 queries posed to GoOLAP 08/2008–10/2014
- Unstructured queries (from referrer)
- Returns factual results
- Large knowledge base [Lös12]
  - 5 million named entities
  - 29 million facts
  - 70 schema types

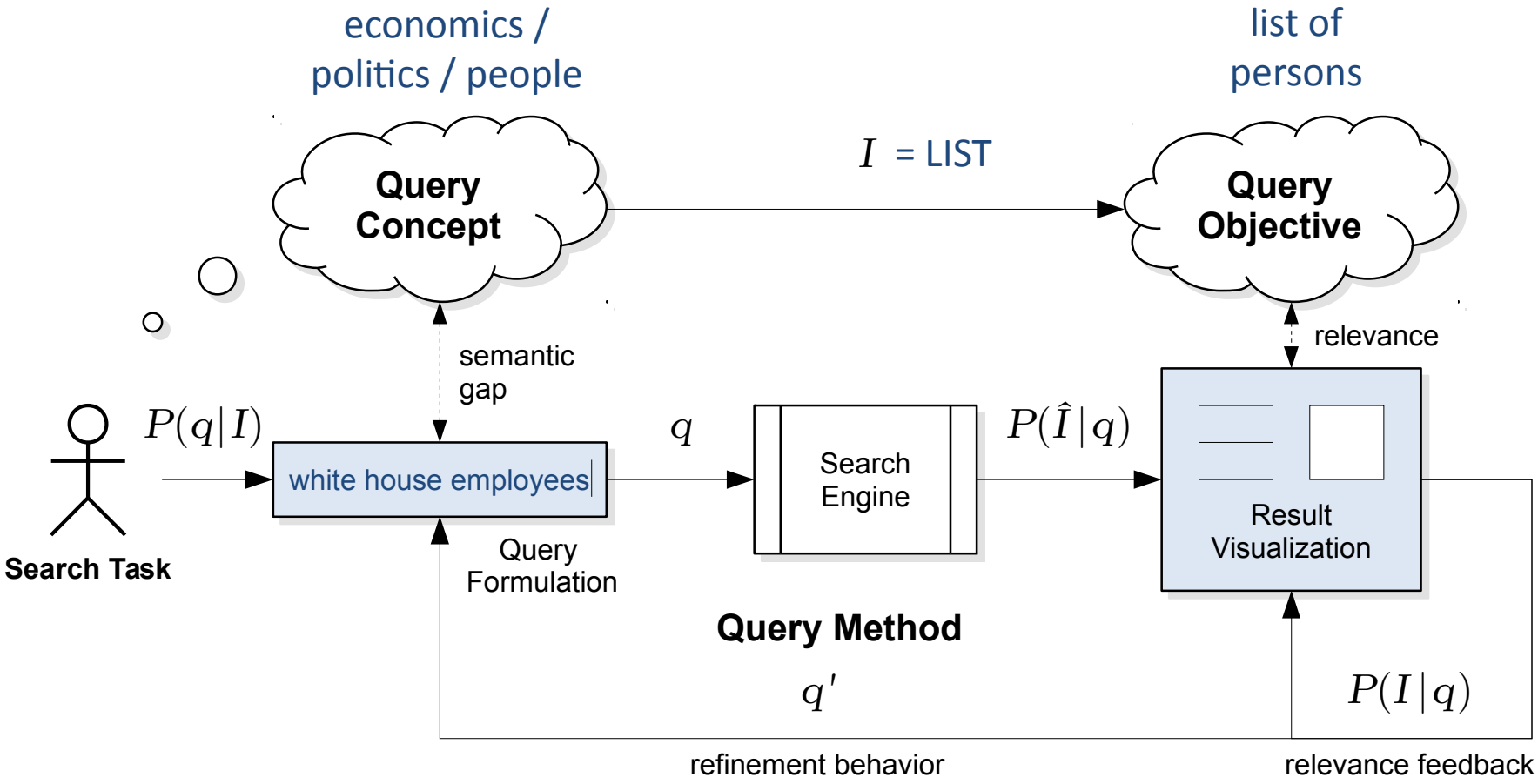


goolap
Search. Explore. Decide.

Outline

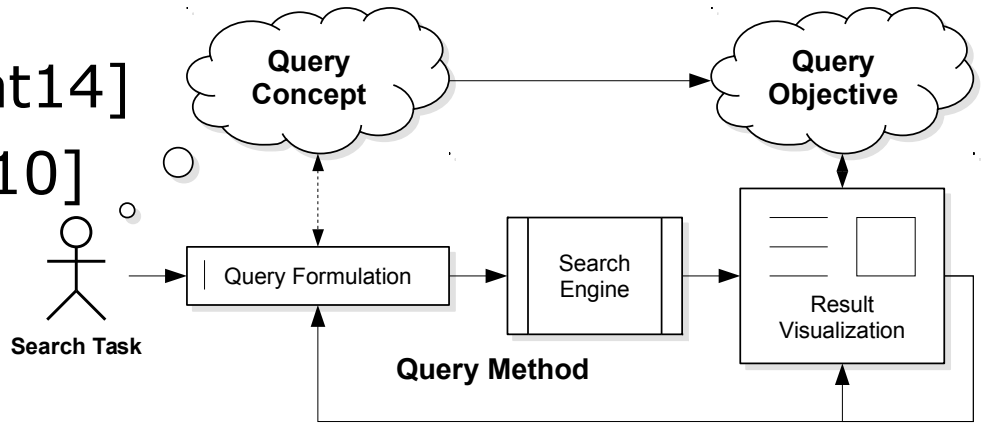
1) Model of Search Intentions
2) **Process of Search**
3) Feature Extraction and Classification
4) Evaluation
5) Summary and Future Work

Process of Informational Search



Dimensions of Informational Search

- **Query Concept**
  - genre, topic, time sensitivity, … [Gon11]
- **Query Objective**
  - specificity *(specific – exhaustive)* [Gon11]
  - explorativeness *(directed – undirected)* [Aul08]
  - task complexity *(lookup – learn – investigate)* [Mar06]
  - result form and size [Kat14]
- **Query Method**
  - query formulation [Kat14]
  - user interaction [Guo10]
  - …

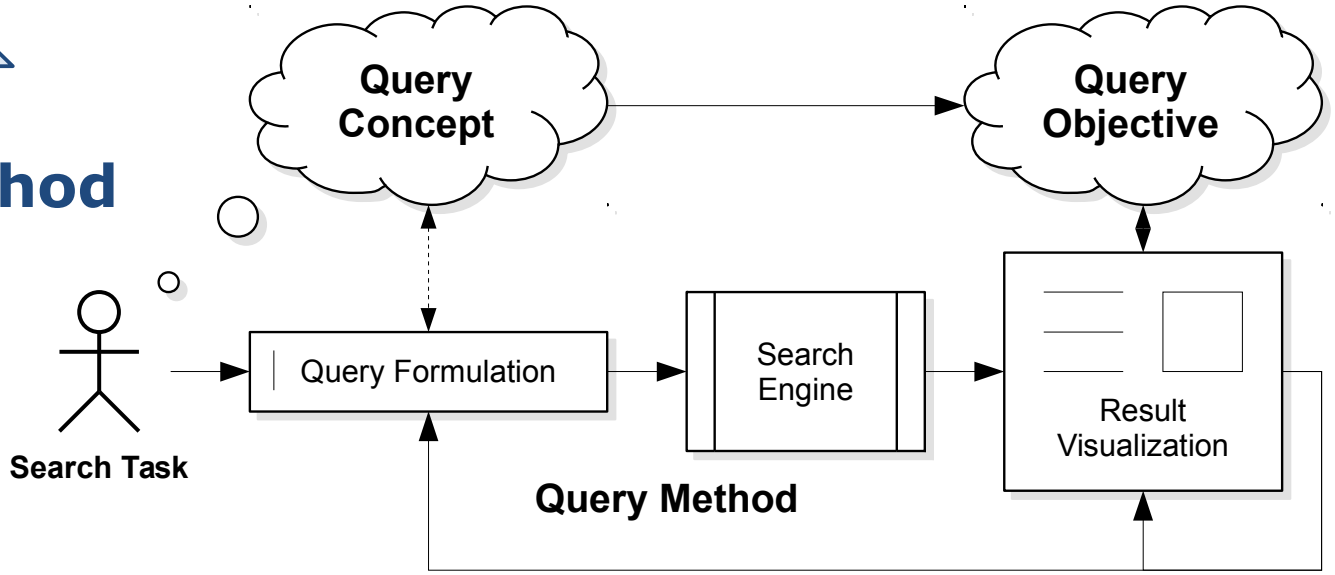


Predict Query Objective

- **Goal: Prediction of Informational Search Intent**

- **Query Objective**
  - EXPLORE / RESOLVE / RELATE
  - LIST / COMPARE / ANSWER

- **Query Method**



Outline

1) Model of Search Intentions
2) Process of Search
3) **Feature Extraction and Classification**
4) Evaluation
5) Summary and Future Work

Feature Extraction (1)

- Let's use a **POS tagger** to extract query syntax!
  - `barack obama`
  - `paris hilton model`
  - `white house employees`
  - `age of bill gates`

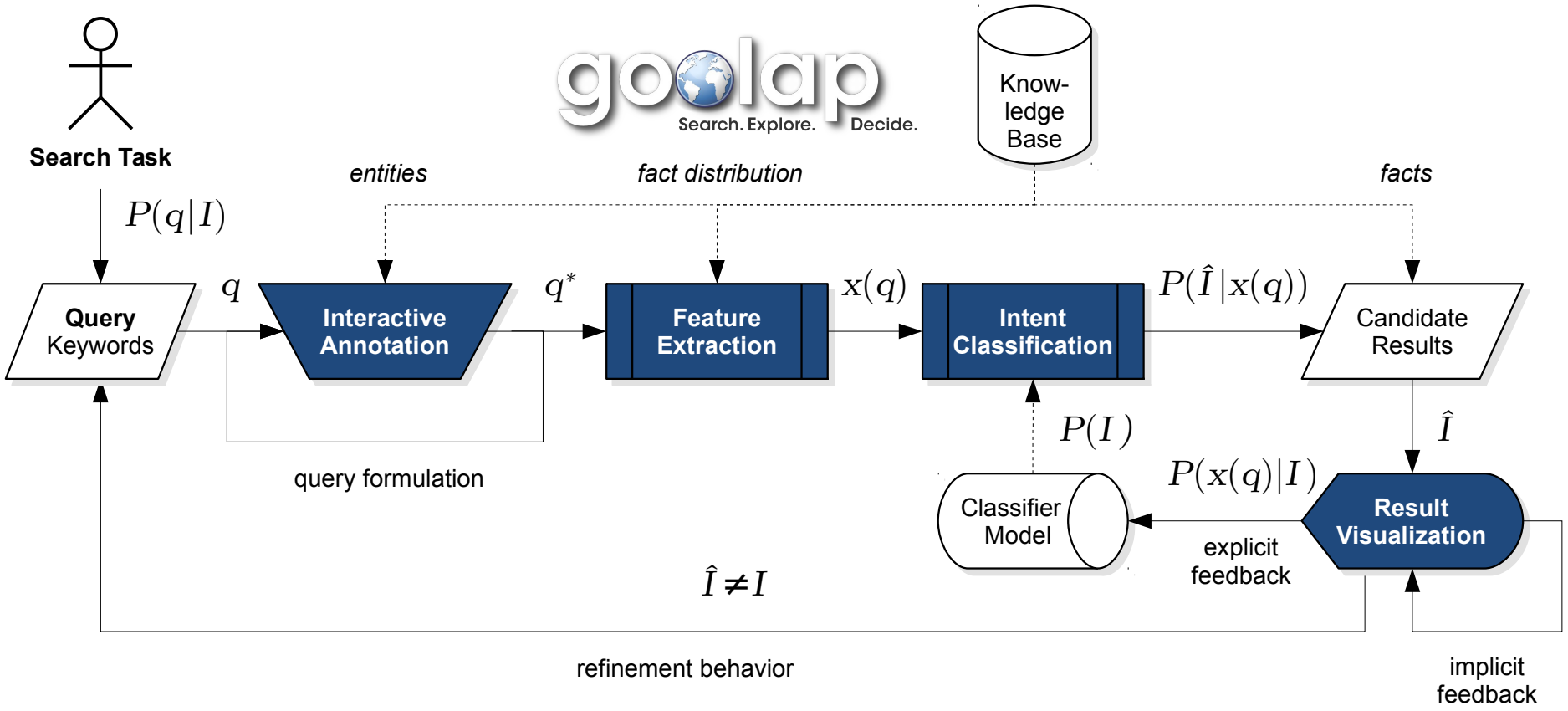
Feature Extraction (2)

- Let's use a **POS tagger** to extract query syntax!
  - `barack obama` → NN NN
  - `paris hilton model` → NN NN NN
  - `white house employees` → JJ NN NNS
  - `age of bill gates` → NN IN NN NNS
- This doesn't work. Queries...
  - ...are short
  - ...are ambiguous
  - ...are not correct English sentences

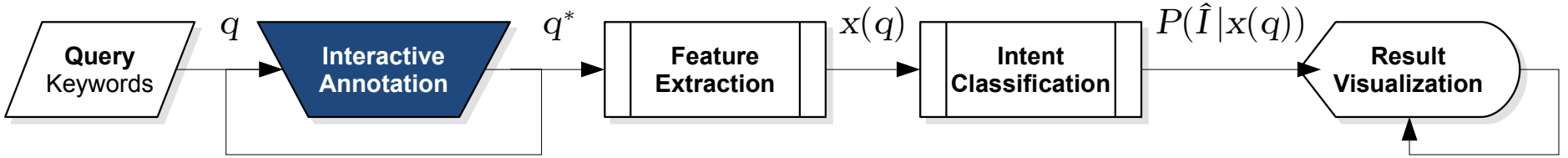
Feature Extraction (3)

- Better idea: **entity recognition** / schema matching
  - barack obama → NNP
    Person
  - paris hilton model → NNP NNP
    Person Position
  - white house employees → NNP NN+
    Organization PersonCareer+
  - age of bill gates → NN of NNP
    Attr of Person
- Requires user interaction or heuristics
  - Query segmentation (NP-hard)
  - Entity recognition
  - Entity disambiguation

Implementation of the Search Process

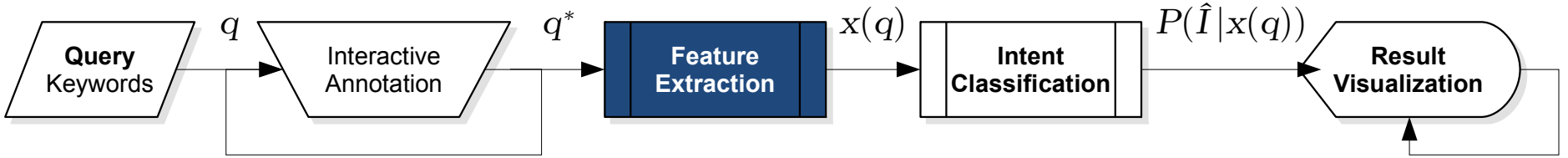


Interactive Query Annotation



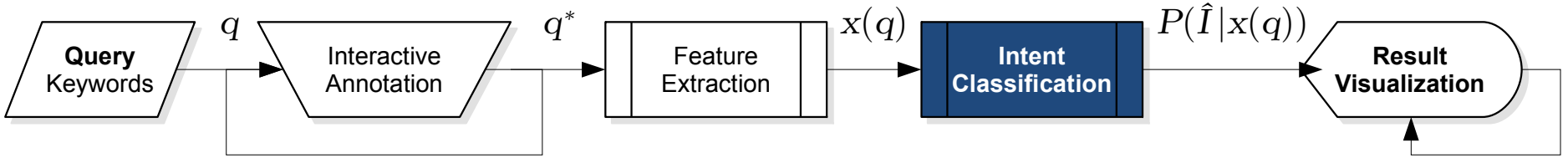
- We incorporate the user into the extraction process
  - Entity disambiguation from GoOLAP knowledge base
  - Right-longest-first auto complete
  - Search field with interactive annotations

Feature Extraction



- Queries are then transformed into simple tag sequences:
  - Example: `white house employees`
    `Organization` `PersonCareer+`
  - SEQpos: NNP NN+
  - SEQtype: Organization PersonCareer+
- Use sequences to produce a feature vector x(q)
  - count tags (e.g. #segments, #NN+)
  - match patterns (e.g. `"Person Person"~2`)

Classification

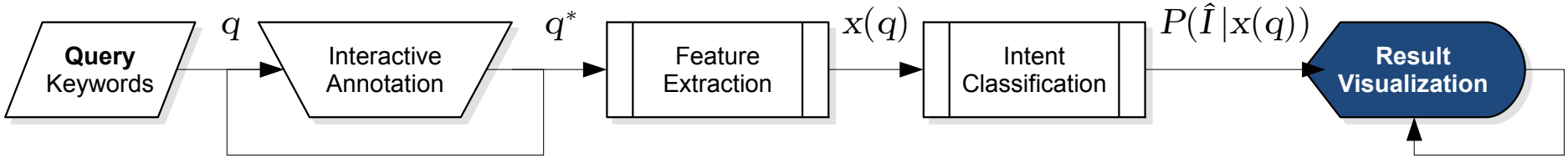


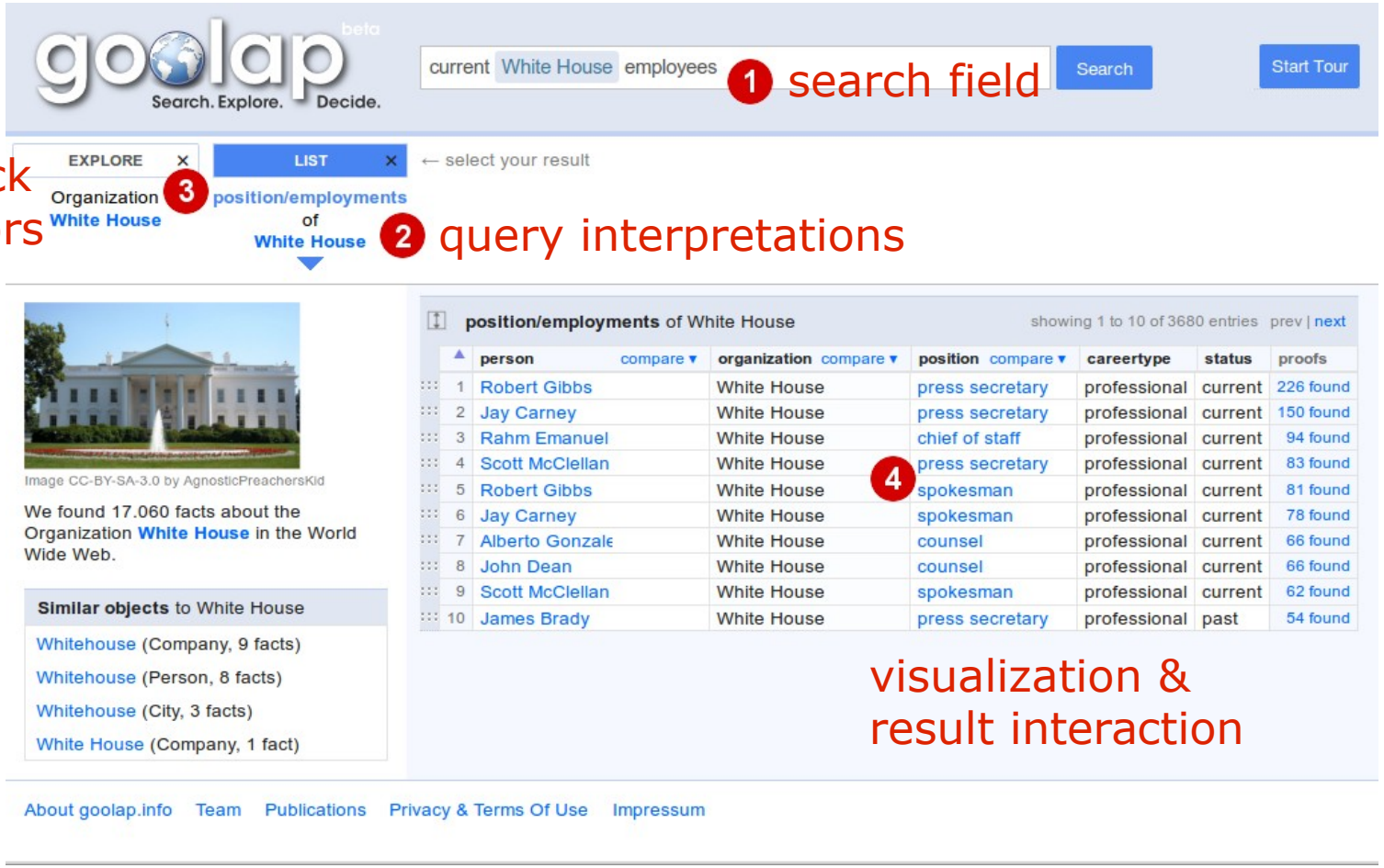
- Estimate $\hat{\mathrm{I}}$ from features x(q)

$$\hat{I} = \underset{k=1,\ldots,K}{\arg\max} P(I = I_k | x(q))$$

- **Labeled data set** Q-EXP (n=477)
  - Prior $\mathrm{P(I)}$, Likelihood $\mathrm{P(x(q)|I)}$
- Implementation of classifier configurations
  - Trivial baseline: always predict EXPLORE
  - Lucene Language Models
  - Naive Bayes supervised (C-SUP)
  - Naive Bayes semi-supervised (C-SEMI) (n=85,430)

Classification

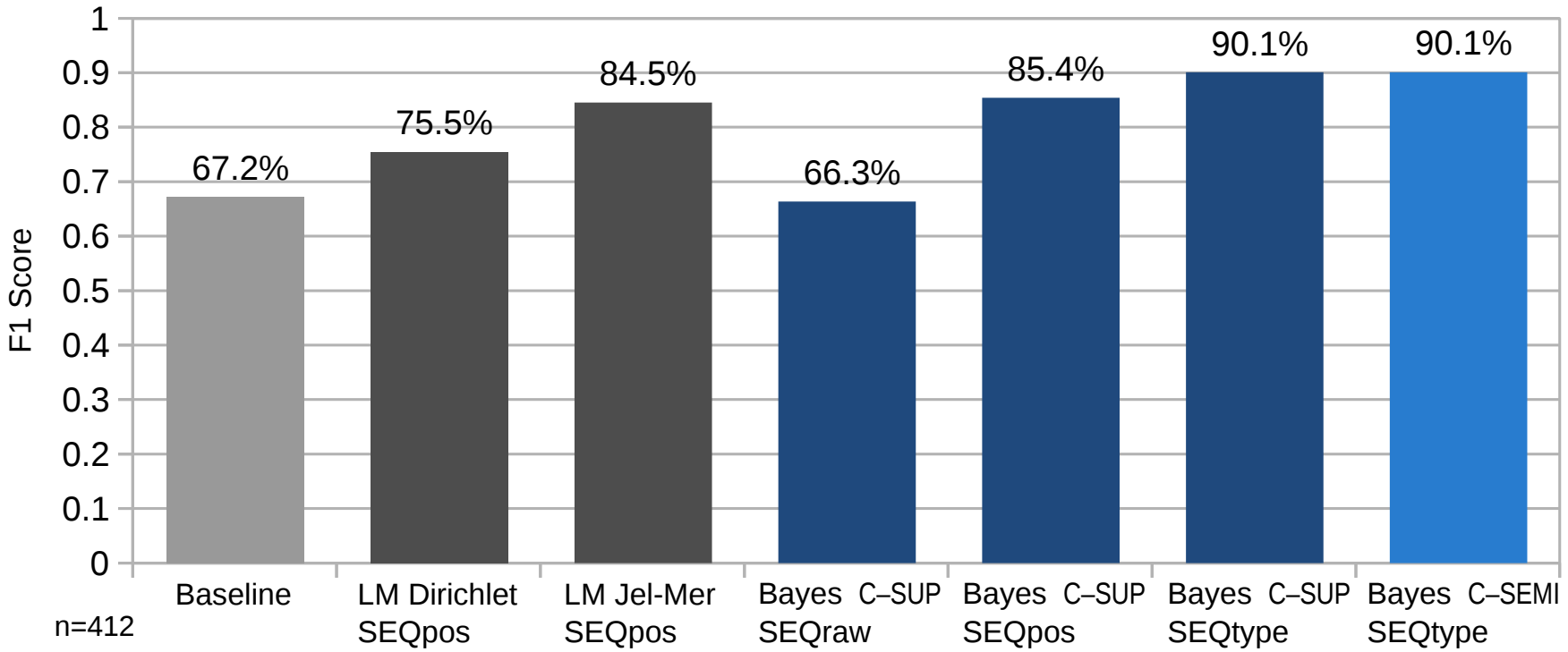




Outline

1) Model of Search Intentions
2) Process of Search
3) Feature Extraction and Classification
4) **Evaluation**
5) Summary and Future Work

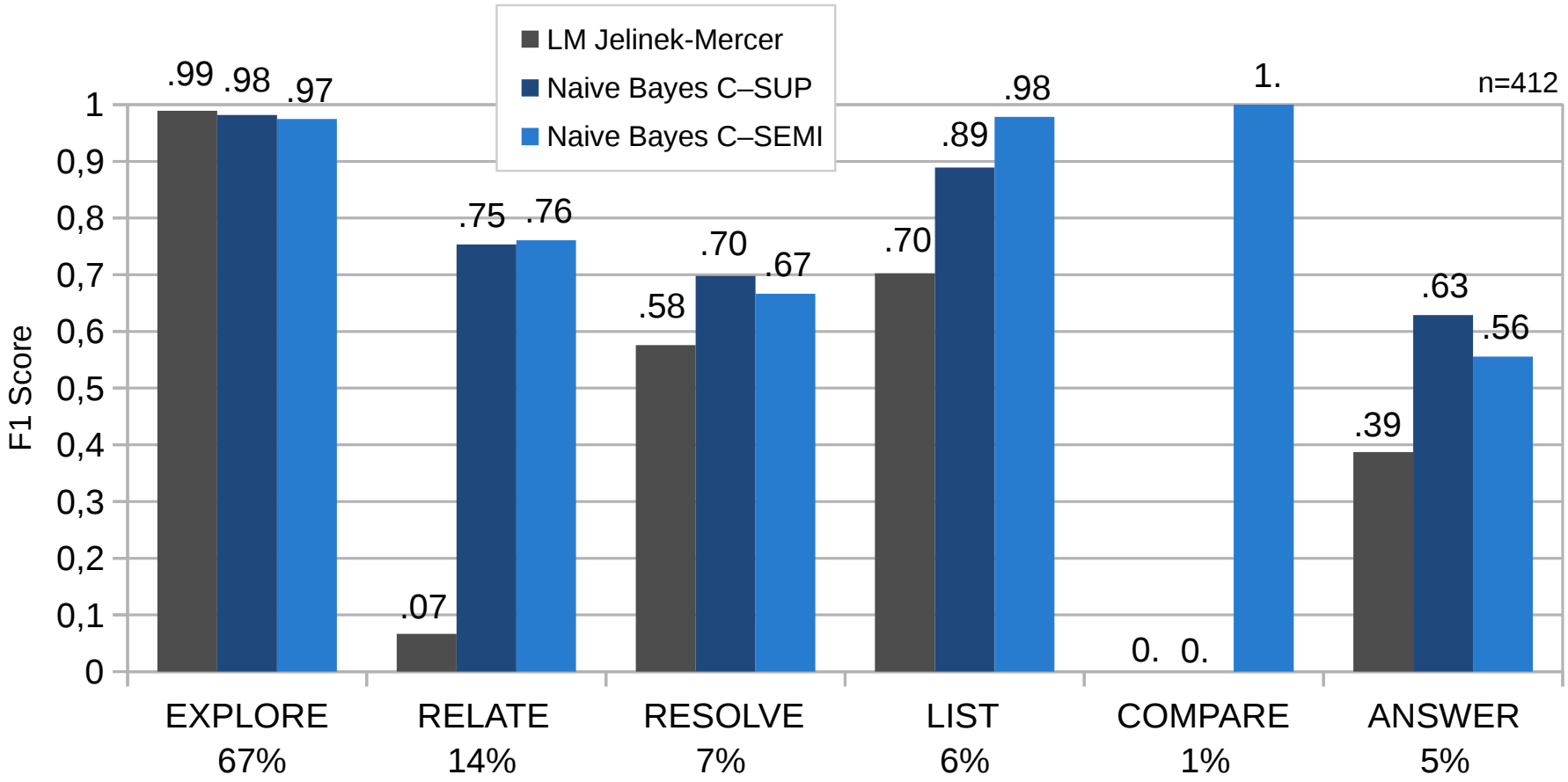
Evaluation Results: Classifiers



- query = `white house employees`
- SEQraw = JJ NN NNS
- SEQpos = NNP NN+
- SEQtype = Organization PersonCareer+

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

Evaluation Results: Classes



Discriminative Query Patterns

intent class	SEQtype	SEQpos	example query	freq
EXPLORE	Person	NNP	barack obama	81.71%
	Company	NNP	microsoft	7.23%
RESOLVE	Person Position	NNP NNP	paris hilton model	3.30%
	Person City	NNP NNP	mike bloomberg new york	2.62%
RELATE	Person Person	NNP NNP	barack obama michelle obama	36.91%
	Person Company	NNP NNP	sebastian vettel ferrari	19.19%
LIST	Person NN+	NNP NN+	homer simpson quotes	10.15%
	NN+ of Person	NN+ of NNP	siblings of george bush	1.29%
COMPARE	VB Company to Company	VB NNP to NNP	compare google to microsoft	3.37%
	Product Product	NNP NNP	coke pepsi	2.25%
ANSWER	Person PersonAttributes	NNP NN	barack obama age	1.97%
	Person PersonRelation	NNP VB	bill gates married	0.95%

Summary

- **Search is a process** that spans multiple dimensions
- There is demand for **factual results** in Web search
- **Informational queries** can be sub-categorized
  - undirected, directed, advice
- We observe **six classes** and specific visualizations
  - explore, resolve, relate, list, compare, answer
- **Feature extraction** from keyword queries needs heuristics
- We **classify search intent** using probabilistic methods
  - supervised Naive Bayes Classifier achieves 90% F1
- We implement the process for the search engine GoOLAP
- Paper submitted to ACM WebDB at SIGMOD 2015 [Arn15]

Future Work

- Extend ontology matching
  - Detection of operator keywords
    `(top, list, download)`
  - Improve recognition of real-world concepts
    `(map, picture, biography)`
  - e.g. WordNet, Biperpedia
- Extend interactive relevance feedback
  - Evaluate implicit measures (mouse movement etc.)
  - Prevent overfitting using result diversification
- Evaluate against representative data set
  - e.g. AOL search leak

References

- [Arn15] S. Arnold, A. Löser, and T. Kilias. Predicting Factual Query Intent from the Long Tail. In *WebDB 2015 at SIGMOD*. ACM, 2015. [IN SUBMISSION].
- [Aul08] A. Aula and D. M. Russell. Complex and Exploratory Web Search. In *ISSS 2008*, Chapel Hill, 2008.
- [Bro02] A. Broder. A Taxonomy of Web Search. In *ACM SIGIR Forum*, volume 36, pages 3–10. ACM, 2002.
- [Gon11] C. Gonzalez-Caro and R. Baeza-Yates. A Multi-Faceted Approach to Query Intent Classification. In *String Processing and Information Retrieval*, 368–379. Springer, 2011.
- [Guo10] Q. Guo and E. Agichtein. Ready to buy or just browsing?: Detecting Web Searcher Goals from Interaction Data. In *SIGIR 2010*, pages 130–137. ACM, 2010.
- [Kat14] M. P. Kato, T. Yamamoto, H. Ohshima, and K. Tanaka. Investigating Users' Query Formulations for Cognitive Search Intents. In *SIGIR 2014*, 577–586. ACM Press, 2014.
- [Lös12] A. Löser, S. Arnold, and T. Fiehn. The GoOLAP Fact Retrieval Framework. In *Business Intelligence*, 84–97. Springer, 2012.
- [Mar06] G. Marchionini. Exploratory Search: From Finding to Understanding. *Communications of the ACM*, 49(4):41–46, 2006.
- [Ros04] D. E. Rose and D. Levinson. Understanding User Goals in Web Search. In *WWW 2004*, 13–19. ACM, 2004.