# Content-Based Methods for Predicting Web-Site Demographic Attributes

Santosh Kabbur
Department of Computer Science
University of Minnesota, Twin Cities
skabbur@cs.umn.edu

Eui-Hong Han
Sears Holdings Corporation
Chicago
han@cs.umn.edu

George Karypis
Department of Computer Science
University of Minnesota, Twin Cities
karypis@cs.umn.edu

***Abstract*—Demographic information plays an important role in gaining valuable insights about a web-site's user-base and is used extensively to target online advertisements and promotions. This paper investigates machine-learning approaches for predicting the demographic attributes of web-sites using information derived from their content and their hyperlinked structure and not relying on any information directly or indirectly obtained from the web-site's users. Such methods are important because users are becoming increasingly more concerned about sharing their personal and behavioral information on the Internet. Regression-based approaches are developed and studied for predicting demographic attributes that utilize different content-derived features, different ways of building the prediction models, and different ways of aggregating web-page level predictions that take into account the web's hyperlinked structure. In addition, a matrix-approximation based approach is developed for coupling the predictions of individual regression models into a model designed to predict the probability mass function of the attribute. Extensive experiments show that these methods are able to achieve an RMSE of 8–10% and provide insights on how to best train and apply such models.**

***Keywords*-Demographic Attribute Prediction, Content Based Models, Regression, Inlink Count, Probability Mass Function**

## I. Introduction

Effective online advertising approaches rely heavily on being able to personalize the advertisements based on information that is known about the individual users. Among this information, demographic attributes (e.g., age, gender, occupation, etc.) about the audience of a web-site (i.e., the set of users viewing the web-pages) play an important role in gaining valuable information about a web-site's users and is used extensively to target online advertisements.

Most of the existing approaches for determining the demographic attributes of a web-site's audience are based on information obtained from user panels. In this approach, which is similar to the methods used to determine the audience characteristics of traditional media (e.g., TV and radio), a set of users with known demographic information is recruited and their web-browsing history is recorded over a period of time. The demographic attributes of the various web-sites are determined by propagating the known demographic information of the panel members based on their browsing histories. For those web-sites that are visited by a sufficiently large number of panel members, this approach leads to reliable estimations. However, in order to cover the large number of web-sites in existence, this approach requires extremely large panels, which makes it impractical. For this reason, machine-learning approaches have recently attracted attention [1], [4], [9], [13] as they can potentially overcome the limitations (and costs) of conducting and monitoring user panels. These approaches employ supervised learning methods to build models for predicting the demographic attributes of a user or a web-site's audience by utilizing different features such as web-page content, web-browsing history, web-search history, and various profile information obtained from registered users. The ongoing research in this area has shown that machine learning approaches represent a viable alternative to user panels and they substantially increase the number of web-sites whose audience demographic attributes can be determined.

In this work we also focus on machine learning approaches for predicting the demographic attributes of a web-site but we restrict ourselves to approaches that compute predictions that do not utilize any directly or indirectly-obtained user information (e.g., web-browsing and web-search histories, registration information, etc.). This is motivated by the observation that users are becoming increasingly more concerned about sharing personal information and behavioral patterns on the web and less willing in having any of their information being used for ancillary purposes. Consequently, approaches that rely on these types of information are less general and can potentially become less applicable.

Within the context of these types of approaches our work focuses on investigating (i) how the performance of regression-based prediction models is affected by the set of features used to represent the different web-pages, and the granularity at which the models are being learned and applied; (ii) how the hyperlink structure of the web and the similarity among the web-site's web-pages can be used to improve the prediction performance; and (iii) how the predictions obtained from a set of regression models can be combined to obtain the probability distribution of the discrete random variable corresponding to the demographic attribute under consideration. Our investigation utilizes a

This work was supported in part by NSF (IIS-0905220, OCI-1048018, IOS-0820730), NIH (RLM008713A), and the Digital Technology Center at the University of Minnesota. Access to research and computing facilities was provided by the Digital Technology Center and the Minnesota Supercomputing Institute

dataset consisting of 8,215 web-sites and focuses on the gender and age demographic attributes. However, we believe that it is equally applicable to other demographic attributes as well.

Our experimental evaluation shows that compared to ground-truth data obtained from Comscore [3], our models achieve an RMSE of 9.97 and 8.26 for the gender and age demographic attributes, respectively, which are better by 21.1% and 11.2% than the corresponding RMSE values obtained by a baseline approach (RMSE of 12.64 and 9.34, respectively). In addition, our analysis of the ground-truth data provided by two commercial suppliers of demographic attribute information (Comscore and Quantcast [11]), shows that the performance of our models is comparable to the differences among their own sets of predictions.

## II. Methods

### A. Demographic Attributes

Even though the methods developed in this paper can be used to predict various demographic attributes, the focus in this paper is to predict the gender and age distribution of a web-site's audience. The gender attribute specifies the male and female percentages of a web-site's audience, whereas the age attribute provides a break-down of a web-site's audience in different age groups. The five age groups that we used in this study are Kid (3–12 years), Teen (13–17 years), Young-Adult (18–34 years), Adult (35–49 years) and Old (50+ years). They closely correspond to age groups that are of interest to advertising agencies.

### B. Overall Approach

We model a demographic attribute as a discrete random variable $X$ whose set of values $\mathcal{S}_{\mathcal{X}}$ correspond to the different population segments of interest. For example, in the case of gender, the corresponding random variable takes the values in the set $\{\text{Male}, \text{Female}\}$, whereas in the case of age, the corresponding set of values is $\{\text{Kid}, \text{Teen}, \text{Young-Adult}, \text{Adult}, \text{Old}\}$. The goal of the demographic attribute prediction problem is to predict the probability distribution of $X$. That is, for each $x \in \mathcal{S}_{\mathcal{X}}$, predict $P(X = x)$.

We followed a supervised learning framework for predicting a demographic attribute. A model is trained using features that are extracted out of a set of web-sites with known probability distributions for a given demographic attribute and such a model is used to predict the probability distribution of unknown web-sites. A key characteristic of the demographic attribute prediction problem is that it requires the prediction of the entire probability distribution of the corresponding discrete random variable. For this reason, the model learning and associated prediction methods that we developed consist of two steps. First, $\epsilon$-support vector regression ($\epsilon$-SVR) [12] is used to estimate the probability for each discrete value of the demographic attribute by treating it as an independent single-value estimation problem. Second, the individual predictions are used as input to a second learning problem whose goal is to estimate the overall distribution of the demographic attribute.

### C. Features

We used two types of features to represent each web-page. The first was designed to capture the web-page's textual content whereas the second was designed to capture the web-page's structure (e.g., organization, style, etc.).

To represent the web-site's textual content, we used the TF-IDF vector-space model from information retrieval [2]. We used a DOM-based approach to analyze the web-pages from each web-site in order to eliminate the "boiler-plate" content of each web-page [6], a stop word list to eliminate certain unimportant words, and the Porter's stemming algorithm to transform each term into its stem. Finally, the web-page's term vector was normalized to be of unit length. We will refer to this as the *T* representation of the web-page.

We also used the semi-structured nature of HTML documents to emphasize the terms that occur in certain HTML tags. Specifically, we focused on the title and section defining tags (TITLE and H1-H6) and modeled the terms that they contain as a separate term vector. The TF-IDF weighting scheme was used to determine the weights of each term and the resulting term vector was normalized to be of unit length. Each web-page was then represented as the concatenation of the original and this new term vector. We will refer to this as the *TH* representation of the web-page.

As shown in [7], visual appearance of a web-page influences the type of users it attracts. The set of features that we extracted were designed to capture the web-page's structure in terms of its style and organization, to represent its complexity. We extracted and used the following information from each web-page: (i) the number of different visual blocks, (ii) the number of hyperlinks, (iii) the number of images, (iv) the number of menus/lists, and (v) the proportion of script in web-page HTML. This information was extracted by counting the corresponding HTML tags (DIV, TABLE, H1-H6, A, IMG, LI) and calculating the ratio of size of text in script to total size of HTML. Overall we used 7 additional structural features. We will refer to this as the *THS* representation of the web-page.

### D. Model Granularity

The goal of the methods developed in this work are to make predictions for a demographic attribute at the web-site level. However, because the primary data corresponds to individual web-pages, it allows for the development of methods in which the training and prediction instances correspond to entire web-sites or individual web-pages.

In the web-site level models, the training and prediction instances correspond to entire web-sites. Each web-site is represented by a feature vector that corresponds to the

unit-length normalized sum of the feature vectors of its constituent web-pages. In the web-page level models, the training and prediction instances correspond to the unit length normalized feature vector of individual web-pages. During prediction, the $\epsilon$-SVR models are used to estimate the probabilities of the different values of the demographic attribute for all the web-pages of a web-site and the web-site-level prediction is obtained by aggregating these web-page-level predictions.

### *E. Aggregating Web-Page Level Predictions*

We developed two different ways to aggregate the web-page level predictions. The first approach, assigns the same amount of importance to each page and computes the web-site level prediction as the (unweighted) average of the predictions obtained at the web-page level (*Avg*). The second approach, uses the number of external inlinks of each web-page as a measure of its importance and computes the web-site level prediction as the weighted average of the web-page predictions using weights derived from the relative number of inlinks. The motivation behind the second approach is that, in general, web-pages that are linked from other web-sites will be some of the first pages a user will visit (as a result of following the corresponding links) and as such they have a higher probability of being viewed by users than the web-pages that are not linked from external web-sites. Thus, the number of external inlinks can be considered as a surrogate of the number of times a web-page is being viewed relative to the other web-pages in that web-site.

We investigated two methods for assigning weights based on the number of external inlinks. The first, assigns a weight that is linear on the number of external inlinks (*LinkLin*), whereas the second, assigns a weight that is logarithmic on the number of external inlinks (*LinkLog*). For those web-pages that have no external inlinks, we investigated two different approaches for assigning weights to them. The first, assigns a weight of one to all such web-pages. The second, assigns a weight that is based on the number of external inlinks of its $k$ most similar web-pages (*LinkKnn*). The external inlink counts are aggregated using a smoothing factor $\alpha$ that controls the amount by which each neighbor influences the inlink count of the web-page. This is analogous to expressing the external inlink count in terms of the amount of traffic forwarded from each neighbor to the given web-page.

### *F. Converting Individual Predictions to Distributions*

The prediction framework that we described so far builds an $\epsilon$-SVR model to estimate the probability for each one of the discrete values of the demographic attribute under consideration. However, since these predictions $\{p_i, \ldots, p_k\}$ are computed independently of each other, they are not guaranteed to form a valid probability distribution. We address this problem by using a simple two-step approach to convert the individual predictions into probability distributions. First, we set to zero any predictions that are negative, and then we linearly scale the predictions so that their sum is one. Note that the above approach is only used for demographic attributes that take more than two values (e.g., age). For variables that take only two values (e.g., gender), we only train a single $\epsilon$-SVR model that is designed to predict the probability for one of those values. If $p_1$ is the prediction obtained by that model, then when $0 \leq p_1 \leq 1$, the probability of the other value is $p_2 = 1 - p_1$. When $p_1 < 0$, $\{p_1, p_2\} = \{0, 1\}$ and when $p_1 > 1$, $\{p_1, p_2\} = \{1, 0\}$.

### *G. Coupling the Individual Models*

A limitation of the above approach is that by estimating the probability for each value of the demographic attribute independently of each other, it fails to take into account certain correlations that exist among the different values of the attribute (i.e., user groups). To address this problem, we developed an approach that builds a second level model that uses as input the predictions obtained by the individual $\epsilon$-SVR models as follows. Let $P$ be a $n \times k$ matrix containing the web-site level predictions produced by the first-level models (using one of the two approaches described in Section II-D), where $n$ is the number of training web-sites and $k$ is the number of values of the discrete random variable under consideration (e.g., 5 for the age attribute). Also, let $A$ be another $n \times k$ matrix that contains the actual probability distributions of the $n$ web-sites in the same order as $P$. The goal of the second-level model is to estimate a $k \times k$ matrix $W$ that minimizes $||A - PW||$. Once $W$ is estimated, a web-site is predicted by first using the $k$ $\epsilon$-SVR models to estimate the probability for each value of the demographic attribute resulting in a $1 \times k$ matrix $p$, then the second model is applied to obtain the prediction $pW$, which is finally converted into a valid distribution using the method described in Section II-F. Matrix $W$ is estimated by using the Moore-Penrose method [8], [10] to obtain the pseudo-inverse $P^{-1}$ of the non-square matrix $P$ at which point $W = P^{-1}A$.

## III. Experimental Evaluation

### *A. Datasets*

The performance of the methods were evaluated on a set of 8,215 web-sites whose audience demographic information was obtained from Comscore and Quantcast. Information on how these web-sites were selected is provided in [6]. These web-sites were used to generate a number of datasets for evaluating different aspects of the methods that we developed. The first dataset, referred to as DS1, was generated by randomly selecting 100 pages from each of the 8,215 websites and is used as the primary dataset for evaluating the performance of the different methods. The second dataset, referred to as DS2, was generated by selecting the subset of web-sites that contained at least 100 pages whose length

Table I
DATA SET STATISTICS

| | DS1 | DS2 | DS3 |
|---|---|---|---|
| No. of web-sites | 8,215 | 3,602 | 7,912 |
| Avg number of pages/site | 100 | 500 | 100 |
| Avg no. of words in the *T* vector | 177 | 478 | 174 |
| Avg no. of words in the *TH* vector | 187 | 504 | 186 |
| Avg % of pages/site with inlinks | 11 | 10 | 10 |
| Avg max number of inlinks/site | 1,452 | 112 | 958 |

Table II
AVERAGE RMSE FOR DIFFERENT TYPES OF FEATURES (DS1).

| | Gender | | Age | |
|---|---|---|---|---|
| Features | Web-Page | Web-Site | Web-Page | Web-Site |
| T | <u>10.25</u> | 10.88 | <u>8.53</u> | 8.76 |
| TH | 10.50 | 11.48 | 8.59 | 8.90 |
| THS | 10.56 | 12.28 | 8.61 | 9.22 |
| Baseline | 12.64 | 12.64 | 9.34 | 9.34 |

Underlined entries correspond to the best performing scheme.

belonged in all the following intervals: 0-100, 100-200, 200-400, 400-800, 800-1600 words. Finally, the third dataset, referred to as DS3, was generated from DS1 by selecting the subset of web-sites (and associated web-pages) for which both Comscore and Quantcast provided values for the two demographic attributes. Various characteristics about these datasets are shown in Table I.

### B. Experimental Methodology

For all experiments, the data set was divided into five folds at the web-site level and five-fold cross validation was performed. The web-site level partitioning of folds ensures that the pages from a given web-site are never in both the training and the test sets. In order to speed up the process of training, instead of training on $k-1$ folds and testing on the remaining fold, we trained on each single fold and tested on the remaining $k-1$ folds. For the distribution prediction approaches based on the pseudo inverse method (Section II-F), matrix $W$ was estimated from $P$ by using a cross-validation approach during training [6]. The SVMlight [5] implementation of $\epsilon$-SVR was used to perform the learning and prediction. All the experiments were performed using a linear kernel function.

### C. Evaluation Metrics

The overall accuracy of the predicted demographic attribute (i.e., distribution) was measured using the root mean squared error (RMSE). The reported result is a percentage value (i.e., probabilities multiplied by 100) and corresponds to the averages over all the web-sites across the five-fold cross validation. The students $t$ test was used to assess the statistical significance of the results.

### D. Baseline Predictions

A simple scheme for predicting a demographic attribute is for each value (e.g., Teen for the age attribute) to compute its average probability over all the web-sites in the training set and use this as the predicted probability for the testing set. This *baseline* method is compared against the methods developed and evaluated in this work using the same five-fold cross validation splits while estimating the average training set probabilities.

## IV. RESULTS

### A. Performance of Different Features

Table II shows the performance achieved for the gender and age prediction tasks for the different features described in Section II-C. These results show that the simplest set of features (T), which corresponds to the web-page's term vector, achieves the best or close to the best results for both the web-page and the web-site level models. Moreover, any additional features that emphasize the set of terms that occur in the title and section HTML tags (TH) or incorporate information about the web-page's structure (THS) does not lead to any improvements.

However, an encouraging observation is that the actual prediction error (as measured by the average RMSE) is rather low. For the gender attribute, the best average RMSE value is 10.25 and for the age attribute, the best average RMSE value is 8.53. Moreover, these RMSE's are considerably lower that the corresponding values of 12.64 and 9.34 that were obtained by the baseline model. These results suggest that the content of the web-sites provide strong information for predicting the demographic characteristics of their audience.

### B. Performance of Model Granularity

Comparing the relative performance of the web-site and web-page level models shown in Table II, we see that the models trained at the web-page level achieve better results than the corresponding web-site level models. Moreover, the relative performance advantage of the web-page level models is quite substantial. These results suggest that by representing the web-pages as individual single training instances, the model is able to capture the web-site's overall characteristics better and achieve more accurate predictions. Due to this clear advantage of the web-page models, the rest of the results in this section will focus on web-page models.

### C. Performance of Different Web-page Lengths

In a typical web-site, the length of each web-page (as measured by the number of words) often varies from tens to thousands of words. To investigate the impact of the web-page length on the quality of the models and their associated predictions, Table III shows the performance achieved by models trained and applied on different length web-pages. These experiments were performed using the DS2 dataset, which was specifically designed for that purpose.

There are two primary observations that can be made by analyzing the results in this table. First, the quality of the models learned does not improve by using training web-pages that have a large number of words. The best (or closed to best) performance is usually achieved by the model that is

Table III
AVERAGE RMSE FOR DIFFERENT WEB-PAGE LENGTHS (DS2).

| Length of training web-pages | Lengths of testing web-pages | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 0–100 | | 100–200 | | 200–400 | | 400–800 | | 800–1600 | |
| | Gender | Age | Gender | Age | Gender | Age | Gender | Age | Gender | Age |
| 0–100 | 11.01 | 8.49 | 10.30 | 8.28 | 10.08 | 8.24 | 9.92 | 8.20 | 9.80 | 8.16 |
| 100–200 | 11.08 | 8.50 | 10.29 | 8.25 | 10.05 | 8.20 | 9.88 | 8.16 | 9.75 | 8.12 |
| 200–400 | 11.13 | 8.53 | 10.36 | 8.27 | 10.10 | 8.20 | 9.91 | 8.15 | 9.77 | 8.10 |
| 400–800 | 11.18 | 8.55 | 10.47 | 8.31 | 10.20 | 8.23 | 10.00 | 8.16 | 9.83 | 8.12 |
| 800–1600 | 11.28 | 8.57 | 10.59 | 8.35 | 10.33 | 8.27 | 10.11 | 8.20 | 9.91 | 8.13 |

Underlined entries correspond to the best performing scheme along each column.

Table IV
AVERAGE RMSE OF DIFFERENT AGGREGATING SCHEMES (DS1).

| Experiment | Gender | Age |
|---|---|---|
| Average (Avg) | 10.25 | 8.53 |
| Log Scheme (LinkLog) | 10.21 | 8.51 |
| Linear Scheme (LinkLin) | 10.00 | 8.42 |
| LinkLin+ LinkKnn($\alpha = 0.10$ and $k = 15$) | 9.97 | 8.41 |

trained using web-pages containing between 100–200 words irrespective of the size of the web-pages used in the testing set. In fact, the relative performance of the models trained on longer web-pages actually degrades. This can potentially be attributed to the fact that due to the higher dimensionality of the longer documents, the models learned may suffer from overfitting. The second observation is that, for both attributes the quality of the prediction improves as longer (testing) web-pages are used to predict the demographic attribute under consideration. A potential reason as to why longer testing documents are better may be due to the fact that by virtue of their length they better cover the web-site's content and as such they can better utilize the models that were learned to relate the web-site's content with the different demographic attributes.

### D. Evaluation of Prediction Aggregation Methods

Table IV shows the performance of the different aggregation techniques discussed in Section II-E. These results show that for both demographic attributes, the use of inlink information during aggregation leads to prediction improvements, with the linear-weighting scheme outperforming both the simple averaging and the log-weighting schemes.

Figure 1 shows the results for using the $k$-nn smoothing technique for web-pages with no inlinks (Section II-E). The experiments were performed for different values of $k$ and linear scheme of aggregation is used to aggregate the web-page level predictions using smoothened inlink counts. Looking at the graphs we can see a pattern where the RMSE value for both gender and age prediction initially decreases as we increase $\alpha$ and then increases. In particular the best RMSE achieved is 9.97 and 8.41 for gender and age respectively when $k = 15$ and $\alpha = 0.10$ (last line of Table IV). This shows that including neighboring web-page inlink information for web-pages having no inlinks further improves the results. Moreover, it also indicates that RMSE is sensitive to both $k$ and $\alpha$. Fine tuning the values of both $k$ and $\alpha$ helps to achieve the best results.

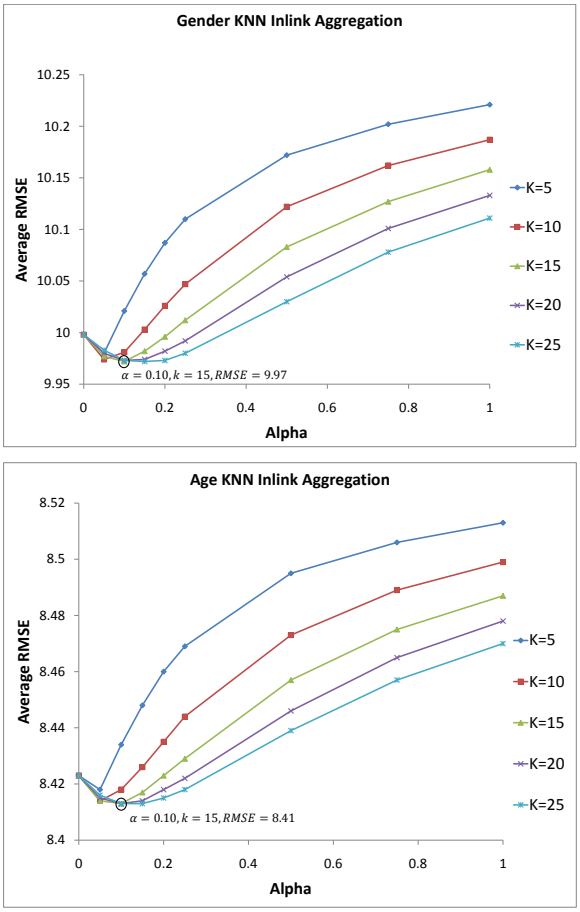


Figure 1. Average RMSE of Gender and Age for $k$-NN smoothing of web-pages with no inlinks.

### E. Performance of the Second Level Model

The average RMSE achieved by the approach that uses the second level model (Section II-G) to couple the predictions obtained by the individual $\epsilon$-SVR models for the age demographic attribute is 8.26 (Table V). The second level model was built using the predictions from the best performing model (web-page level model coupled with aggregation done using linearly weighted inlinks scheme with $\alpha = 0.10$ and $k = 15$). Comparing these results we see that the use of the second-level models leads to performance improvements. Note that the second level model cannot be applied to the gender attribute as it only takes two values.

Table V
AVERAGE RMSE OF SUCCESSIVELY IMPROVED MODELS (DS1).

| Model | Gender | Age |
|---|---|---|
| Baseline | 12.64 | 9.34 |
| $\epsilon$-SVR with prediction averaging | 10.25 | 8.53 |
| $\epsilon$-SVR with inlinks & smoothing | 9.97 | 8.41 |
| $\epsilon$-SVR with inlinks & smoothing & 2nd level model | - | 8.26 |

Underlined entries correspond to the best performing model.
The performance of each successive model is statistically significant than the preceding model at $p < 0.005$.

Table VI
AVERAGE RMSE OF COMPARISON WITH COMSCORE AND QUANTCAST DATA (DS3).

| | Average RMSE | |
|---|---|---|
| | Gender | Age |
| Comscore vs Quantcast | 9.74 | 8.87 |
| Panopia vs Comscore | 9.97 | 8.43 |
| Panopia vs Quantcast | 6.20 | 6.42 |

### F. Comparison with Comscore and Quantcast Predictions

The ground-truth information about the gender and age demographic attributes of the different web-sites correspond to estimates that were obtained using different methods (e.g., user panels and/or tracking cookies). As a result, the demographic attribute information obtained from different sources is expected to be different. We used the DS3 dataset to compare the ground-truth information obtained from Comscore and Quantcast with each other and also against the predictions obtained by our best model. Table VI shows three sets of RMSE values. The first set shows the average RMSE between the Comscore and Quantcast values for the gender and age demographic attributes. These RMSE values indicate that there is a considerable degree of disagreement between the two companies as to the distributions of these attributes. These differences can be attributed to varying data collection methodologies employed by them and indicates an inherent degree of uncertainty or error in the estimations. The second and third set show the RMSE of the predictions produced by our best model (underlined entries in Table V), referred to as *Panopia*, when compared to Comscore and Quantcast, respectively. These results show that the RMSE values between Panopia and Comscore are comparable to the corresponding values between Comscore and Quantcast (higher RMSE for gender and lower for age), whereas the RMSE values between Panopia and Quantcast are considerably lower than the corresponding RMSE values between Comscore and Quantcast. Overall these comparisons are very encouraging, as they indicate that once the inherent differences between sources as to what are the ground-truth distributions is taken into account, the predictions produced by our methods are quite good.

## V. CONCLUSION

In this paper we developed and studied regression-based methods for predicting demographic attributes for web-sites that do not rely on any personal and behavioral information. The successively more complex models that we developed (whose performance is summarized in Table V) are able to achieve increasingly better results, with the best models achieving an RMSE of 9.97 and 8.26 for the gender and the age demographic attributes, respectively. These RMSE values represent a 21.1% and 11.2% improvement of the corresponding RMSE values of the baseline model and are significant at $p < 10^{-5}$. Moreover, the RMSE values obtained by our methods are comparable to the RMSE values between the ground-truth information provided by different commercial sources. These results indicate that content-based information can be used quite effectively for predicting the demographic attributes of web-sites without relying on any information that can potentially be intruding on users' privacy. In addition, our study showed that based on the characteristics of the web-pages, different strategies can be utilized that build and use different models during prediction (e.g., a T- and TH-based model) or select longer and more inlinked web-pages to compute the predictions that can lead to further improvements in accuracy.

## REFERENCES


[1] Adar, E., Adamic, L., Chen, F.: User profile classification by web usage analysis. In: US Patent Publication number 2007/0073682 A1 (2007)
[2] Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. Addison Wesley Longman Publishing Co. Inc (1999)
[3] http://www.comscore.com/
[4] Hu, J., Zeng, H.J., Li, H., Niu, C., Chen, Z.: Demographic prediction based on user's browsing behavior. In: WWW '07: Proceedings of the 16th international conference on World Wide Web. pp. 151–160. ACM, New York, NY, USA (2007)
[5] Joachims, T.: Text categorization with suport vector machines: Learning with many relevant features. In: ECML '98: Proceedings of the 10th European Conference on Machine Learning. pp. 137–142. Springer-Verlag, London, UK (1998)
[6] Kabbur, S., Han, E.H., Karypis, G.: Content-based methods for predicting web-site demographic attributes. Tech. rep., Department of Computer Science, University of Minnesota (2010)
[7] Michailidou, E., Harper, S., Bechhofer, S.: Visual complexity and aesthetic perception of web pages. In: SIGDOC '08: Proceedings of the 26th annual ACM international conference on Design of communication. pp. 215–224. ACM, New York, NY, USA (2008)
[8] Moore, E.H.: On the reciprocal of the general algebraic matrix. Bulletin of the American Mathematical Society 26, 394–395 (1920)
[9] Murray, D., Durrell, K.: Inferring demographic attributes of anonymous internet users. In: Lecture Notes in Artificial Intelligence. pp. 7–20. Springer-Verlag (1999)
[10] Penrose, R.: A generalized inverse for matrices. In: Proceedings of the Cambridge Philosophical Society. vol. 51, pp. 406–413 (1955)
[11] http://www.quantcast.com/
[12] Vapnik, V.: The Nature of Statistical Learning Theory. Springer-Verlag (1995)
[13] Zhang, B., Dai, H., Zeng, H., Qi, L., Najm, T., Mah, T., Shipunov, V., Li, Y., Chen, Z.: Predicting demographic attributes based on online behavior. In: US Patent Publication number 2007/0208728 A (2007)