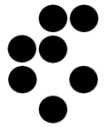
User Modeling Combining Access Logs, Page Content and Semantics

Blaž Fortuna
Dunja Mladenić
Marko Grobelnik

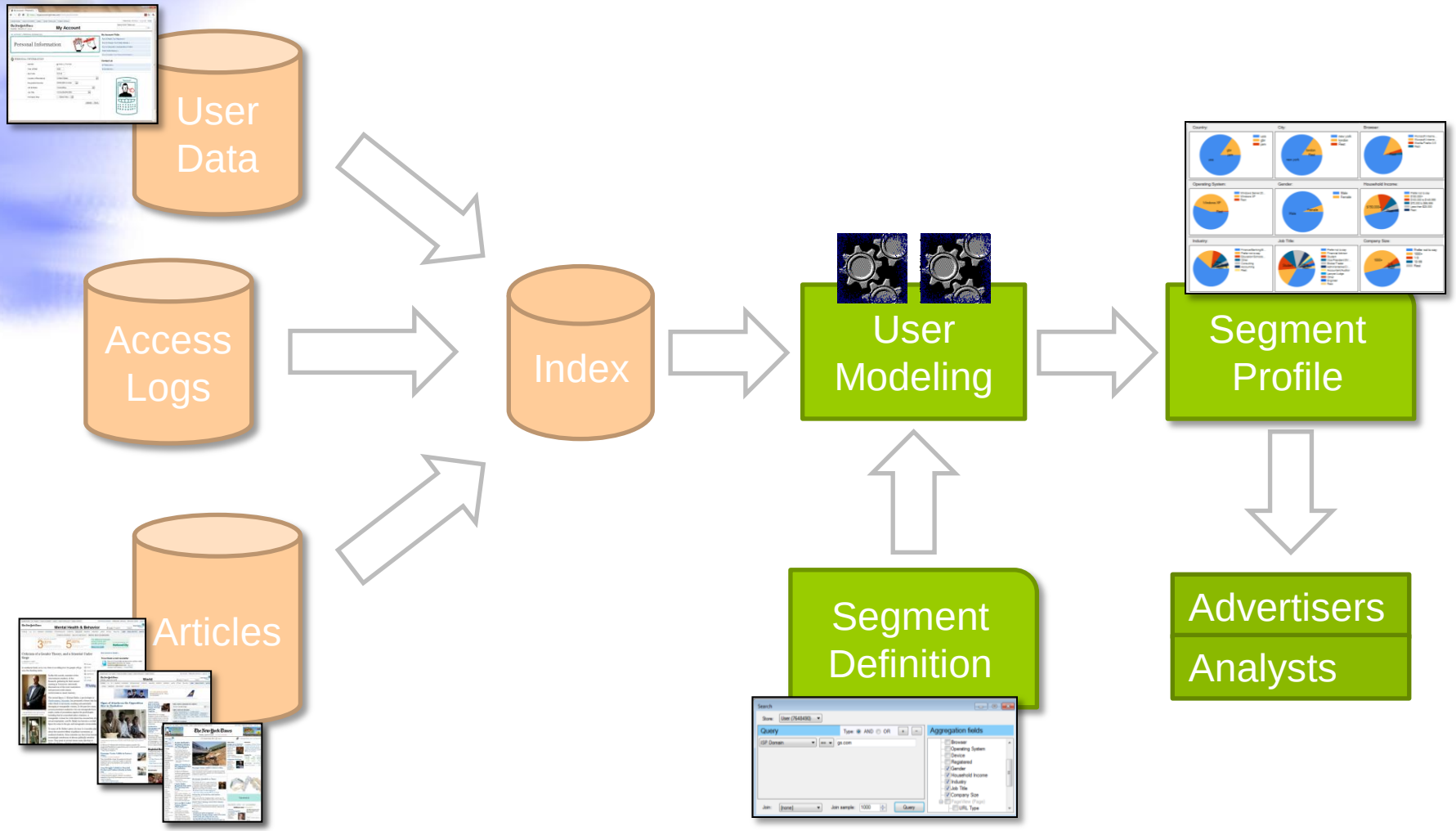
Artificial Intelligence Laboratory
Jožef Stefan Institute



Outline

- System Overview
- Data sources
- Defining segments
- User modeling
- Experiments
- Conclusions

System Overview



Access Logs

- User interactions with the website
- Each page-view described with:
  - **User ID**
  - **Date and Time**
  - **Location** (from IP address)
  - **Requested page**
  - **Referring page**
  - **Search query** (from Referring page)
  - **Browser, Operating System, Device** (from User agent)
- Users tracked using cookies
  - Tag with unique ID at the first visit

Example

User ID cookie: `1234567890`

IP: `123.123.123.123` (Beijing, China)

Requested URL:
http://www.nytimes.com/2009/08/23/weekinreview/23baker.html

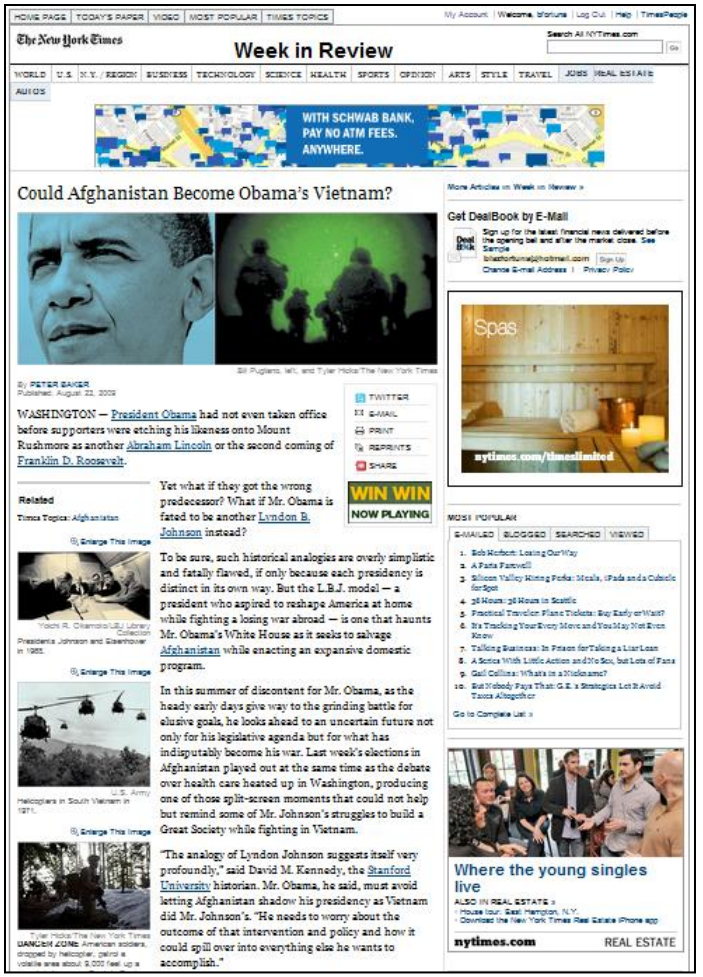
Referring URL:
http://query.nytimes.com/search/sitesearch?query=obama

Date and time: `2009-08-25 08:12:34`

User agent: `Mozilla/5.0 (Windows; U; Windows NT 5.1; en) AppleWebKit/526.9 (KHTML, like Gecko) Version/4.0dp1 Safari/526.8` (Safari, Windows, PC)

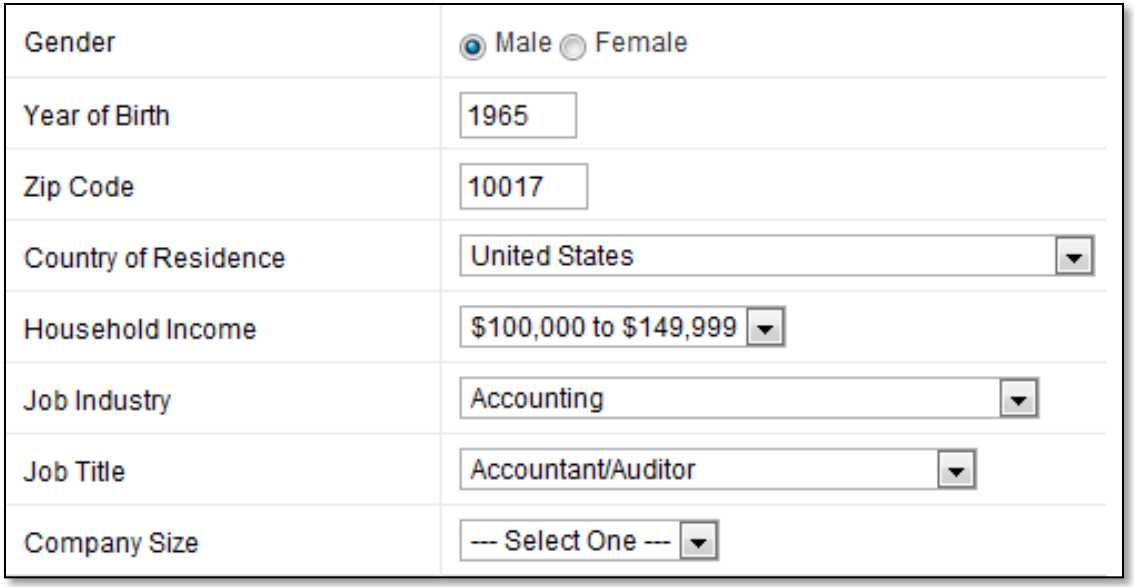
Articles

- Content and Semantics about requested pages
- Each page described with:
  - **Content**
  - **Annotations**
    - **Named Entities** (e.g. Obama, Mount Rushmore, Afghanistan, Vietnam)
    - **Topics** (e.g. politics, opinion, sports)
  - **Content meta-data** (e.g. author, publish date, editorial desk)
  - **Page meta-data** (e.g. article, home-page, section-front)



User Data

- Provided only for registered users
  - ~20% unique users in our case
  - Can generalize to all using machine learning
- Each registered users described with:
  - **Gender**
  - **Year of birth**
  - **Household income**
- Noisy



User Segment

- User segment:

> Subset of website visitors sharing some common characteristics

- Example:
  - [Gender = Male]
  - [Age ≥ 40]
  - [Referring domain = facebook.com]
  - [Requested page topic = Travel]
  - …

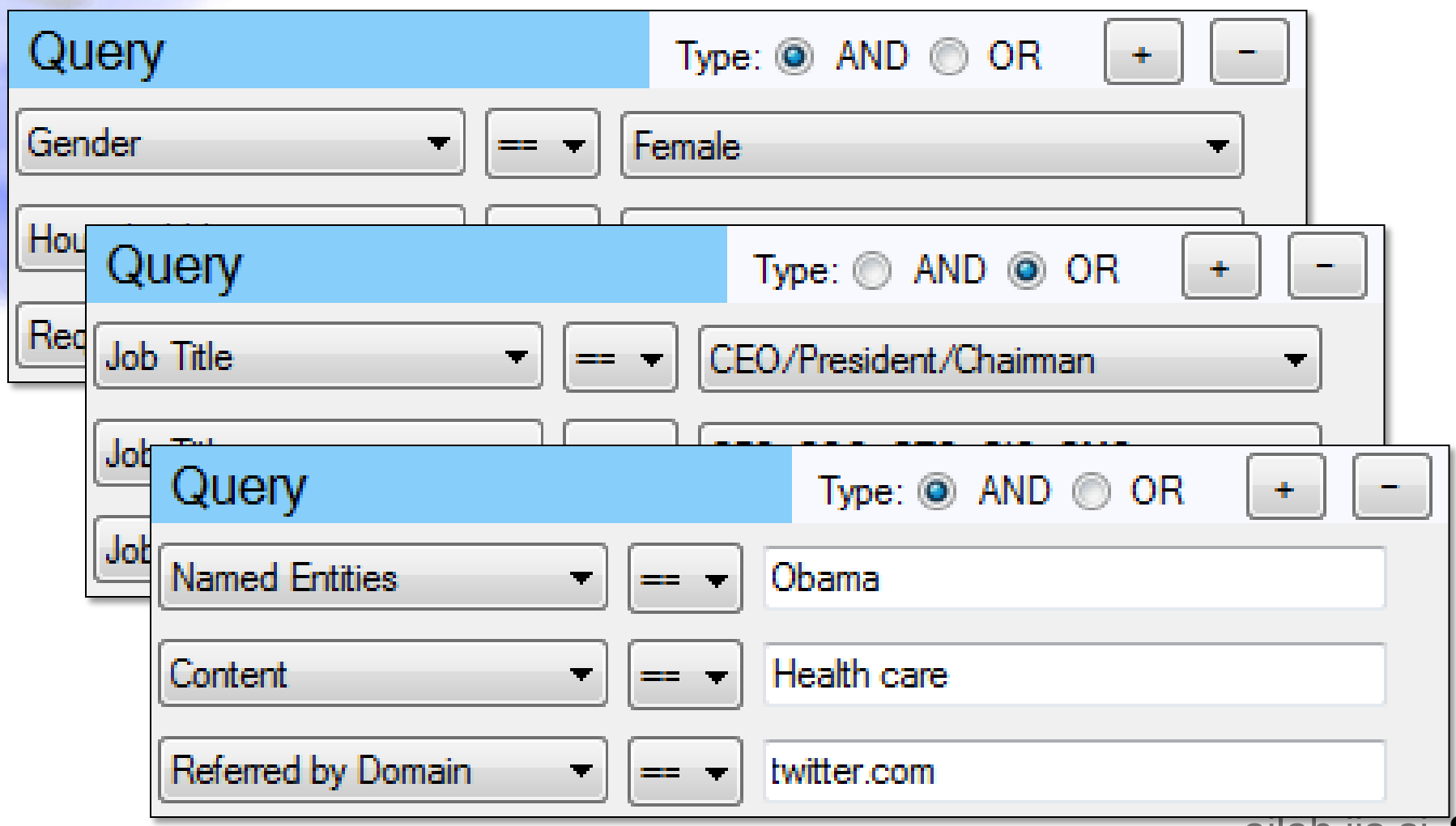
Defining Segments

- Must be simple enough so it can be used by domain experts
- Our solution
  - Index all users using inverted index
  - Segment definition equals faceted search query over users
  - Ad-hoc segment definitions

Indexed fields:

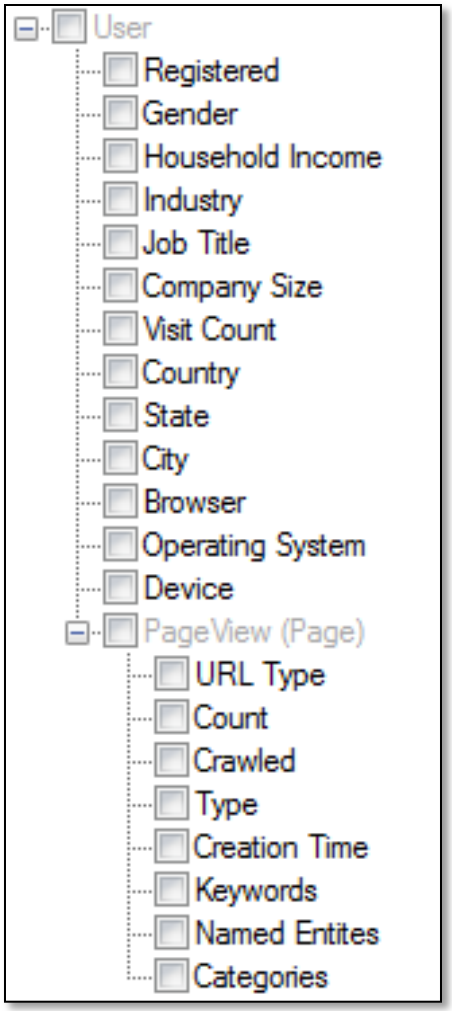
- Domain
- Sub-domain
- Page URL
- Page Meta Tags
- Page Title
- Page Content
- Named Entities
- Referring Search Term
- Referring Domain
- Referring URL
- Country (from IP)
- State (from IP)
- City (from IP)
- Date
- Day of the Week
- Hour of the day
- User Agent
- Income
- Age
- Gender

Example


Query
Type: AND OR
+
−
Gender
==
Female
Query
Type: AND OR
+
−
Job Title
==
CEO/President/Chairman
Query
Type: AND OR
+
−
Named Entities
==
Obama
Content
==
Health care
Referred by Domain
==
twitter.com

User Modeling

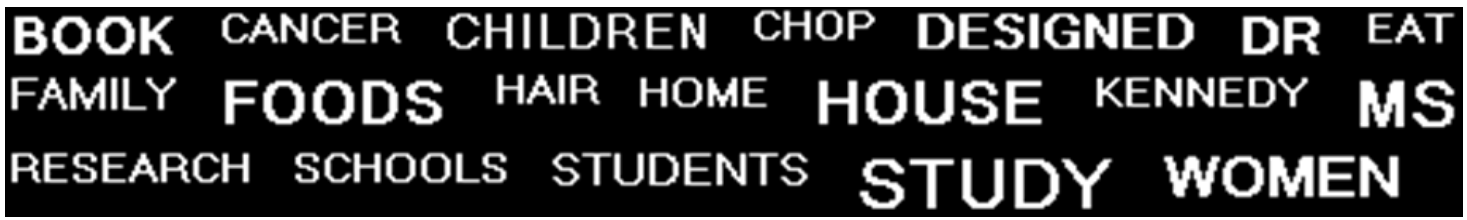
- Feature space
  - Extracted from subset of fields
  - Using vector space model
  - Vector elements for each field are normalized
- Training set
  - One visit = one vector
  - One user = a centroid of all his/her visits
  - Users from the segment form positive class
  - Sample of other users form negative class
- Classification algorithm
  - Support Vector Machine
  - Good dealing with high dimensional data
  - Linear kernel
  - Stochastic gradient descent
    - Good for sampling



Segment visualization

- Using SVM for feature selection
- Visualize a segment by displaying keywords significant for correct classification
- Useful information for the website editors

Gender = female
Income ≥ $100,000
Meta Data = Category Style



Experimental setting

- Real-world dataset from a major news publishing website
  - 5 million daily users, 1 million registered
- Tested prediction of three demographic dimensions:
  - Gender, Age, Income
- Three user groups based on the number of visits:
  - ≥2, ≥10, ≥50
- Evaluation:
  - Break Even Point (BEP)
  - 10-fold cross validation

Category	Size
Male	250,000
Female	250,000

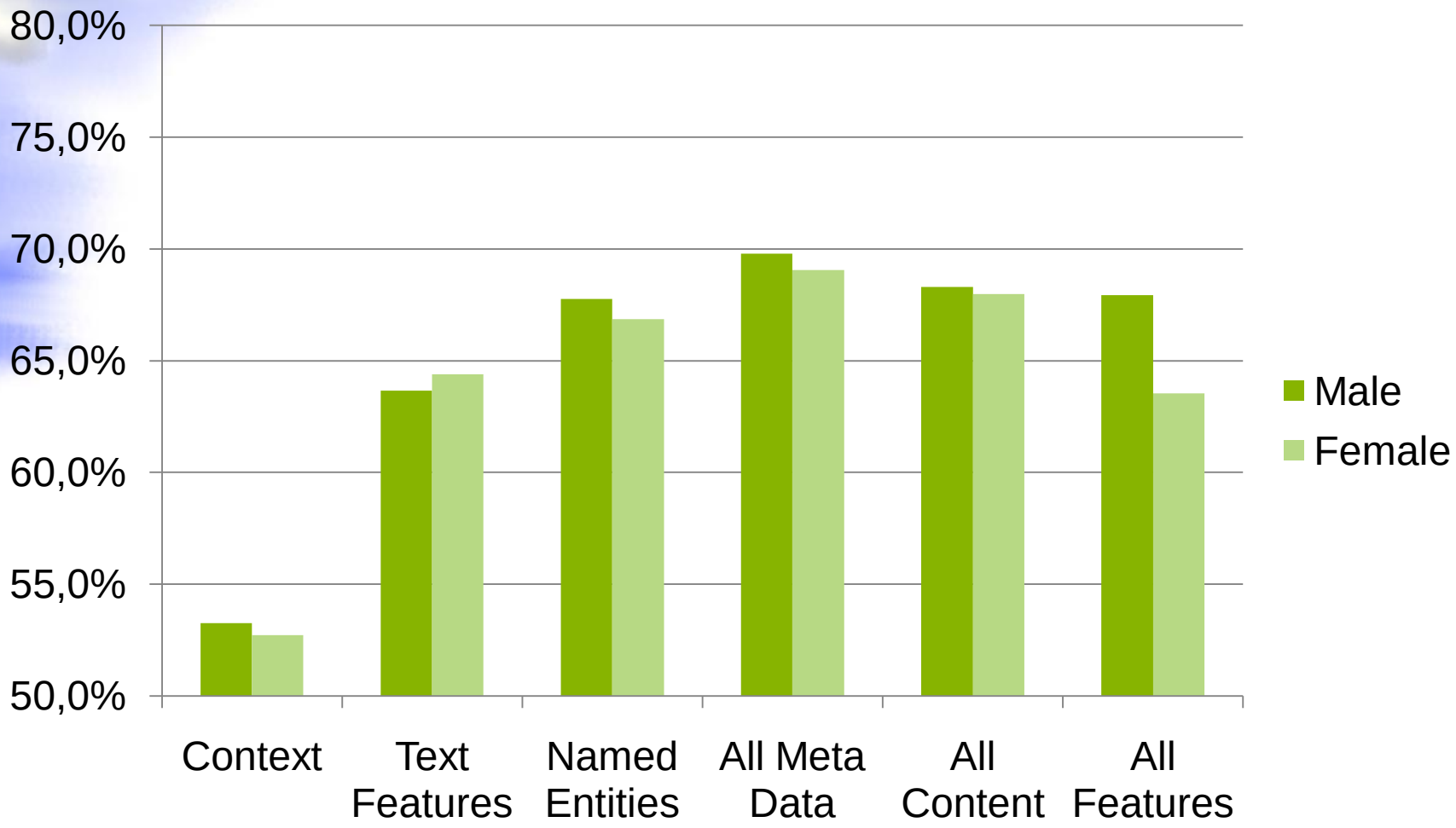
Category	Size
21-30	100,000
31-40	100,000
41-50	100,000
51-60	100,000
61-80	100,000

Category	Size
0-24k	50,000
25k-49k	50,000
50k-74k	50,000
75k-99k	50,000
100k-149k	50,000
150k-254k	50,000

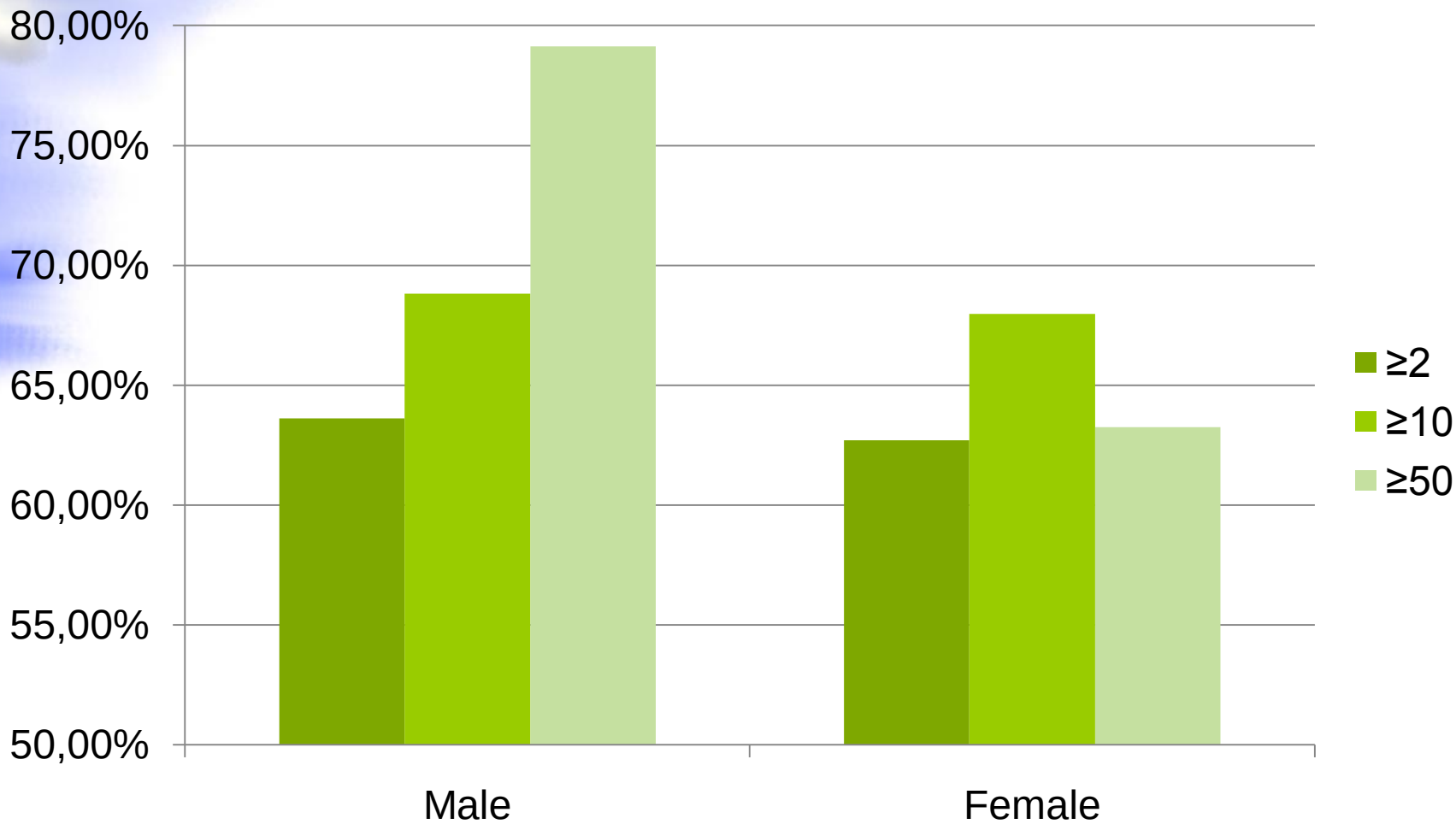
Combining Features

- **Context** – features that can be obtained from access logs, such as time, referring page, location and device.
- **Content features**:
  - **Text Features** – keywords extracted from the articles
  - **Named Entities** – automatically extracted named entities
  - **All Metadata** –assigned to the article by the authors and editors
    - byline; topics; main keywords; people, organization and countries mentioned in the article; publish date.
- **All Content** – combination of text features, named entities and metadata features.
- **All Features** – combination of all above features.

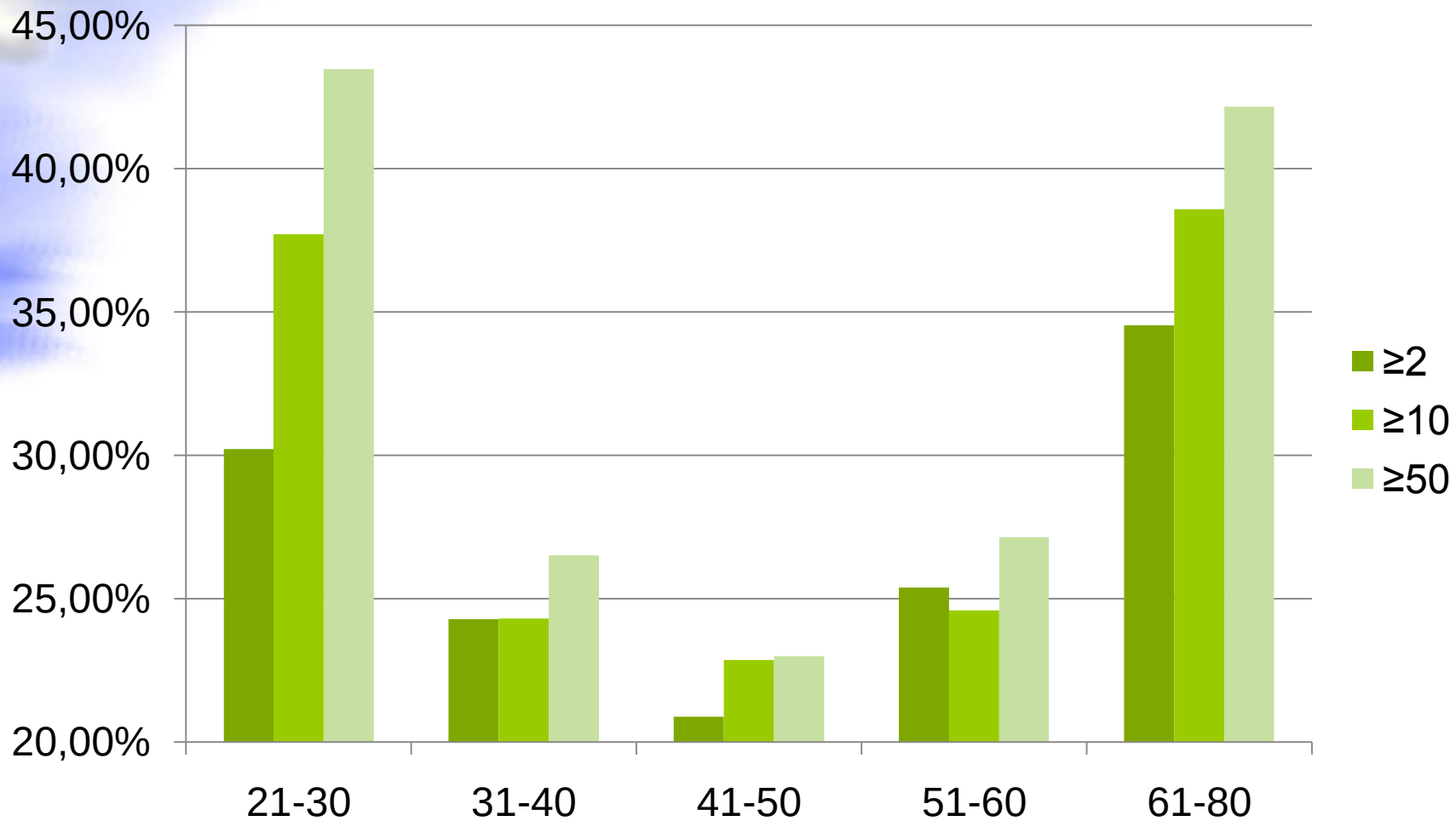
Gender (≥10 visits)



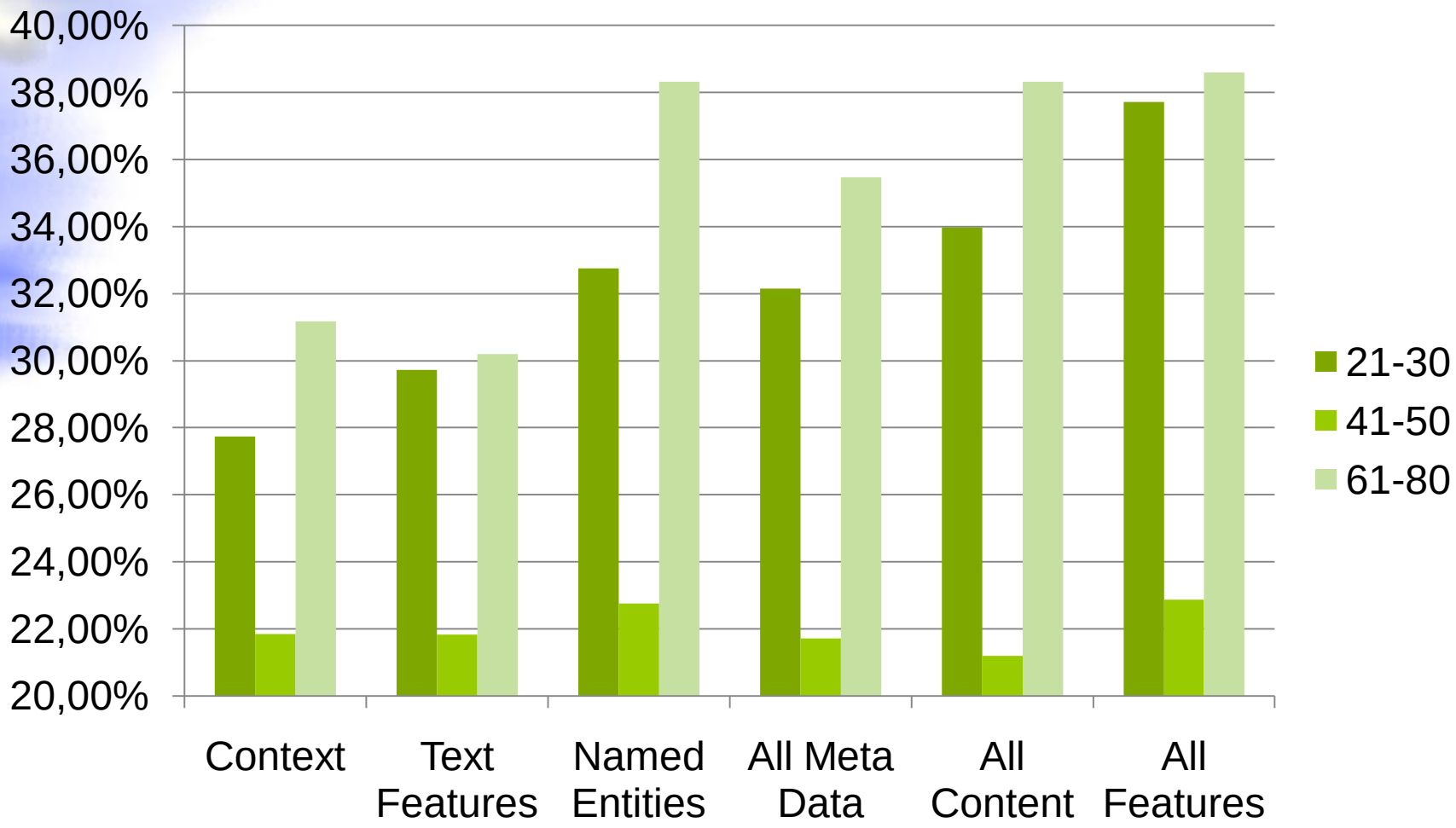
Gender



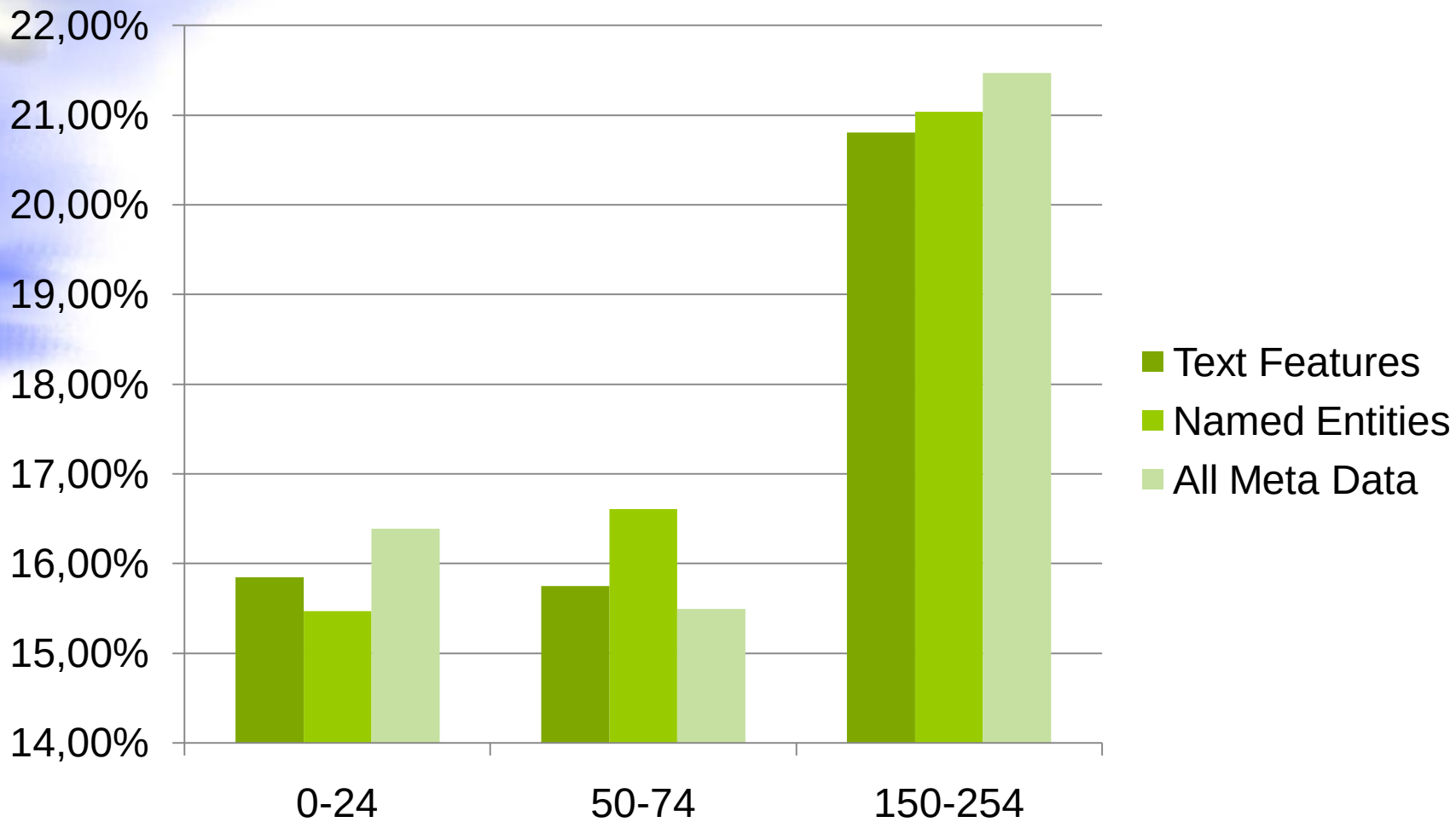
Age (all features)



Age (≥10 visits)



Income (≥10 visits)



Conclusions

- Modeling user segments
  - User friendly way to define complex segments
- Combining several data sources
  - Usage logs, content and semantics
- Tomorrow (related work):

  SemSearch – “Learning to Rank for Semantic Search”
    - Using Wikipedia usage data for ranking in RDF datasets

  LDOW – “Automatically Annotating Text with Linked Open Data”

Thank you

- Questions?