How much do age and gender affect news topic preferences?

Nemanja Djuric[†], Mihajlo Grbovic[†], Dilan Görur[‡*], Vladan Radosavljevic[†]
[†]Yahoo! Labs, Sunnyvale, CA, USA, {nemanja, mihajlo, vladan}@yahoo-inc.com
[‡]Microsoft, Sunnyvale, CA, USA, dilangrr@gmail.com

ABSTRACT

Despite the recognized relevance and significant resources invested in making an online experience more personalized and enjoyable, there still remains a problem of personalization for a large number of online services in a presence of limited user information. In this paper, we recognize cold start problem related to presentation of personalized articles on Yahoo news stream, and address it as a topic ranking problem. For active users news articles are commonly ranked according to their historical news topic preferences, while for new users article ranking is often based on the globally popular topics, which may result in diminished user experience. To mitigate this problem, we exploit demographic information to learn separate topic preferences for different age-gender user groups, and evaluate this simple, efficient approach on a large data set comprising more than 750,000 users and their news stream activity logs from Yahoo homepage. Interestingly, the results show that gender is more indicative of topic preferences than age, and confirm that demographics carry a strong useful signal for personalizing content in a topic ranking setting, where by using age and gender information we obtained 5% improvement in Kendall's tau over the baseline predicting globally popular topics.

1. INTRODUCTION

Personalization has become increasingly important in the recent years. It is defined as "the ability to proactively tailor products and product purchasing experiences to tastes of individual consumers based upon their personal and preference information" [3], which may lead to improved user experience and directly translate into financial gains for online businesses [14]. In addition, personalization fosters stronger bond between users and companies, and can help in increasing user loyalty and retention [2]. For these reasons it has been recognized as an important strategic task of internet companies [7, 13], and is a focus of significant research efforts. Personalized content has already become an integral part of many popular online services, a trend likely to continue in the future [15].

Considering the importance of providing more personalized content for improved online experience, personalization of central online services, such as news websites, is of particular interest [7, 9]. Assuming each article is assigned a list of topic tags (e.g., sports, politics), personalized news stream can be obtained by casting the task into topic ranking problem [4, 10, 11] and learning user topic preferences in a form of a total ranking. For example, as the counts of read articles are directly correlated with the user preferences for article topics, preferences can be obtained by sorting the counts. On the other hand, there is an evident cold start problem for new users. Activity logs are not available for users who signed up recently, and it is not clear which content would maximize user satisfaction. Due to difficulty of the problem, popular heuristics addressing this issue include suggesting random content or globally popular content. However, these simple methods can be suboptimal, and in the following we explore how limited demographic data, such as age and gender, can improve user experience. Such approach is particularly of interest as it allows very efficient model for web-scale personalization. We evaluate the approach on large, real-world data set comprising more than 750,000 Yahoo homepage users.

2. METHODOLOGY

We view the personalization task as a topic ranking problem, where the input space is defined by a feature vector $\mathbf{x} \in \mathcal{X}$, and the output is defined as a ranking of topics $\pi \in \Pi$. If by π_i we denote the topic at the i^{th} position in the topic ranking π, then for any $i < j$ it follows that topic π_i is preferred over topic π_j. Given a training sample from the underlying distribution $D = \{(\mathbf{x}_n, \pi_n)\}_{n=1,\ldots,N}$, where $\mathbf{x}_n$ is a d-dimensional feature vector and π_n is a vector containing either a total or a partial order of a finite set $\mathcal{Y}$ of L topics, the goal is to learn a model that maps new example $\mathbf{x}$ into a total topic order.

2.1 Mallows model for topic ranking

We represent each user by a input vectors $\mathbf{x}$ which encode user's age and gender, and π represents topic preference ranking. We partition the users in several groups according to their demographic data, by splitting the whole age span in equal segment of width r and learning separate topic ranking model for each age-gender user group. Note that this approach was earlier considered for personalization in [1, 5], although not in the topic ranking setting. Due to the large

[*] This work was done while the author was at Yahoo! Labs.

1988 female Arts & Entertainment:341 Celebrities:284 Politics & Government:170 Media:111 Family & Relationships:107 Crime & Justice:106 Music:39 Sports & Recreation:37 Travel & Tourism:34 Health:32 Society & Culture:30 Television:29 Movies:28 Fashion:26 WIKI_Kim_Kardashian:24

1988 male Sports & Recreation:286 Politics & Government:140 Arts & Entertainment:92 Ice Hockey:84 Finance:76 Baseball:71 Media:57 American Football:54 Investment & Company Information:49 Budget, Tax & Economy:44 WIKI_National_Hockey_League:43 Basketball:38 Celebrities:37 Business:35 WIKI_New_York_Rangers:32

1933 female Politics & Government:151 Health:65 Arts & Entertainment:62 Crime & Justice:60 WIKI_Barack_Obama:48 Disease & Medical Conditions:43 Celebrities:41 Society & Culture:40 Family & Relationships:37 Government:36 WIKI_China:33 WIKI_Avian_influenza:27 Public Health:27 Unrest, Conflicts & War:26 Science:26 WIKI_Boston:22 Military & Defense:22 Foreign Policy:22 WIKI_White_House:21

Figure 1: Example of user preferences

001. Crime & Justice
002. Celebrities
003. Media
004. Unrest, Conflicts & War
005. Government
006. Religion & Beliefs
007. Company Legal & Law Matters
008. Military & Defense
009. Cultural Groups
010. American Football
· · ·
170. Transportation Services
171. Nanotechnology
172. Lacrosse
173. Wrestling
174. Animal Science
175. Genealogy
176. Entrance Examinations
177. Fantasy Sports
178. Politics & Government Events
179. Computer Science

Figure 2: Global YCT topic ranking

scale of the personalization task, we consider Mallows model [12] for learning topic preference ranks for each user group, which allows efficient training and inference.

The Mallows ranking model is a distance-based probabilistic model for permutations. The probability of permutation π of L topics is given in terms of a permutation distance d,

$$\mathbb{P}(\pi \mid \theta, \rho) = \frac{\exp\big(-\theta\ d(\pi, \rho)\big)}{Z(\theta, \rho)}, \tag{2.1}$$

where π is a candidate permutation, ρ is a central ranking, $\theta \in \mathbb{R}$ is a dispersion parameter, d is, in our case, the Kendall's tau distance d_τ, and $Z(\theta, \pi)$ is a normalization constant computed as

$$Z(\theta, \pi) = \sum_{\rho \in \Pi} \exp\big(-\theta\ d(\rho, \pi)\big). \tag{2.2}$$

The maximum probability is assigned to the central ranking ρ, which can be found using the maximum likelihood principle. Given a population of N topic rankings $\{\pi_n\}_{n=1,\dots,N}$, probability that we observe training topic rankings is

$$\mathbb{P}(\{\pi_n\}_{n=1,\dots,N} \mid \theta, \rho) = \prod_{n=1,\dots,N} \mathbb{P}(\pi_n \mid \theta, \rho). \tag{2.3}$$

The maximum likelihood estimate (MLE) of ρ is the one that minimizes average Kendall's tau, where the solution can be found by exhaustive search. The MLE of θ is found from the mean observed distance from ρ, $\frac{1}{N}\sum_{n=1,\dots,N} d_\tau(\pi_n, \rho)$ by line search.

To find globally popular topics we can simply consider the entire training set as the training population, while for each age-gender group we consider only topic rankings of users belonging to that specific user group. However, the Mallows model has high computational complexity of $\mathcal{O}(L!)$, and also it cannot directly model incomplete topic ranks. In order to improve the efficiency of the model training, we make use of the approximate solution [4] that uses a simple Borda count algorithm [8]. The central rank is found by voting, where each π_n votes in the following manner. Topic π_1 receives L votes, π_2 receives $(L-1)$ votes, and so on. Finally, all the votes are summed up and the topic with the most votes is ranked first in ρ, the topic with the second most votes is ranked second in ρ, and so on. The approximation is valid as it has been shown that Kendall's tau is well approximated by Spearman's rank correlation [6], whose median can be computed using Borda. To account for missing topics, the Borda count is modified in [4] such that partial rankings π_n of $L_n < L$ topics vote in the following manner: the j^{th} ranked topic receives $(L_n - j + 1)(L+1)/(L_n + 1)$ votes, the missing topics receive $(L+1)/2$ votes, and the topic ranking is obtained by sorting the votes.

2.2 Data set description

We selected a subset of 757,371 Yahoo homepage users over 3 months during an undisclosed period of time, as well as the news articles and their associated Wiki and Yahoo Content Taxonomy (YCT) tags[1]. For each user, we aggregated the counts of topics associated with the articles they read, and derived a topic ranking π (comprising 101,186 Wiki topics and 304 YCT topics) by ordering the topics in the descending order.

In Figure 1 we give an example of user topic preferences. We randomly sampled two younger and one more senior user, and show their preferred topics. Following age and gender information, we give a list of preferred topics with a number of read articles annotated with a specific topic tag. We can see that the gender feature carries a strong signal regarding user preferences. While the young male user was mostly interested in sports (e.g., "Baseball", "Ice Hockey", "Basketball" YCT topics, and "NHL" and "NY Rangers" Wiki topics), the young female user was mostly

[1] http://developer.yahoo.com/search/content/V2/contentAnalysis.html, accessed June 2014

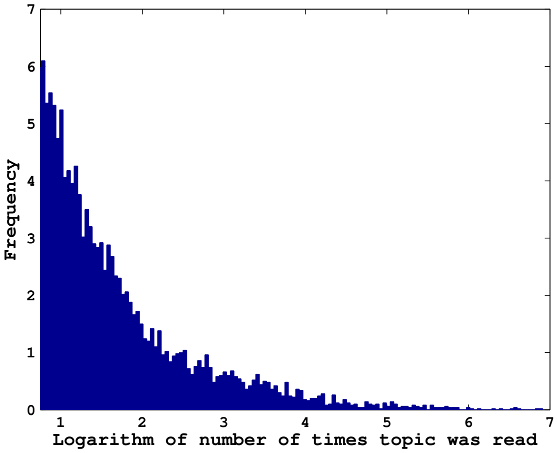

Figure 3: Histogram of topic popularity

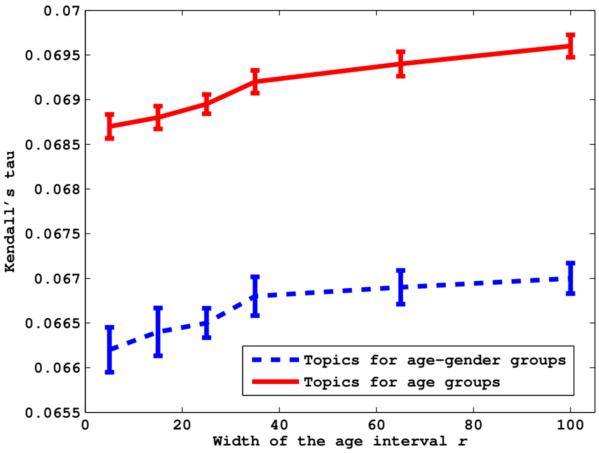

Figure 4: Comparison of different approaches

interested in celebrities (”Arts & Entertainment”, ”Celebrities” YCT topics, and ”Kim Kardashian” Wiki topic). If we also consider age information, we can see that younger users are more interested in entertainment topics, while older user also expressed strong interest in more serious topics, such as ”Disease & Medical Conditions”, ”Public Health”, and ”Unrest, Conflicts & War” YCT topics. We can see that age and gender provide a strong signal useful for improved personalization, indicating a strong potential of news personalization approach which uses demographic information.

Furthermore, we explored the popularity of Wiki and YCT topics, shown in Figure 3 (values on axes were rescaled to remove sensitive information). The distribution roughly follows the Zipf's law, where a small number of topics is frequently read by the users (e.g., ”Celebrities”, ”Barack Obama”), and a large number of rarely read topics are specific to each user's preference (e.g., ”Lacrosse”, ”Salzburg”). Thus, if only the most popular topics are recommended to a user for further reading, which is a commonly used approach illustrated in Figure 2, some of less frequently read topics might not be included in the news stream which could otherwise help improve user experience.

3. EXPERIMENTS

For the empirical evaluation of the proposed approach we considered 179 YCT topics from the second level of the YCT hierarchy. We randomly split the data set, and used 657,371 users for training and remaining 100,000 for testing. As a baseline, we considered the approach that always predicts generally most preferred topics, obtained by fitting a single Mallows model to the training set. To compare different models, we increased r from 5 to 100, and measured Kendall's tau between true and predicted rankings π_n and $\hat{\pi}_n$. We report an average normalized Kendall's tau,

$$loss_{LR} = \frac{1}{N} \sum_{n=1}^{N} \frac{2 \cdot d_\tau(\pi_n, \hat{\pi}_n)}{L \cdot (L-1)}, \qquad (3.1)$$

as well as confidence intervals of two standard deviations after 10 experiments. Similarly to results in [1, 5], in Figure 4 we can see that by considering demographics we obtained improved performance over the baseline (shown by full line for $r = 100$). Including only gender (i.e., dashed line for $r = 100$) already led to significant drop in $loss_{LR}$, which further decreased as we segmented the users using finer age resolutions (i.e., as we decreased r).

Next, we considered topic preferences for different user age-gender groups. In Figures 5 and 6 we give top 10 topic preference rankings for younger and older populations in male and female groups, respectively. Interestingly, we can see that topics such as ”Celebrities”, ”Crime & Justice” and ”Media” are popular across ages and genders. We believe that this is in some part due to imbalance in topic distributions (i.e., news articles from ”Celebrities” topic are found more often than articles from ”Computer Science” topic), which will be investigated in our future work. In addition, certain topics are preferred only for certain gender (e.g., ”Military & Defense” and ”Foreign Policy” for male users or ”Parenting” for female users) or age group (e.g., sports topics for younger users or more serious topics such as ”Investments” or ”Medical Conditions” for older users).

4. CONCLUSIONS AND ONGOING WORK

In this paper we reported our initial findings on using demographic information for predicting news topic preferences on a data set comprising more than 750,000 Yahoo users. While simple, this approach is commonly used in practice as it results in very good performance that is hard to outperform with more complex methods. More specifically, we segment the users by their age-gender information and learn a different Mallows topic ranking model for each user group. The results on large real-world data set confirm that the approach leads to significant improvement over the baseline which computes a single globally popular topic ranking for all users. In addition, we found that gender is more informative than age information when predicting user's news preferences, resulting in significant performance improvements.

To further improve personalization of news stream, we are currently exploring efficient non-linear ranking models to improve the effectiveness of news stream personalization while not sacrificing efficiency. In addition, we are consider-

male, aged 21-25
01. Crime & Justice
02. Celebrities
03. Media
04. Unrest, Conflicts & War
05. Military & Defense
06. Basketball
07. American Football
08. Government
09. Foreign Policy
10. Cultural Groups

male, aged 76-80
01. Crime & Justice
02. Celebrities
03. Media
04. Unrest, Conflicts & War
05. Government
06. Military & Defense
07. Religion & Beliefs
08. Foreign Policy
09. Environment
10. Investment & Company Information

Figure 5: YCT ranking for male age groups

female, aged 21-25
01. Celebrities
02. Crime & Justice
03. Media
04. Fashion
05. Parenting
06. Unrest, Conflicts & War
07. Religion & Beliefs
08. Cultural Groups
09. Company Legal & Law Matters
10. Athletics, Track & Field

female, aged 76-80
01. Crime & Justice
02. Celebrities
03. Media
04. Religion & Beliefs
05. Government
06. Unrest, Conflicts & War
07. Parenting
08. Death & Funeral
09. Disease & Medical Conditions
10. Cultural Groups

Figure 6: YCT ranking for female age groups

ing exploiting additional user data, such as shopping history and data from other available sources, to better personalize the news stream to each user's individual preferences.

References

[1] D. Agarwal, B.-C. Chen, P. Elango, N. Motgi, S.-T. Park, R. Ramakrishnan, S. Roy, and J. Zachariah. Online models for content optimization. In *Advances in Neural Information Processing Systems*, pages 17–24, 2008.

[2] J. Alba, J. Lynch, B. Weitz, C. Janiszewski, R. Lutz, A. Sawyer, and S. Wood. Interactive home shopping: Consumer, retailer, and manufacturer incentives to participate in electronic marketplaces. *The Journal of Marketing*, pages 38–53, 1997.

[3] R. K. Chellappa and R. G. Sin. Personalization versus privacy: An empirical examination of the online consumerŠs dilemma. *Information Technology and Management*, 6(2–3):181–202, 2005.

[4] W. Cheng, J. Hühn, and E. Hüllermeier. Decision tree and instance-based learning for label ranking. In *Proc. of the 26th International Conference on Machine Learning (ICML-09)*, pages 161–168, 2009.

[5] W. Chu and S.-T. Park. Personalized recommendation on dynamic content using predictive bilinear models. In *Proceedings of the International Conference on World Wide Web*, pages 691–700. ACM, 2009.

[6] D. Coppersmith, L. K. Fleischer, and A. Rudra. Ordering by weighted number of wins gives a good ranking of weighted tournaments. *ACM-SIAM Symposium on Discrete Algorithms*, pages 776–782, 2006.

[7] A. S. Das, M. Datar, A. Garg, and S. Rajaram. Google news personalization: Scalable online collaborative filtering. In *Proceedings of the International Conference on World Wide Web*, pages 271–280. ACM, 2007.

[8] J. C. de Borda. *Memoire sur les Élections au Scrutin.* Histoire de l'Academie Royale des Sciences, 1781.

[9] G. M. Del Corso, A. Gullí, and F. Romani. Ranking a stream of news. In *Proceedings of the 14th International Conference on World Wide Web*, pages 97–106. ACM, 2005.

[10] M. Grbovic, N. Djuric, and S. Vucetic. Supervised clustering of label ranking data. In *SIAM International Conference on Data Mining*, pages 94–105. SIAM, 2012.

[11] M. Grbovic, N. Djuric, and S. Vucetic. Multi-prototype Label Ranking with Novel Pairwise-to-Total-Rank Aggregation. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*. AAAI Press, 2013.

[12] C. L. Mallows. Non-null ranking models. *Biometrika*, 44:114–130, 1967.

[13] U. Manber, A. Patel, and J. Robison. Experience with Personalization on Yahoo! *Communications of the ACM*, 43(8):35–39, 2000.

[14] D. Riecken et al. Personalized views of personalization. *Communications of the ACM*, 43(8):27–28, 2000.

[15] A. Tuzhilin. Personalization: The state of the art and future directions. *Business Computing*, 3:3, 2009.