Audience Segmentation Based on Topic Profiles

Ljubljana, 9 October 2017

Matic Kladnik, Luka Stopar, Blaz Fortuna,
Dunja Mladenić

Motivation

- Web portals can offer frequently updated content:
  - News articles
  - Market data
  - Financial data
- First-time Users -> Returning Users -> Loyal Users
- Serve interesting content and advertisement to users

Motivation

- Audience Segmentation: dividing (website's) audience into smaller groups

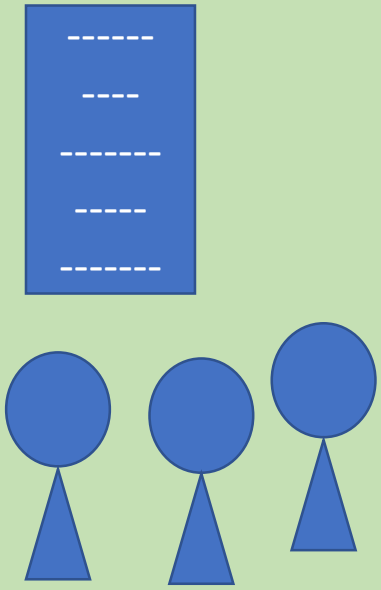
- Audience segmentation a key activity in audience analysis

Outline

- Data description
- Approach overview
- Architecture of the proposed approach
- Evaluation

Data Description

- Visit logs of users
- Content from almost 3000 pages (crawled)
- Anonymized user data from over 500k users from visit logs
- All page content (text) in English



Data Description

- Each web page represented as a Bag-of-Words (BoW) with TF-IDF weights

- Content Labels for each page
  - Assigned by editorial team
  - Used to annotate users
  - Examples: Brexit, jobs, Europe, markets

- User properties
  - Demographics
  - Page content labels
  - (pages visited)

Approach Overview

- Audience segmentation commonly groups users by their common interest, behavior, etc.

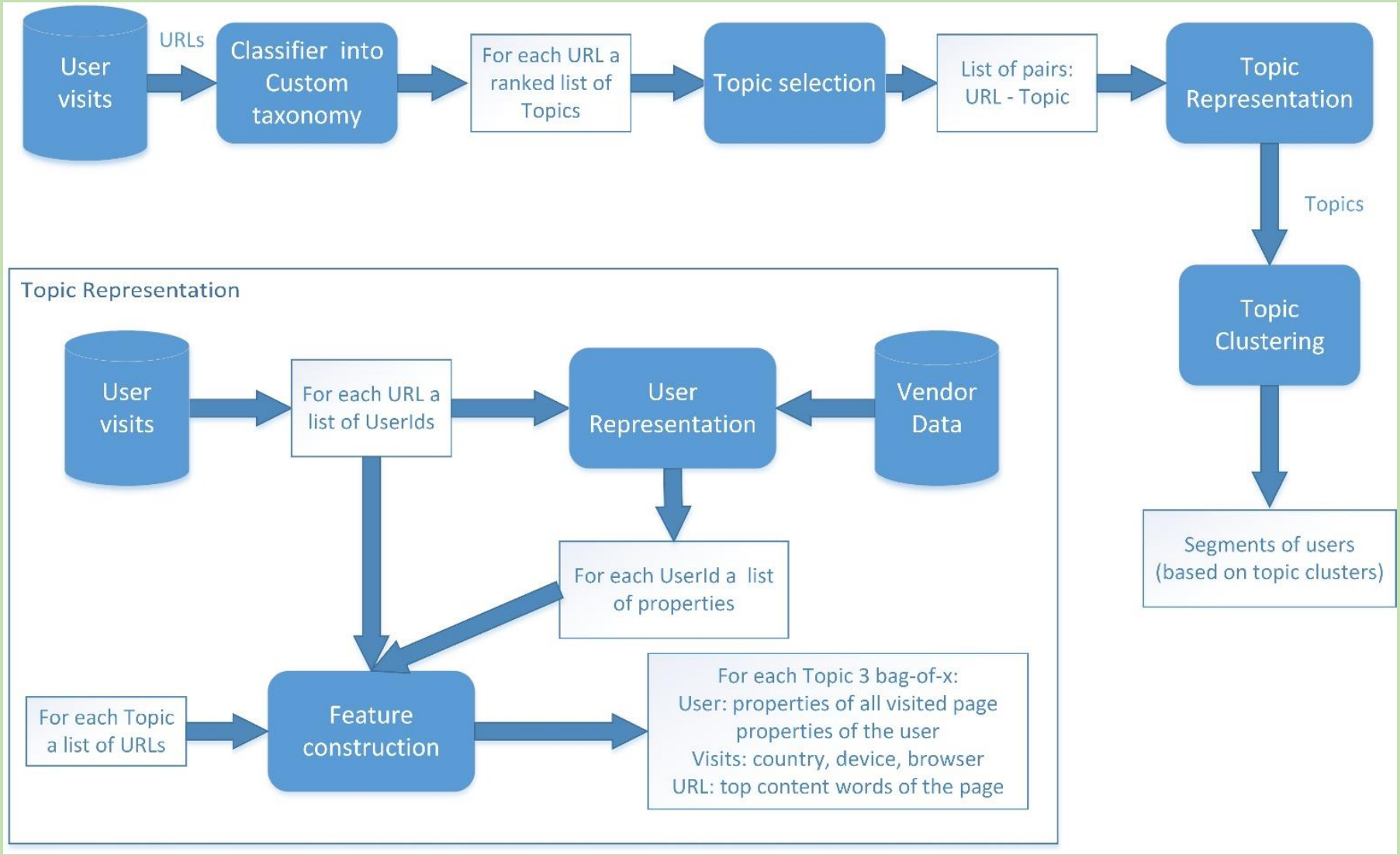
- However:
  - Users may have several interests
  - Exhibit different behavior (current focus)

- This approach might group users that only share some interests with a third user

Approach Overview

- Our approach: audience segmentation based on similarity of topics that users are interested in
- Obtain segments of users through topics of the visited web pages
- Same user can appear in several segments

Architecture of the Proposed Approach



Page Classification

- DMOZ classifier with Custom Taxonomy
- Classify each web page into a hierarchical content topic
- Pages classified to any level of hierarchy

Page Classification

- Upper level topics give context

- Science/Technology/Aerospace

- Business/Aerospace_and_Defense/Aeronautical

- More information about content topic

Features for Audience Segmentation

- In experimental evaluation we combine two sources of data

Source	Description	No. of values
Web page	BoW - Words from the Web pages	**59929**
User interest	**Content Labels of the visited Web pages**	**1268**

Cluster Dispersity Measure

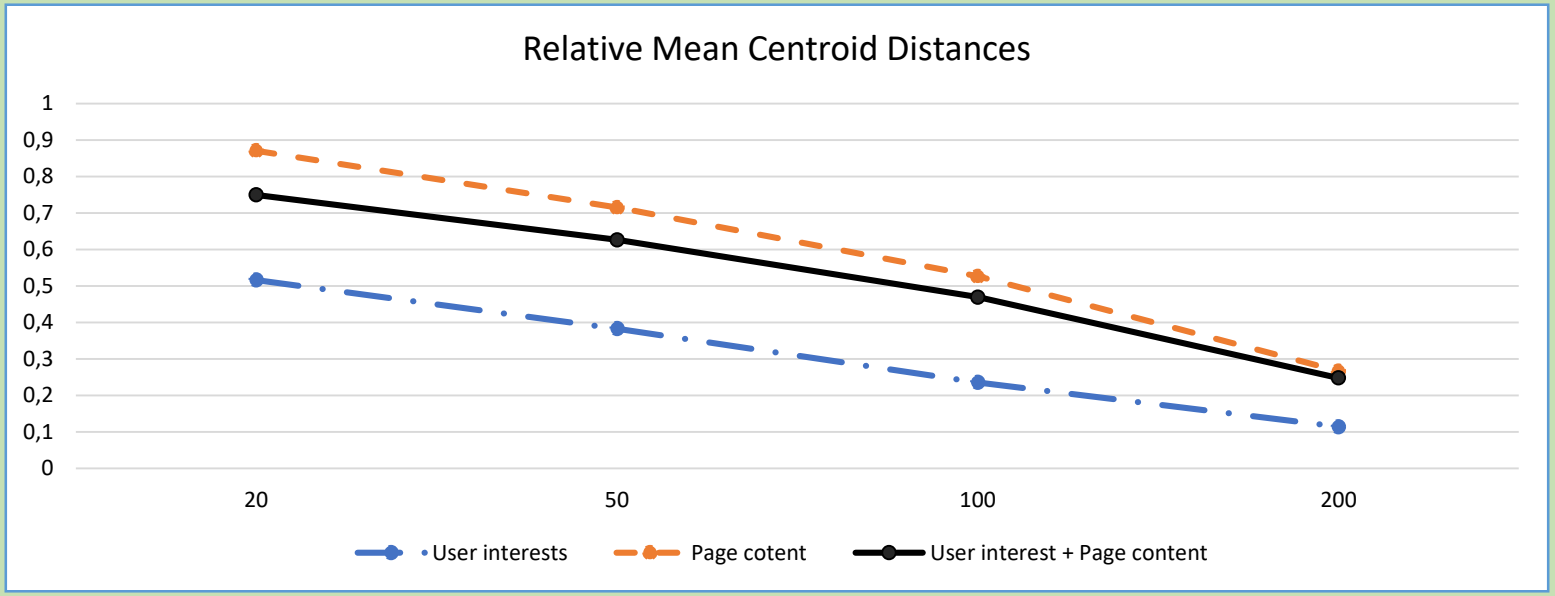
- Used to compare influence of different feature sources
- Weighted average distance between examples and their centroid normalized by average distance to the global mean

$$D = \frac{\sum_{i=1}^{k} \frac{n_i}{n} \sum_{j=1}^{n_i} \frac{1}{n_i} d(\mu_i, x_j)\Big)}{\frac{1}{n} \sum_{j=1}^{n} d(\mu, x_j)} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} d(\mu_i, x_j)}{\sum_{j=1}^{n} d(\mu, x_j)}$$

- Examples in more compact clusters lie closer to the centroid

Evaluation

- Comparing relative mean distances for different data representations over different number of clusters (k)



Evaluation: example clusters

- From 50 clusters
- Some clusters with a broad range of topics

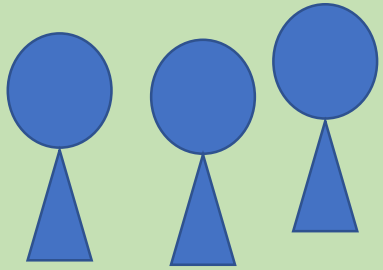
Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Business/Biotechnology_and_Pharmaceuticals	Recreation/Models	Home/Personal_Finance	Health/Addictions	Business/Financial_Services/Venture_Capital/Regional
Health/Child_Health	Science/Astronomy		Recreation/Drugs	Business/Transportation_and_Logistics/Bus
Health/Conditions_and_Diseases/Cancer			Business/Chemicals/Wholesale_and_Distribution	Business/Transportation_and_Logistics/Rail
Health/Conditions_and_Diseases/Immune_Disorders			Business/Food_and_Related_Products/Beverages	Government/Agencies
ealth/Conditions_and_Diseases/Infectious_Diseases			Science/Biology/Bioinformatics	Recreation/Autos/Makes_and_Models/Honda
Health/Pharmacy			Society/Issues/Gun_Control	Science/Environment
				Science/Environment/Carbon_Cycle, …

Evaluation

- Comparing average size of the Audience Segments in relation to the granularity of segmentation

No. of segments	Average size	Median size
20	43592.35	589.5
50	17436.94	301
100	8718.47	229.5
200	4359.235	193

Conclusion



- Topic profiles: user properties + content of web page
- Audience segmentation based on topic profiles of web pages
- Topics obtained by classifying pages with custom taxonomy
- Small scale experiments show some promise
- Future work:
  - Large scale experiments
  - Add additional topic features as in proposed architecture: user demographics etc.