Towards Identifying the User's Query Intent in Web Search

Liliana Calderón-Benavides
Web Research Group
Universitat Pompeu Fabra
mcalderon@unab.edu.co

Joined work with Ricado Baeza-Yates & Cristina González-Caro

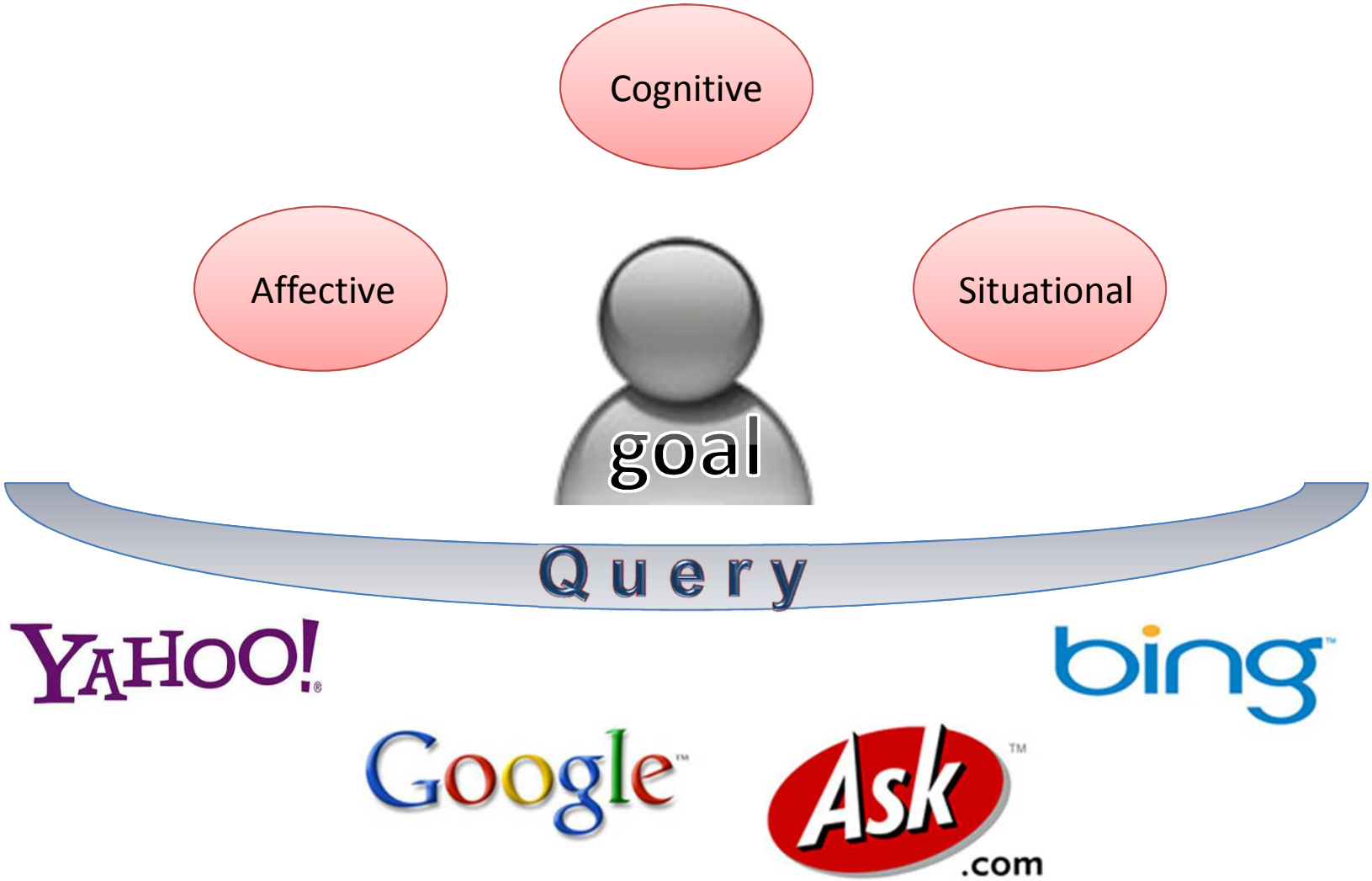
This work was partially suported by the Universidad Autónoma de Bucaramanga and the Ministry of Science and Innovation of Spain under grant numbers TIN2009-14560-C03-01 and TIN 2006-15536- C02-01.

Agenda

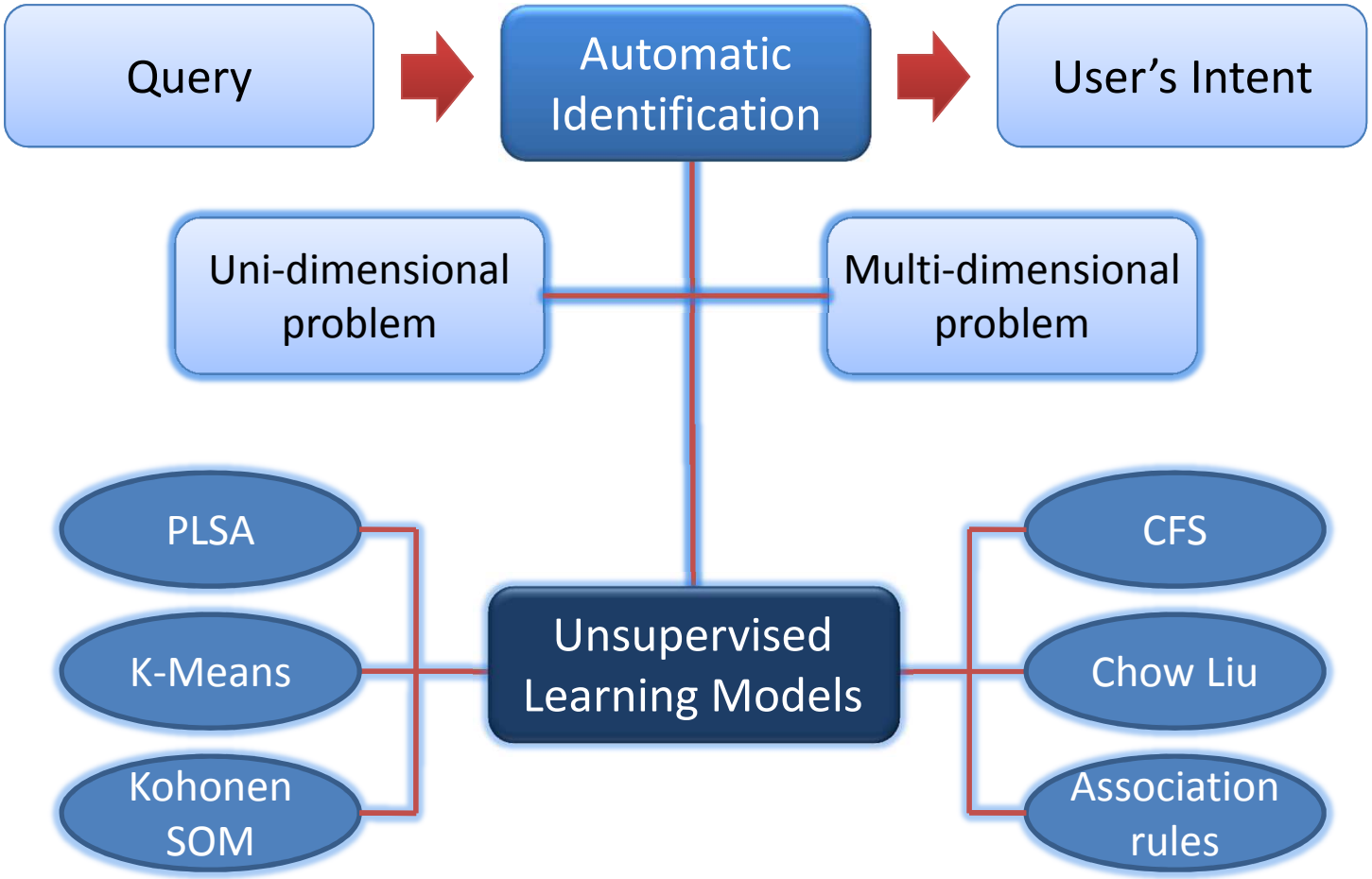
- Motivation
  - User's Intent
  - Objective of this work
  - How to determine the user's intent?
- A New Perspective of the User's Intent
- Multi-dimensional Perspective of the User's Intent
- Towards an On-line Identification
- Conclusions and Future Work

Motivation

User's Intent



Objective


Query
Automatic Identification
User's Intent
Uni-dimensional problem
Multi-dimensional problem
PLSA
K-Means
Kohonen SOM
Unsupervised Learning Models
CFS
Chow Liu
Association rules

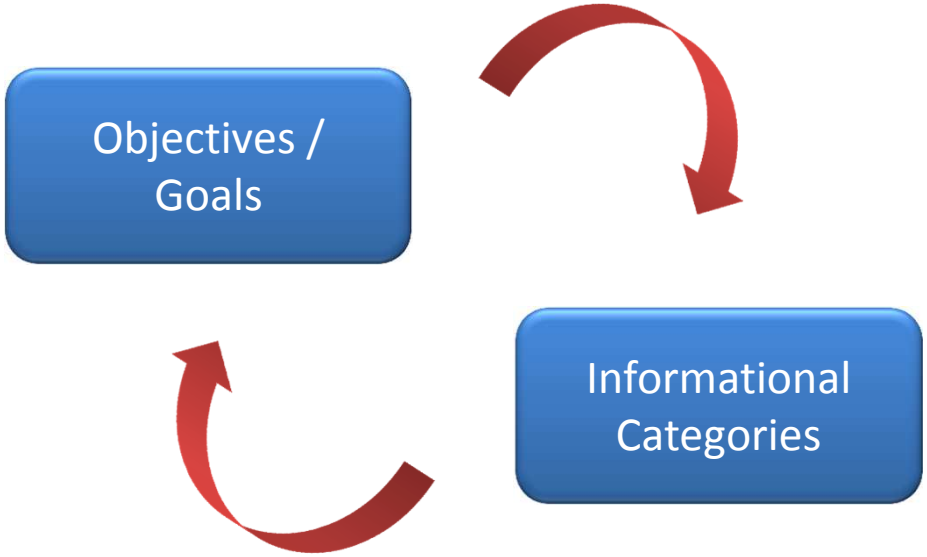
How to Determine the User's Intent?

Statistical Analysis	Silverstein et al., 1999	Jansen et. al., 2000	Spink et al., 2001
Manual Analysis	Broder, 2002 (400 queries)	Rose & Levinson, 2004 (1500 queries)	Rojas et al., 2010 TREC collection
Automatic	Kang & kim, 2003	Lee et al., 2005	Jansen et al., 2008

A New Perspective of the User's Intent

Objective

- Develop a model for the identification of the user's intent, using their past interaction with a Web search engine



A new characterization of Web user's intents

- Intents:
  - Informational
  - Not informational
  - Ambiguous

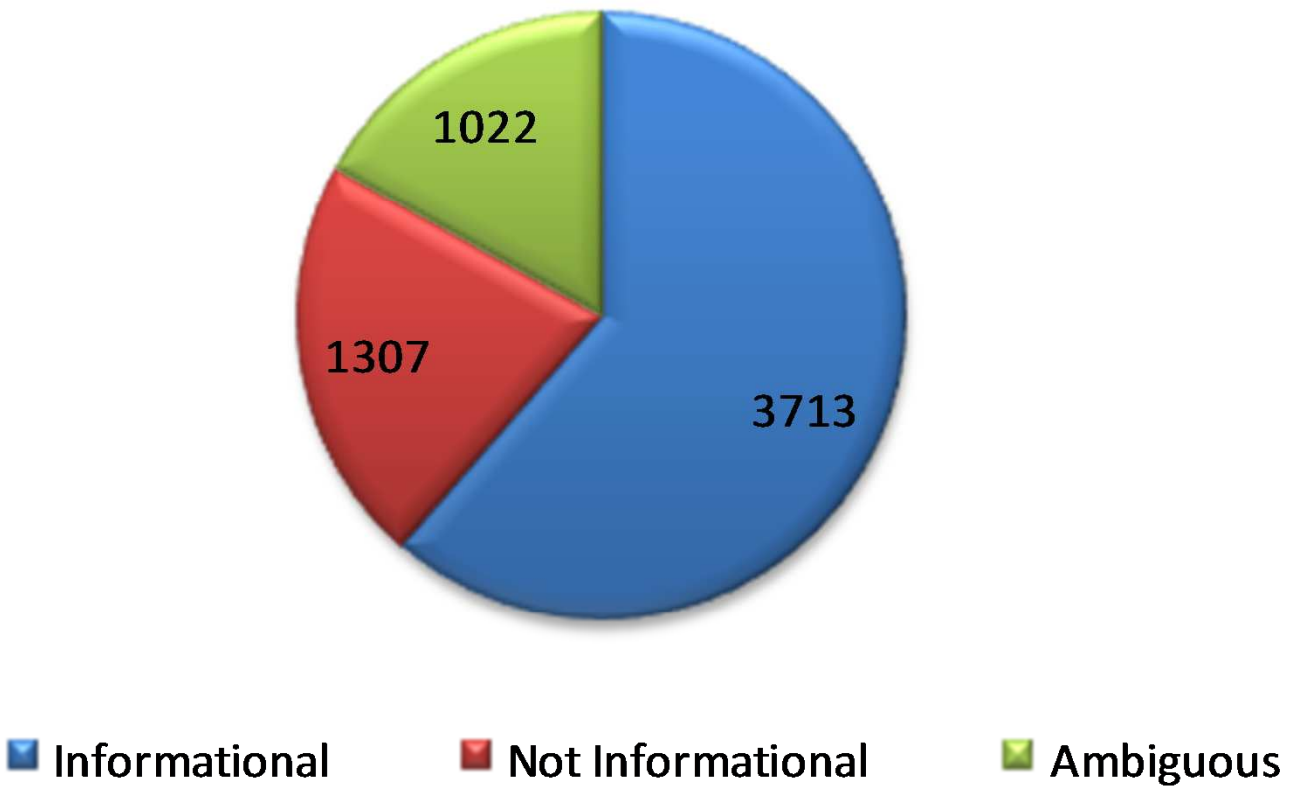
- Categories:
  - Open Directory Project, ODP
  - Adult, Various and Others

Semi-manual Classification

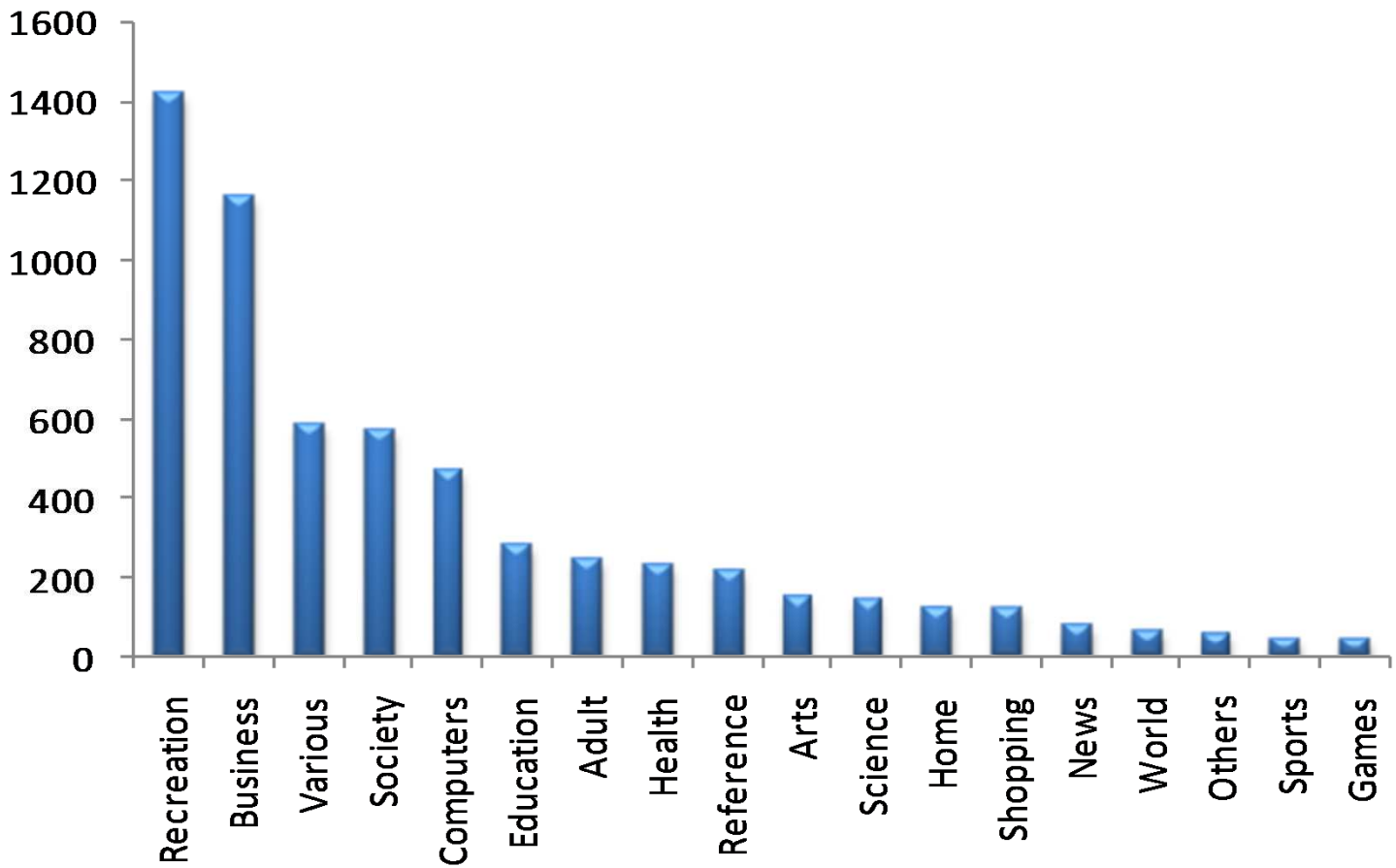
- Pre-processing...
  - Data Set:
    - More than 6,000 unique queries from todocl TODO CHILE EN INTERNET
  - Query representation
    - Words contained in the pages that users selected as answer to his/her query
  - Clustering of Queries
    - **K-Means model**
    - 600 clusters

UNIVERSITAT POMPEU FABRA

Semi-manual Classification: Goals

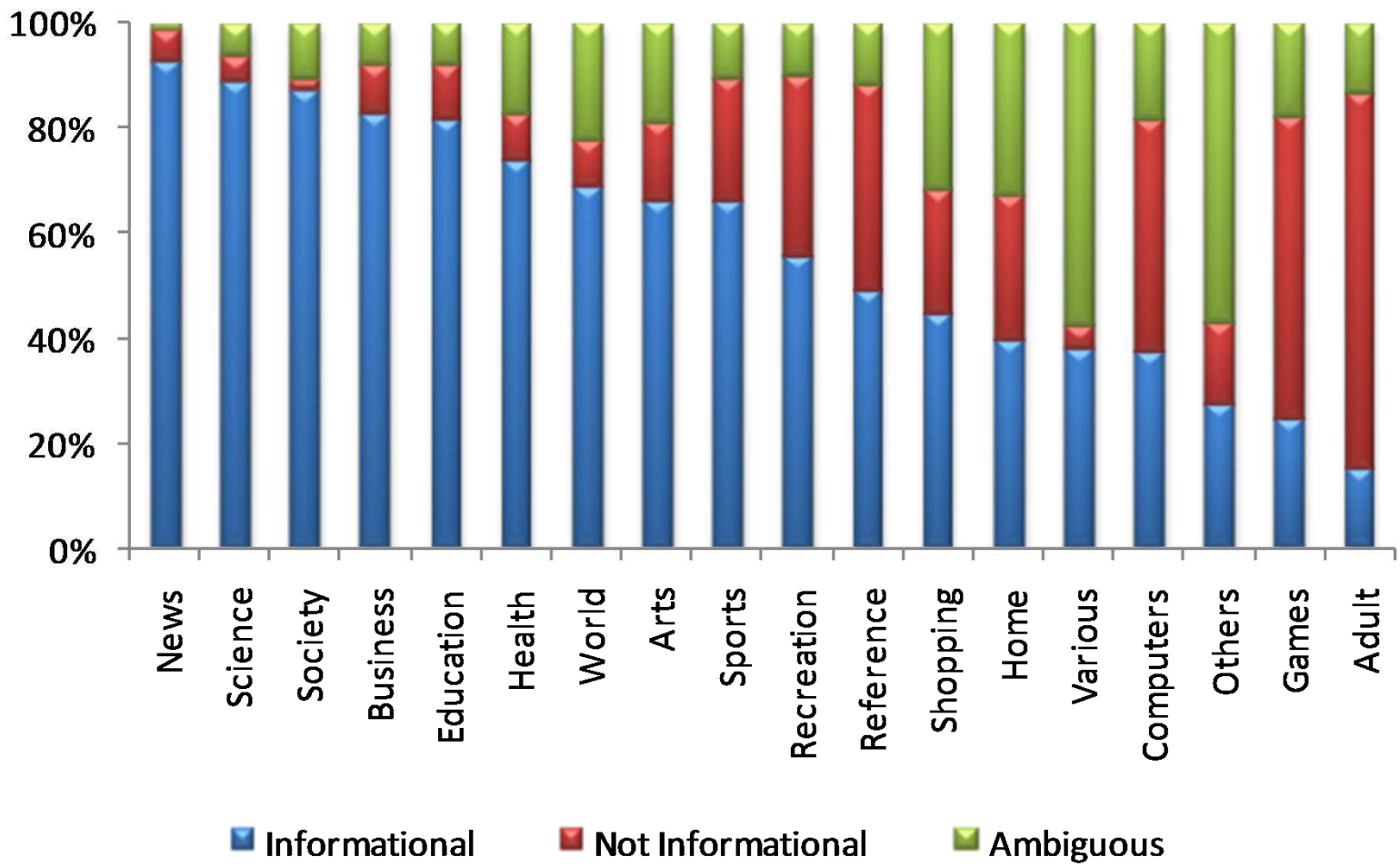


Semi-manual Classification: Categories



UNIVERSITAT POMPEU FABRA

Semi-manual Classification: Goals - Categories



UNIVERSITAT POMPEU FABRA

Automatic Classification of Queries

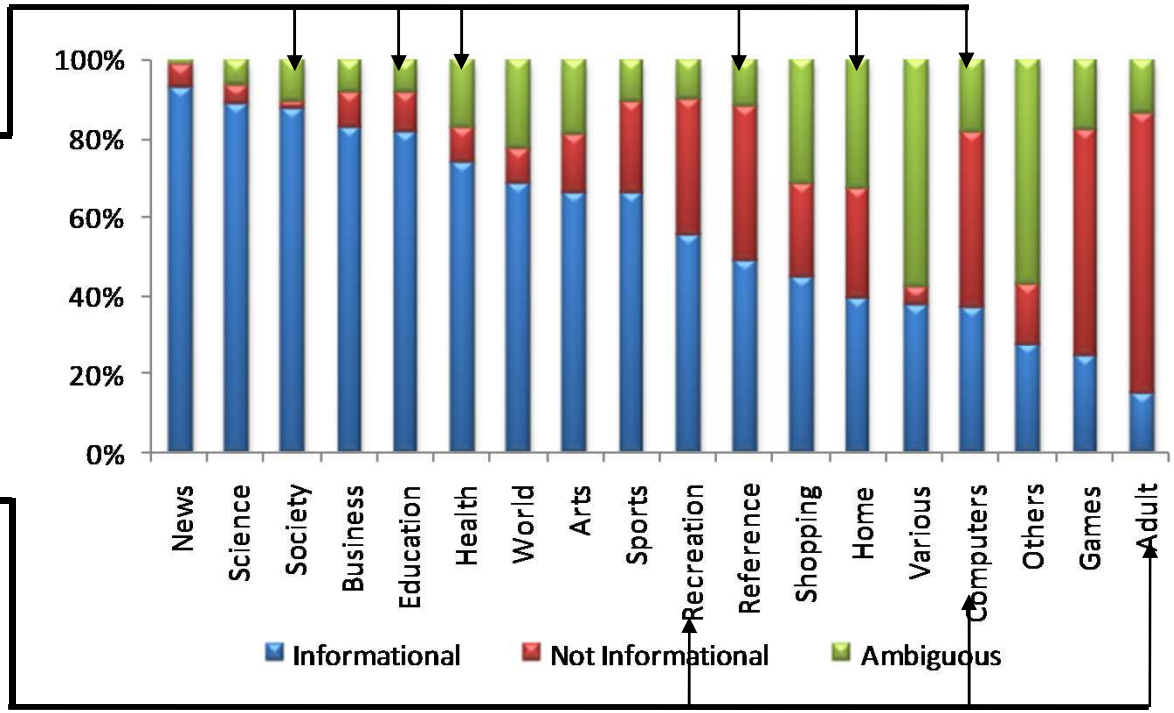
- **Probabilistic Latent Semantic Analysis (PLSA)**
  - Generative model of latent class
  - PLSA uses the Expectation Maximization algorithm in order to find the parameters of the model
  - PLSA models have been applied to recommendation problems and information retrieval, among others

Automatic Classification of Queries

- Clear identification of categories such as Adult, Recreation, Business, Reference, or Health
- Arts, Sports, Science, or Games?
- New categories appear: Law – Cars
- Relationships between query categories

Automatic vs. Manual Classification

- Informational: Approx. 76% of queries were related with: Society, Education, Health, Reference, Home, and Computers
- Not Informational: Approx. 70% of queries were related with: Recreation, Computers and Adult
- Ambiguous: Approx. 73% of queries were related with Adult or Recreation



Automatic Classification Results

Semi-manual Classification Results

UNIVERSITAT POMPEU FABRA

Contributions

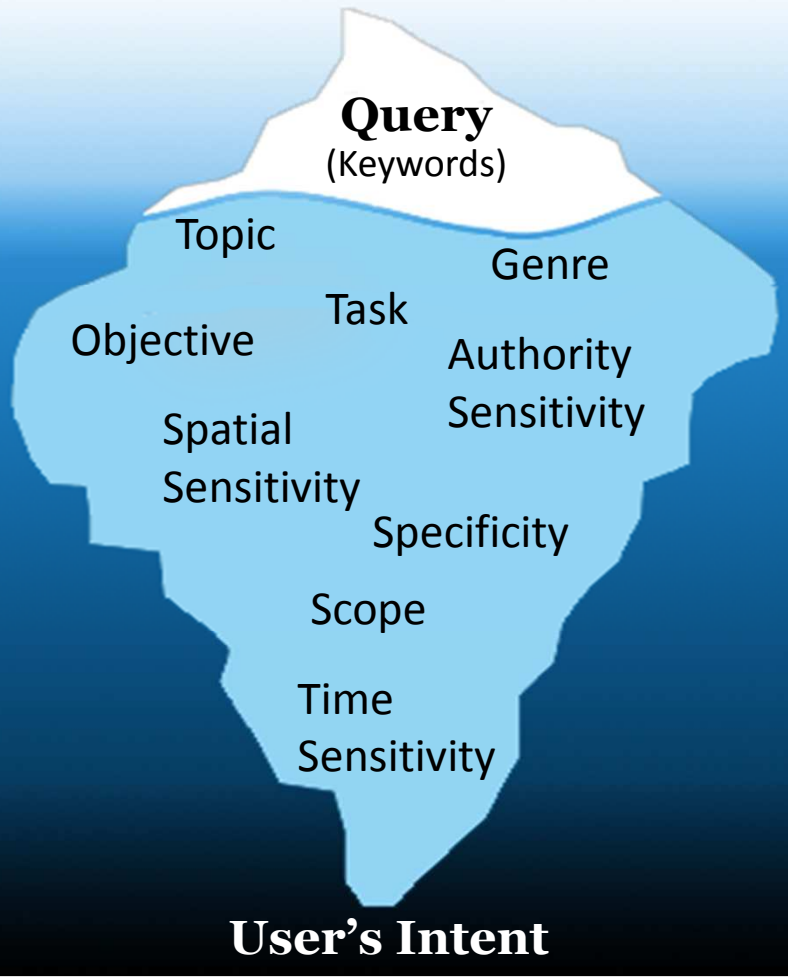
- A new automatic categorization of the user's intent
- The relationship between intents and informational categories
- The use of a large sample of real user queries

Multi-Dimensional Perspective of the User’s Intent

Objective

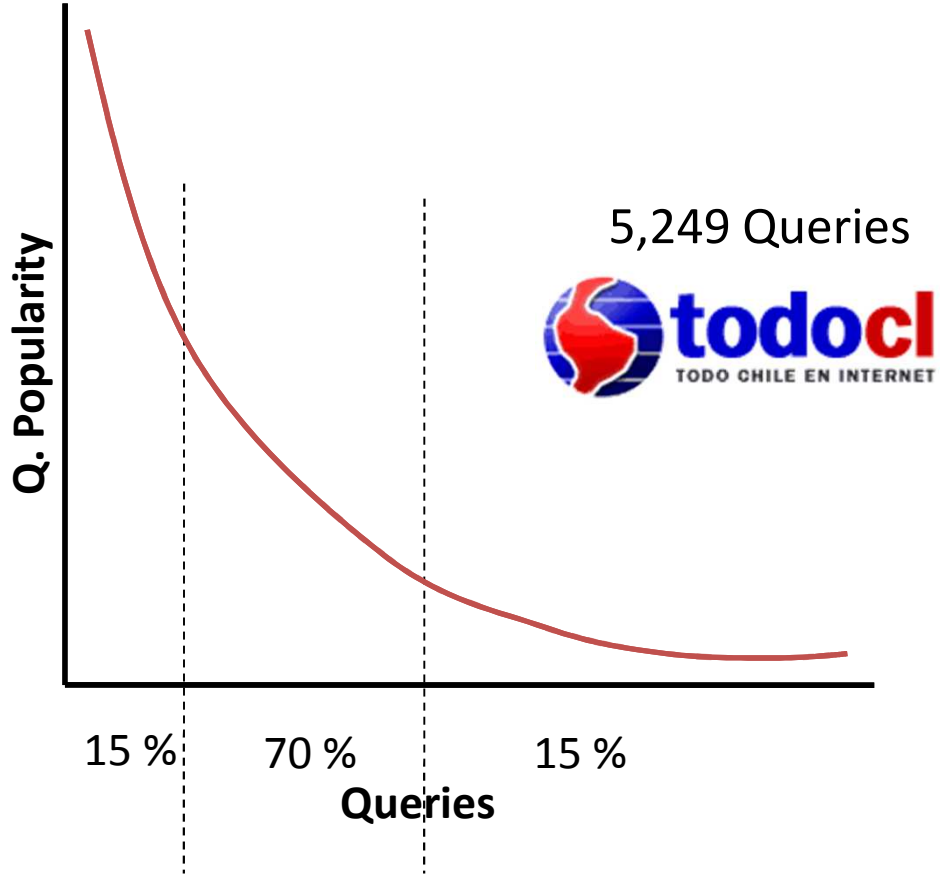
- Explore and analyze the user's intent from a multi-dimensional perspective
  - Propose a set of dimensions
  - Delineate the relationships and dependencies that exist among the dimensions involved in establishing the user's intent

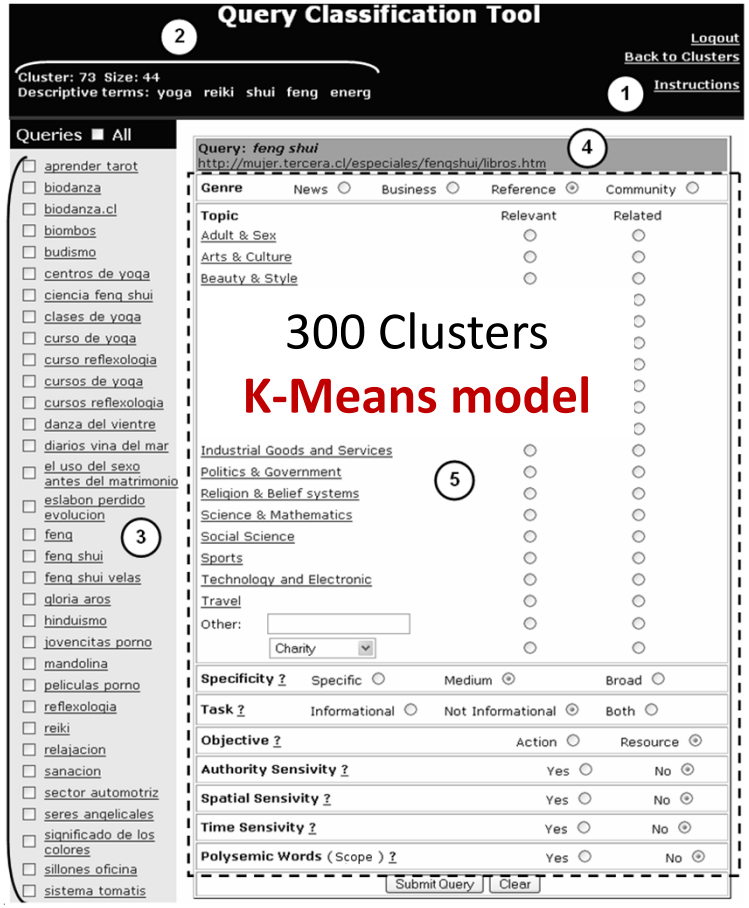
Dimensions of the User's Intent



- Based in previous work
- Focus in:
  - Completeness
  - Coverage
  - Simplicity

Semi-Manual Classification of Queries





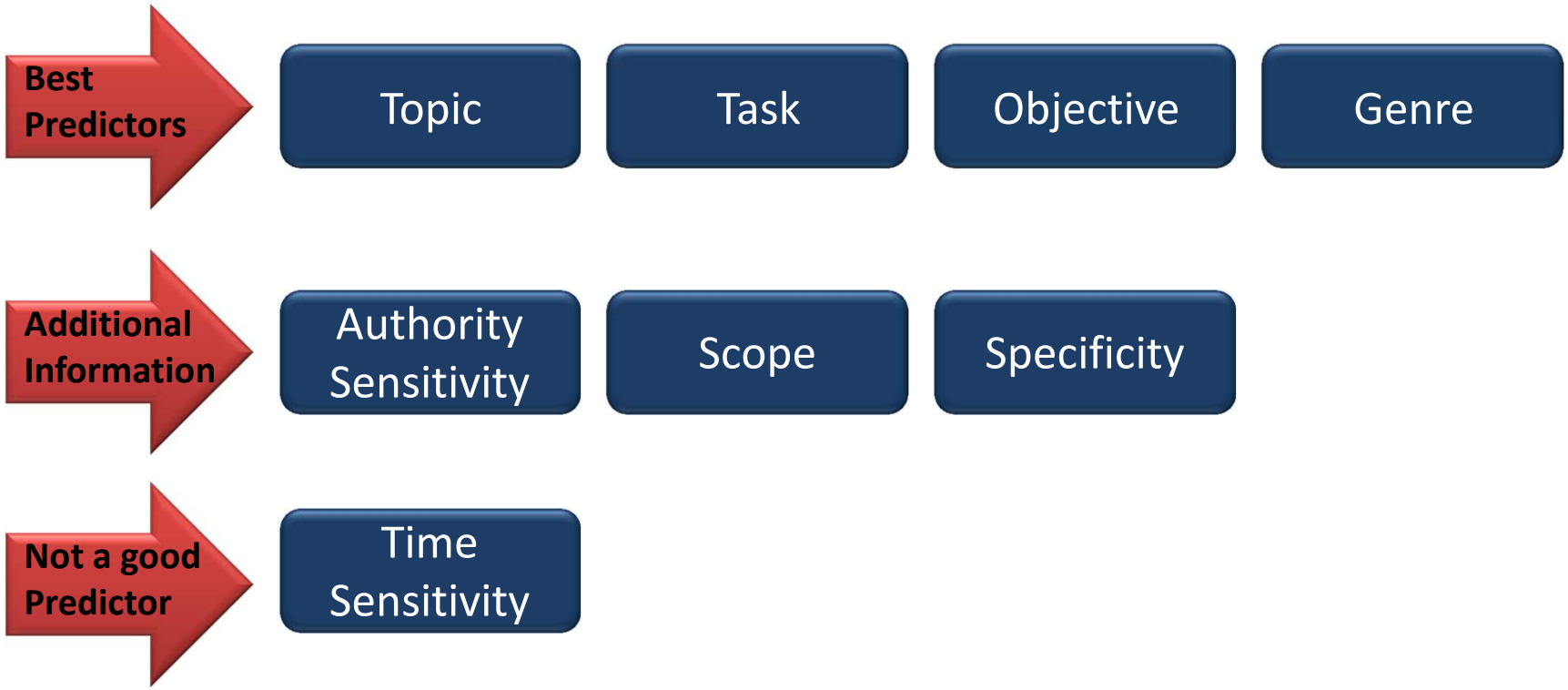
UNIVERSITAT POMPEU FABRA

Consistency of the Classification

Dimension	Overall Agreement	Kappa Coefficient	
Time S.	99.23	0.98	Almost perfect
Scope	96.74	0.93	
Objective	92.54	0.85	
Authority S.	84.32	0.69	Substantial
Topic	68.26	0.66	
Task	75.71	0.63	
Spatial S.	81.07	0.62	
Genre	65.00	0.53	Moderate
Specificity	55.44	0.33	Fair

Classification Capability of the Dimensions: Results

- **Correlation Based Feature Selection Algorithm, CFS**



Association Rules: Results

- **Apriori Algorithm**

	Support	Confidence
if task: *Informational* objective: *Resource*		
if genre: *Business*, topic: *Travel* specificity: *Medium* task: *Not-Informational*		
if topic: *Health*, genre: *Reference* objective: Action authority Sensitivity: *Yes*		

Barcelona map

Barcelona to Madrid

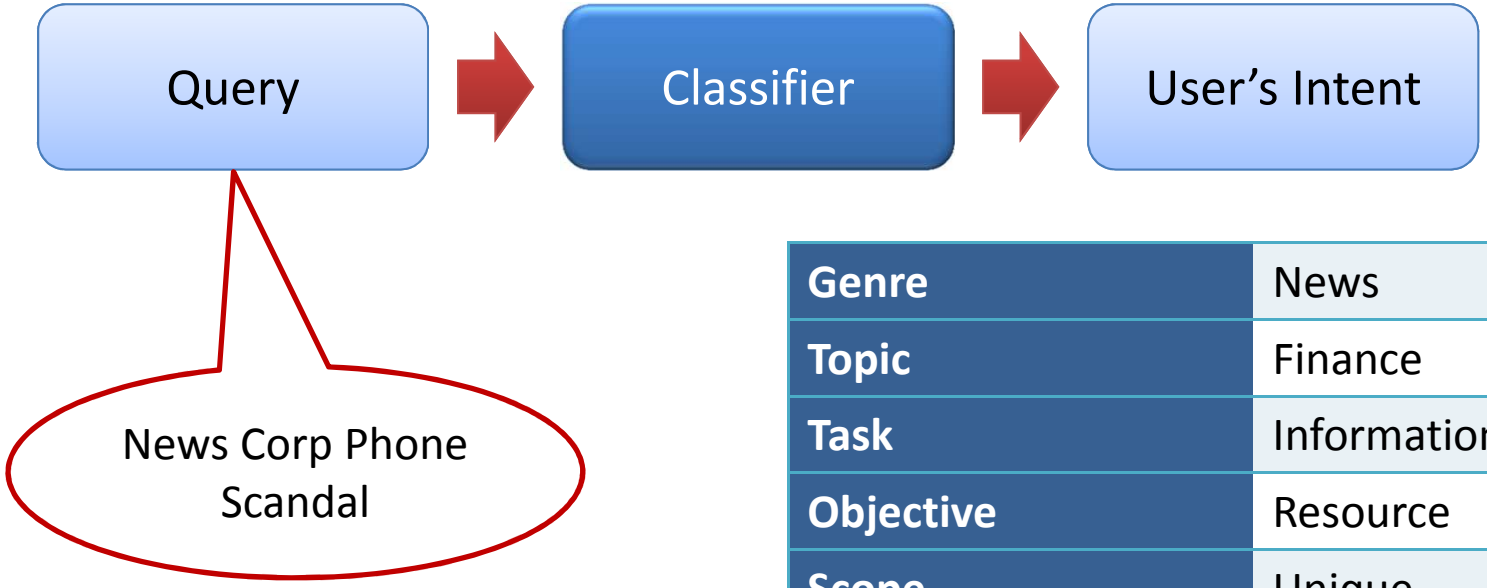
Aspirine

Contributions

- A new vision of the user's search intent as a problem of multi-dimensional nature
- Analyze nine dimensions involved in the user need, four of which are introduced by us
- To the light of the machine learning models applied, we point out the prediction capability of each dimension

Towards On-Line Identification of the User’s Intent

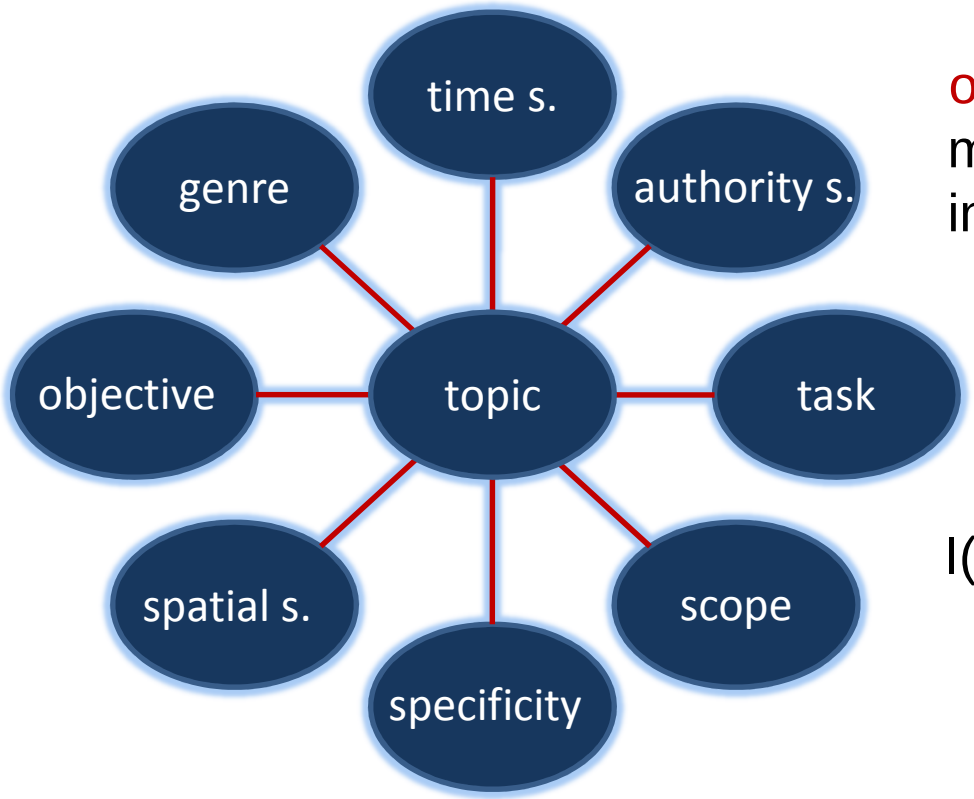
Objective



Genre	News
Topic	Finance
Task	Informational
Objective	Resource
Scope	Unique
Specificity	Specific
Authority Sensitivity	Yes
Spatial Sensitivity	No
Time Sensitivity	Yes

Challenge	Our Approach
User Intent dimension Inter-relationships	Tree Structured Distribution over user intent dimensions
Query Word – User Intent relationship	Conditional independent query words given user intent
New query words	Semantic relationships between words using the WordNet database

Tree Structured Distribution over user intent dimensions [ChowLiu, 1968]



- Find tree that maximizes the total mutual information between the user intent facets based on training data

$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

I(X;Y) = 0 iff X and Y are independent.

$$P(\vec{f}) = \frac{\prod_{(i,j)\in\Theta} \Theta_{ij}(f_i, f_j)}{\prod_i \Theta_i(f_i)}$$

UNIVERSITAT POMPEU FABRA

Conditional distribution of query words

- Each query word is independent of other query words given the facets (of user intent)

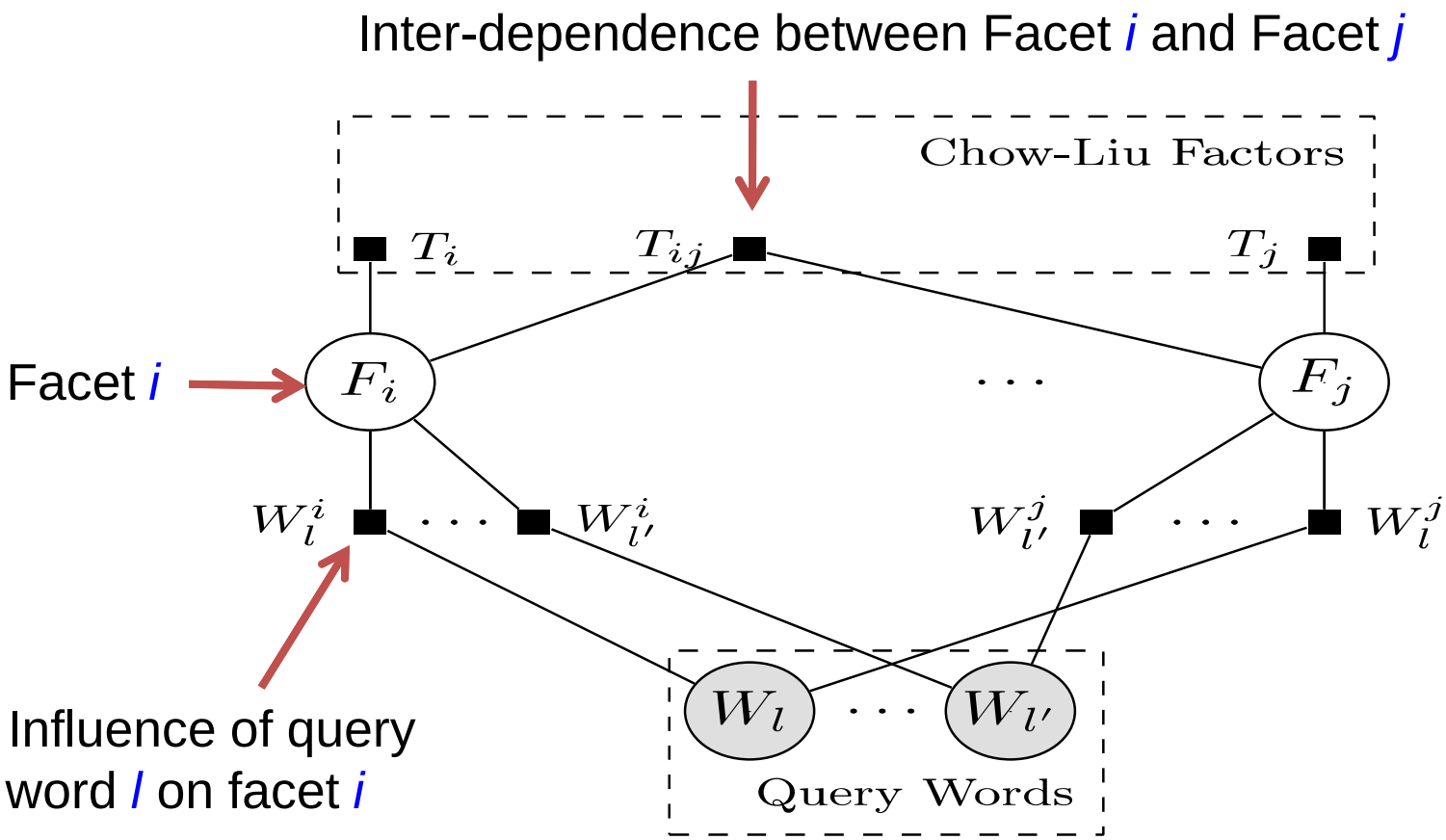
$$P(q|\vec{f};\Lambda) = \prod_{w \in q} P(|\vec{f};\Lambda)$$

- Learning (Maximum Likelihood Estimate)

Λ Query word – User Intent conditional distribution parameter

Θ Facet Inter-relationship distribution parameter

The FastQ Inference Algorithm



UNIVERSITAT POMPEU FABRA

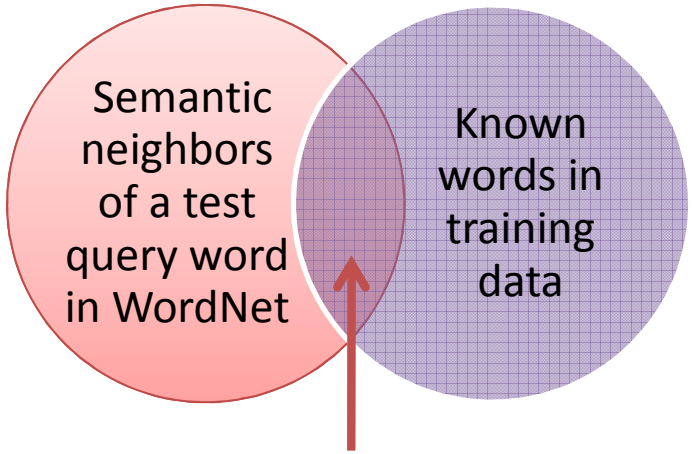
Extending FastQ using WordNet

- Inference based on semantically related words in the training data for a test query word not present in training database

WN search time

Maximum depth of search in WordNet L

Average number of neighbors in WordNet c_{WN}

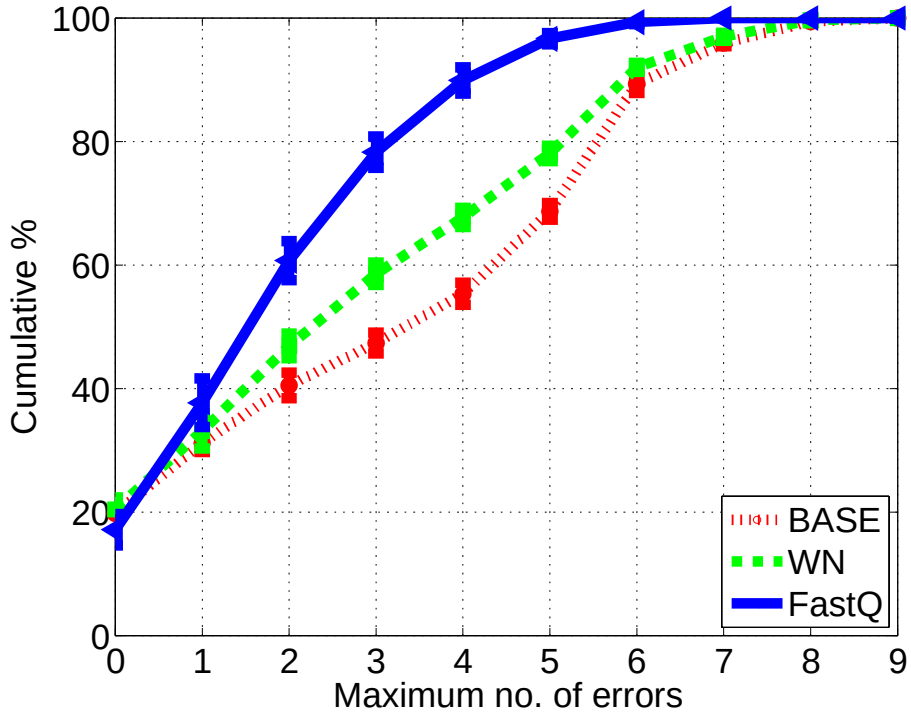


Words used for inference in FastQ in lieu of unknown test query word

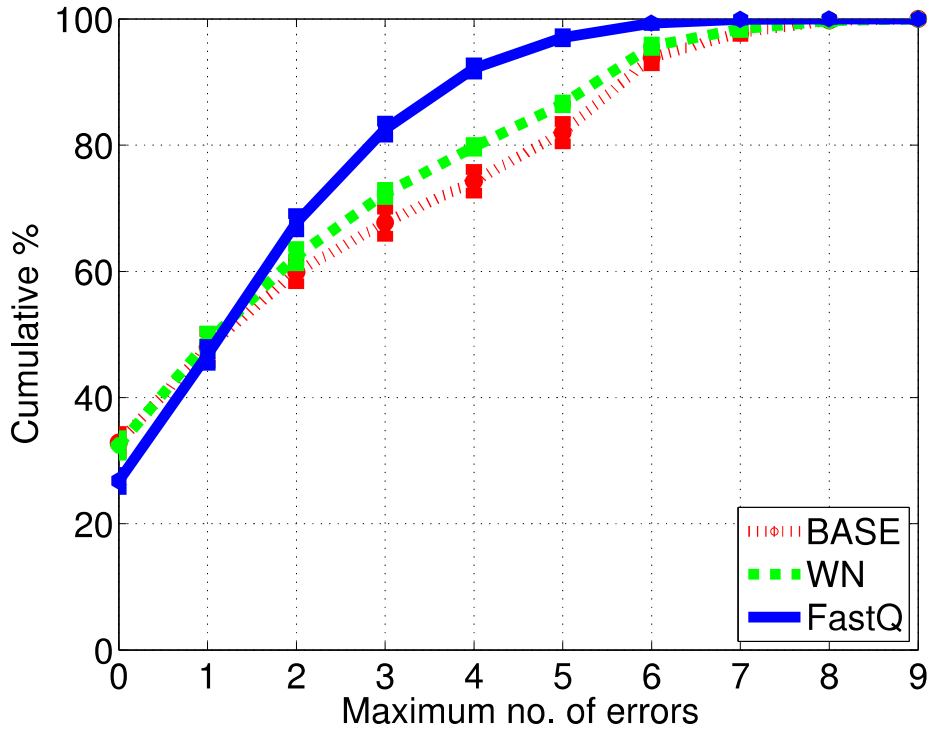
Evaluation

- Three models
  - BASE
    - Treats each facet independently
  - WN
    - Incorporates semantic information into BASE
    - Uses *synsets* to find semantically related words
  - FastQ
    - Incorporates facet inter-relationships into WN

Results: Facet classification (Hamming error)



10 % training data



50 % training data

Accuracy of Individual Facets

% Tr. data						
Model	BAS					FASTQ
author s.	50.(					**81.70**
genre	43.1					**65.02**
objective	**86.(**					85.59
scope	46.(					**97.28**
spatial s.	52.8					**67.98**
specificity	77.2					**78.14**
task	**75.2**					76.04
time s.	44.(					**98.43**
topic	**27.57**	20.90	29.07	54.44	**56.00**	32.90

Dimension	Overall Agreement
Authority S.	84.32
Genre	65.00
Objective	92.54
Scope	96.74
Spatial S.	81.07
Specificity	55.44
Task	75.71
Time S.	99.23
Topic	68.26

UNIVERSITAT POMPEU FABRA

Single Facet Identification (Task)

Method	Size	Accuracy
Topic/homepage	200	91%
FastQ (task)	**5,249**	**76%**
Rule-based hierarchy	1,500,000	74%
SVM	**6,042**	**63%**
Query-log based	50	54%

Task	True	Pred.	P	R	F1
Informational	1,382	1,584	0.79	0.91	0.85
Not Informational	645	635	0.73	0.72	0.72
Ambiguous	212	20	0.89	0.08	0.15

- FastQ performs better than other methods
- Average classification time 2-3 ms, parallelizable

Contributions

- We cast the problem of identifying user intent from Web search queries as an inference problem, and present a tree structured graphical model for modeling the user intent based on just the query words

- We incorporate WordNet, a lexical database, to idenfify new words present in the search query which are absent in the existing training vocabulary

- We propose a generic framework which can be extended by using other dimensions or query features, and thus enabling it use in other settings

Conclusions & Future Work

Conclusions

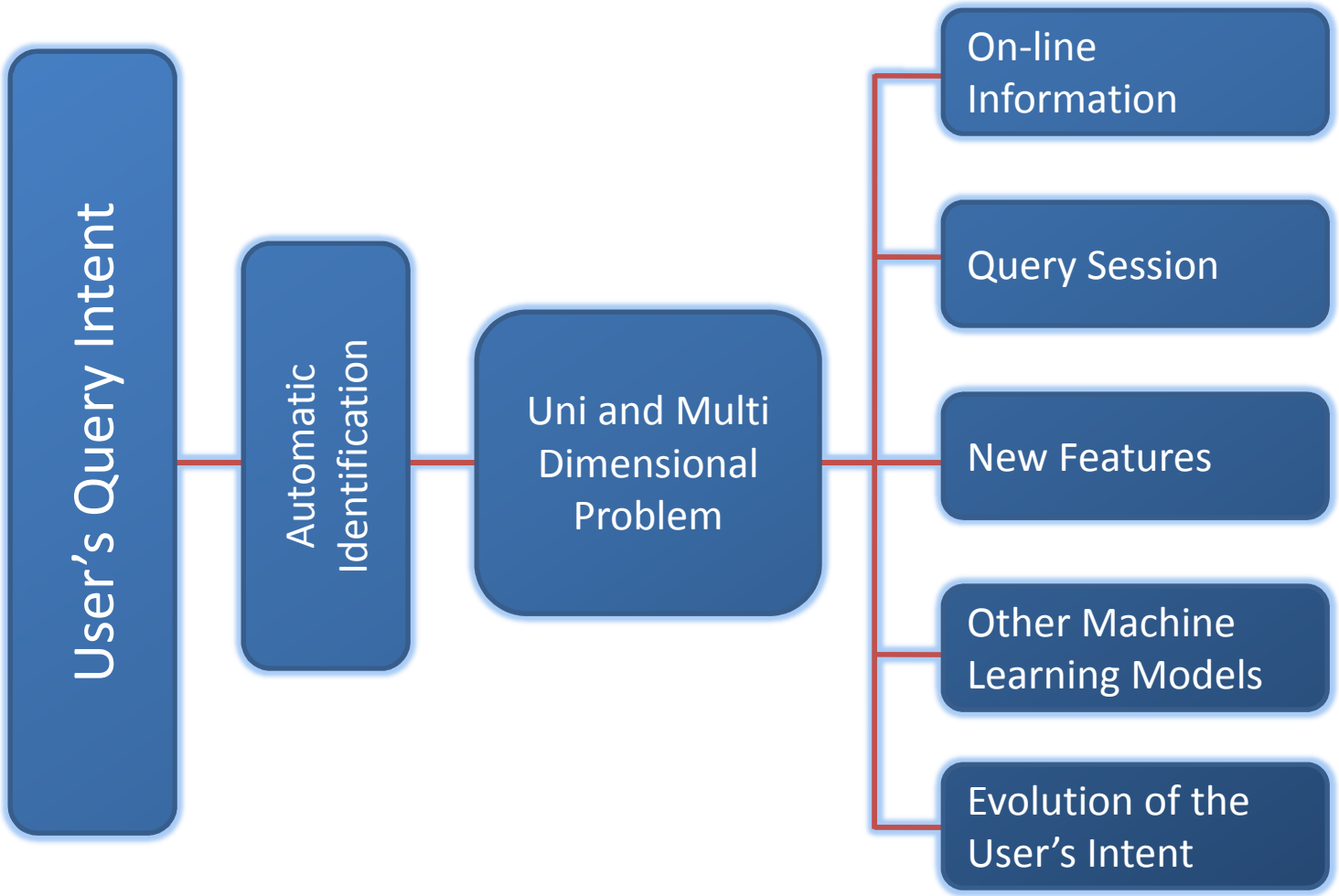
- We have applied several *Unsupervised Learning* models for the automatic identification of the user's intent
- From the applied models we proved that it is possible to identify the user's intent in an automatic way
- Unsupervised Learning models offer the possibility to discover the user's intent without any additional information, avoiding the cost of having labeled data
- The natural way to group the data with unsupervised learning techniques allows to find new patterns reflecting a closser view of the user's intent

Overall Comparison

	Supervised			Semi-supervised			Unsupervised		
	Prec.	Recall	F1	Prec.	Recall	F1	Prec.	Recall	F1
Inf.	0.64	0.98	0.78	**0.79**	0.77	0.78	0.62	0.72	0.68
Not Inf.	**0.84**	0.22	0.35	0.55	0.60	0.58	0.34	0.42	0.38
Amb.	**0.48**	0.02	0.04	0.44	0.42	0.46	0.47	0.27	0.35

- **Supervised:** Support Vector Machines
- **Semi-Supervised:** Semi-supervised Linear SVM.
- **Unsupervised:** Probabilistic Latent Semantic Analysis

Paths for Future Work



Queries?

Liliana Calderón-Benavides
Department of Communication and Information Technologies
Universitat Pompeu Fabra
liliana.calderon@upf.edu