Audience Segmentation Based on Topic Profiles

Matic Kladnik, Luka Stopar, Blaz Fortuna, Dunja Mladenić
Jožef Stefan Institute
and
Jožef Stefan International Postgraduate School
Jamova 39, 1000 Ljubljana, Slovenia
e-mail: matic.kladnik@ijs.si

ABSTRACT

Audience segmentation is often applied on Web portals to gain insights into the audience, support targeted marketing and in general provide on-line recommendations to the users. We propose an approach to audience segmentation that is based on using machine learning on topic profiles of the visited content. Our preliminary experiments on a small sample of log-data show that the proposed approach is promising and the proposed combination of features capturing short term and long-term user interest gives better results than using only the short-term interests of the user.

1. INTRODUCTION

Large number of retuning users regularly visiting the same Web portals offer an opportunity to apply audience modeling considering descriptions of the visited content, different characteristics of the user and the user behavior. Instead of the usual audience segmentation based on user profiles that is also commonly adopted in recommendation systems [1], we propose an audience segmentation approach that is based on the topic profiles of the visited content. As the users potentially have interests in different topics we allow the same user to occur in several segments.

In this paper, we described the proposed approach and on a small sample of real-world log-data test the hypothesis that the proposed combination of features capturing the content of the recently visited pages and the properties of all the pages visited by the user improves the quality of the segmentation.

The rest of this paper is structured as follows. Section 2 describes the problem setting and the dataset used in the experiments. The proposed approach is described in Section 3 together with the description background knowledge that we have used for mapping form the space of users into the space of topic profiles. Section 4 gives the results of the preliminary experiments, while the conclusions are presented in Section 5.

2. PROBLEM SETTING AND DATA

High-quality Web portals that offer regularly updated content, such as market data, news articles or financial data, attract many loyal users [2]. Today, vendors offer user data obtained by third-party cookies that cover a whole range of user properties including demographics, interest, geography. The problem that we are addressing is automatic audience segmentation where potentially vendor data on the users is available in addition to the usual Web log files and content of the visited pages. In addition we propose to use background knowledge in the form of pre-trained machine learning model that classifies Web pages into a predefined custom taxonomy. In this way, each Web page is based on its textual content assigned a ranked list of content topics of different granularity. For instance, the assigned topics may be Business, Business/Financial_Services/Medical_Billing, Business/News_and_Media, Society/Issues/Gun_Control.

The dataset that we have used to test the proposed approach was obtained from an international media company. Almost 3 000 Web pages were crawled from the company Web site. The anonymized user data was obtained for more than half a million users that have visited the Web site within one selected day. All the considered text is in English language.

We have pre-processed the data to remove references to Web pages that have limited textual content or un-standard formatting. The Web pages were processed in a standard way to extract the textual content, remove the standard English stop-words and represent each Web page as a bag-of-words (BoW) with TFIDF weight. In addition to the content, each page has metadata including a set of manually assigned content labels done by the editorial team. For instance, brexit, Europe, money, davos, jobs, London, markets. These content labels were historically used to annotate the users visiting the pages. Each user is thus described by a set of properties including demographics and the content labels of the pages visited over a longer period of time.

3. APPROACH DESCRIPTION

Audience segmentation is commonly based on grouping the users by their common interests and some other e.g., demographic properties and behavioral similarity. However, the same user may have several interests and exhibit different behavior depending on the current focus. This may result is grouping together the users that do not have much in common except that they share some (but not the same) interests with the third user.

Thus we propose an approach to audience segmentation based on the similarity of the topics that the users are interested in. The idea is to view the problem through topics of the visited Web pages and based on that obtain segments of the users.

Architecture of the proposed approach is shown in Figure 1. The whole pipeline consists of several steps:

1. From the log file of the user visits we obtain a list of visited pages (URLs).
2. By using background knowledge in the form of machine learning model for classifying documents into a predefined custom taxonomy, we assign a ranked list of topics to each URL.
3. For each URL we select one or a few topics with the highest rank and form a collection of URL – Topic pairs.
4. Representation of the topics is based on a list of URLs that were assigned the topic and the list of UserIds of the users that visited the URLs.
5. Topic profiles are processed by a clustering algorithm to obtain segments of the users.

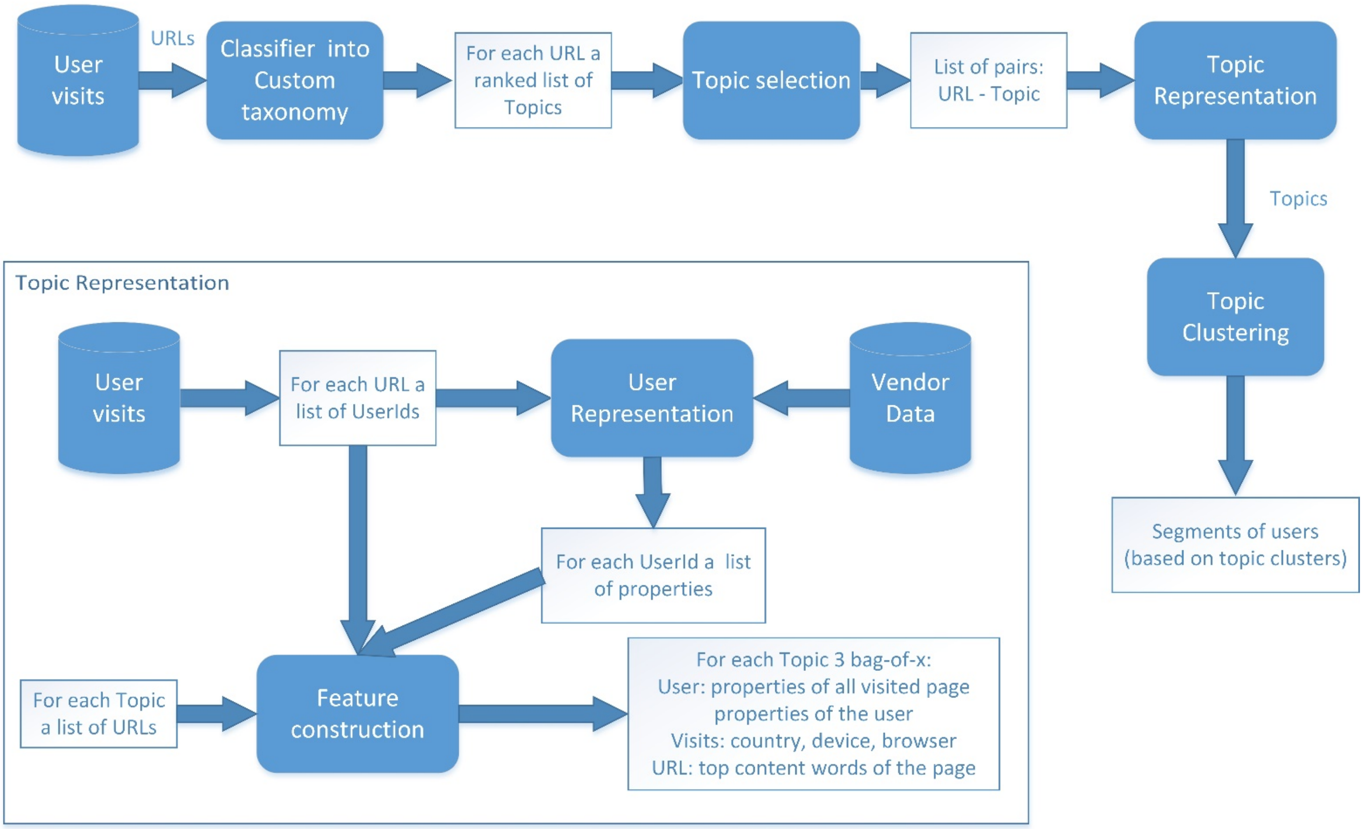
User visits
URLs
Classifier into Custom taxonomy
For each URL a ranked list of Topics
Topic selection
List of pairs: URL - Topic
Topic Representation
Topics
Topic Clustering
Segments of users (based on topic clusters)
Topic Representation
User visits
For each URL a list of UserIds
User Representation
Vendor Data
For each UserId a list of properties
For each Topic a list of URLs
Feature construction
For each Topic 3 bag-of-x:
User: properties of all visited page properties of the user
Visits: country, device, browser
URL: top content words of the page

Figure 1. Architecture of the proposed approach to audience segmentation. The user visits of the Website and combined with the visited content and properties of the users to obtain segments of the users based on the topic clusters.

As topic profiles are built around the URLs that were assigned the topics using background knowledge, we separately keep a list of UserIds for each URL (obtained from the log file). When constructing features for each topic, we combine different data source:

- content of the Web pages visited by the users,
- properties of the visited pages,
- properties of the users, and
- information about the visits.

Background knowledge that we have used in the experiments is based on a subset of a large custom taxonomy [2]. The subset was defined by the domain expert from the company of the Web portal and consists of several hundred of topics.

We use DMOZ classifier with custom taxonomy to classify each Web page into a hierarchical content topic. Pages can be classified into topics on different levels of hierarchy, where lower levels give more specific classification. Upper levels also give context to the lower levels of classification. If we compare the following two topics that mention aerospace in their hierarchy:

- Science/Technology/Aerospace,
- Business/Aerospace_and_Defense/Aeronautical,

we can see that the first content topic is put into the context of Science and Technology, whereas the second is put into the context of Business and Aerospace and Defense. This approach gives us more information about the content topic.

In our experiments, content of the Web pages is obtained by crawling the Web portal. User data including properties of the visited pages is obtained from the Vendor data.

4. EVALUATION

In the experimental evaluation we combine two sources of data: URLs from the log file and the user data from the user's history. The two features sets used for data representation correspond to these two data sources: bag-of-words from the Web page corresponding to the URL; the user interest in the form of a collection of content labels of the Web pages visited by the user over a longer period of time (see Table 1).

Table 1. Features used for audience segmentation.

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

To compare the influence of different feature sources on the obtained segmentation we use a cluster dispersity measure. Specifically, we measure the weighted average distance between the examples and their centroid normalized by the average distance to the global mean (i.e. center of the data). The formula is given below:

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

D represents the dispersity, d is a distance measure (in our case cosine distance), n_i and μ_i are the size and centroid of the i-th cluster, x_j is the j-the example and μ is the global mean. Intuitively, examples in more compact clusters will lie closer to the centroid and so will contribute less to the dispersity than more disperse clusters.

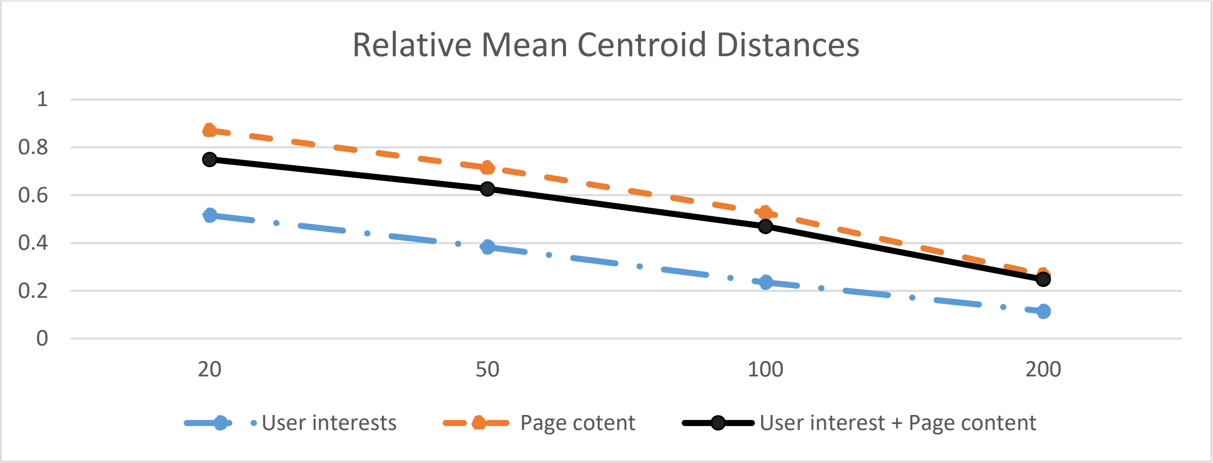

Figure 2 Experimental results comparing relative mean centroid distances for the three data representations over different number of clusters (*k* ranging from 20 to 200).

The experiments on clustering topic profiles were performed on three different data representations. One that considers only content of the visited pages, one that considers only user interest and a combination of the two feature sets. We have applied k-means clustering, varying the value of parameter *k* (the number of clusters) form 20 to 200. As the results of the applied clustering method depends on the random seed for choosing the initial clusters, we have repeating the process five time for each value of k. Figure 2 shows the results of the experiments averaged over five runs.

We can see that the smallest distance of topic clustering is obtained when the data is represented only by user interests, which represents a long-term interest of the user on an aggregated level. This can be particularly attributed to the fact that the number of different content labels is much smaller than the number of different words from the page content. Combining user interest (capturing history of the user) and page content (of pages visited in the considered log file) gives better results than using only page content.

Looking at the topic clusters that we have obtained, we can see that the similar topics are clustered together (see Table 2). For instance, the topics such as Health/Addiction, Business/Chemicals/Wholesale_and_Distribution, Recreation/Drugs are in the same cluster.

Table 2. Illustrative example of some clusters obtained when generating 50 clusters using both feature sets. For a few selected clusters we show the topics that belong to the cluster.

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, …

To obtain audience segments from the clustering of the topic profiles, we map the topic clusters onto a set of UserIds based on the user visits of the URLs that are classified to each of the topic in the cluster. In this way we obtain non mutually exclusive audience segments. The average number of users per segment is given in Table 3. From the table we can see are assigned to the same segment, due to similarity with the other users form the same segment.

Table 3. Average size of the audience segments in relation to the granularity of the 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

5. CONCLUSION

We have proposed an approach to audience segmentation based on topic profiles of the visited Web pages instead of the commonly used user profiles. The topic are obtained by classifying the visited Web pages into a custom taxonomy. The classification is performed automatically using a pre-trained machine learning model. The topic profiles are formed from properties of the users visiting the Web page that are classified into the topic, and the content of the Web page.

Preliminary experiments on a small sample of log-data show that the proposed approach is promising, grouping together similar topics and based on that segmenting the audience into reasonably populated segments. Namely, one of the issues with audience segmentation when the users have multiple interests is that many users that are not similar to each other

Larger scale experiments are needed in the future work to confirm the results and provide additional insights into the other properties of the users form the same segment, such as demographics, geography, job.

6. ACKNOWLEDGMENTS

This work was partially supported by the Slovenian Research Agency.

References

[1] Adomavicius, G. and Tuzhilin, A., Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions, *IEEE Transactions on on Knowledge and Data Engineering*, Vol. 17, No. 6, June 2005.

[2] Fortuna, B., Fortuna, C., Mladenic, D., Real-time news recommender system. In *Proceedings of Machine learning and knowledge discovery in databases : European conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010* (Lecture notes in computer science, ISSN 0302-9743, Lecture notes in artifical intelligence, vol. 6323). Berlin; Heidelberg; New York: Springer. 2010, vol. 6323, pp. 583-586.

[3] Grobelnik, M., Bank, J., Mladenic, D., Novak, B., Fortuna, B.. Using DMoz for constructing ontology from data stream. In Proceedings of the 28th International Conference on Information Technology Interfaces, June 19-22, 2006, Cavtat/Dubrovnik, Croatia, (IEEE Catalog, No. 06EX1244). Zagreb: University of Zagreb, SRCE University Computing Centre. cop. 2006, pp. 439-444.