Technical Report

Department of Computer Science
and Engineering
University of Minnesota
4-192 EECS Building
200 Union Street SE
Minneapolis, MN 55455-0159 USA

TR 10-021

Content-Based Methods for Predicting Web-Site Demographic Attributes

Santosh Kabbur, Euihong Han, and George Karypis

September 17, 2010

Content-Based Methods for Predicting Web-Site Demographic Attributes

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

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

George Karypis
Department of Computer Science
University of Minnesota
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], [5], [6] 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 utilizing 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 work makes a number of contributions that provide insights on how privacy-preserving information can be best used to predict demographic attributes and on the methods for deriving these predictions. Specifically, our work shows that:

- The set of words that are present in a web-page is the most important feature and that the incorporation of features that takes into account the structure of the web-page does not lead to any improvements.
- Incorporating words that occur in title and sectioning-defining HTML tags as additional features tend to improve the performance of the models, when these tags are used extensively within the web-pages of a web-site.
- Projecting the demographic attributes of a web-site to its individual web-pages and constructing the models using individual web-pages as training instances outperforms models in which the training instances correspond to entire web-sites.
- The length of the web-pages (in terms of the number of words) used to train the models and derive the predictions plays an important role. For training, shorter web-pages tend to perform better, whereas for deriving the predictions, longer pages perform better.
- Good prediction performance is achieved by using a relatively small number of pages for each web-site and that the prediction performance does not improve by using more web-pages.
- The intra-website links can be used to improve the prediction accuracy by emphasizing the pages that are linked from external web-sites and as such have a higher probability of being visited by the web-site's audience.
- For demographic attributes that correspond to discrete random variables with more than two values, considerable improvements can be obtained by using a second-level model to learn a probability mass function that couples the single-value predictions.

Our experimental evaluation shows that compared to ground-truth data obtained from Comscore [14], 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[15]), shows that the performance of our models is comparable to the differences among their own sets of predictions.

II. Related Work

Previous research on demographic attribute prediction has focused on methods that take into account user's web usage pattern. Adar et al [6] focused on predicting the demographic attributes of users based on the analysis of accessed web-pages. Web pages accessed by the users were used to construct feature vectors and vector similarity measure was used to calculate the similarity between users. Nearest neighbor approach was used to predict the demographic attributes of unknown users. They also proposed an approach which assigned bias values to web-pages for each demographic attribute based on the known users who have accessed those web-pages. The demographic attributes of an unknown user were then predicted based on these bias values of the web-pages accessed by the user. Hu et al [1] developed a different set of methods to predict the demographic attributes of users. They made use of web-page click through data and a set of known users to learn a model that associates web-pages with users. Their models used two different kinds of features that were derived from the content of the web-pages and the categories to which they belonged. They used this classification model to predict the demographic attributes of unknown users using a Bayesian framework. In addition, they used user browsing data to determine the similarity between users and web-pages. Then, for an unknown user, they extracted the features from the recently accessed web-pages of the user to predict the demographic attributes. For demographic attributes that take more than two values (e.g., age), they used multiple classifiers to classify each user to each of the age groups. In a follow-up work [4], they extended their method to also include search terms entered by users in a search engine and the snippets from the search results page as features along with the content and category-based features. In another research work, Murray and Durrell [5] used search terms entered and web-pages accessed by the user to predict the demographic attributes of the user. They constructed a reduced vector space using Latent semantic analysis (LSA) to represent the web-usage data of the known users. Then they used this vector space as input to build a neural network based model and subsequently used the model to predict the demographic attributes of unknown users. Similar to other research studies, they use k-way classification techniques to predict demographic attributes that take more than two values.

A common characteristics to all of these methods is that besides information about the content of the various web-sites, they also utilize different types of users' web-usage data for deriving their predictions (e.g., web-page click-through data, search terms, etc.). In contrast, the methods developed in this paper are based entirely on information that can be extracted from the different web-pages and are designed to operate in the absence of any user-supplied (or derived) information. In addition, all of these methods

formulate the demographic attribute prediction problem as a k-way classification problem. In contrast, our formulation treats the problem as that of predicting a probability distribution, which we believe better captures the problem's underlying characteristics.

III. Methods

A. Demographic Attributes

Even though the methods developed in this paper can be used to predict a wide range of demographic attributes, the focus in this paper is predicting the gender and age distribution of a web-site's audience. The choice of these two demographic attributes was primarily dictated by the availability of data, as there are a number of services providing the gender and age distributions of the audience for various web-sites. 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 $P(X = x)$. Note that in order for these predictions to correspond to a valid probability distribution, $\sum_{x \in \mathcal{S}_{\mathcal{X}}} P(X = x) = 1$.

The approach that we investigated for predicting a demographic attribute follows a supervised learning framework. Within this framework, a set of web-sites with known probability distributions for the demographic attribute under consideration are used as the training set, a set of features for these web-sites is extracted, and a model is learned to predict the probability distribution of the demographic attribute based on these features.

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. This is different from most of the traditional value estimation problems that focus on building models to estimate a single value. For this reason, the model learning and associated prediction methods that we developed consist of two steps. First, regression techniques are 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. This second-level model can be considered as learning the probability mass function of the corresponding random variable.

The individual regression models are estimated using ϵ-support vector regression (ϵ-SVR) [7], whereas the individual estimations are coupled using an approach based on matrix approximation. Details about the different features that we investigated, the granularity at which the individual regression models were learned and applied, and the methods used to obtain the probability distributions are provided in the remainder of this section.

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). Both of these features are determined entirely by analyzing the web-page itself and they do not rely on any information about the users visiting the corresponding web-pages and web-sites. This was done by design, as one of the primary goals of this work is to develop methods that can accurately predict the demographic characteristics of a web-site's audience without relying on any data that directly or indirectly intrudes on a user's private information.

1) Web-page's Textual Content: To represent the web-site's textual content, we used the popular vector-space model from information retrieval [10]. In this model, each web-page is represented as a sparse vector in the space of the distinct terms that exist in the collection. The non-zero entries of that term vector correspond to the terms that are present in the web-page. We used the standard TF-IDF term weighting scheme that assigns a weight to each term that is linearly related to its occurrence frequency in the web-page and inversely related to the number of web-pages in which it occurs. Note that the size of the document collection (i.e., web-pages) used in the IDF component is equal to the number of web-pages across the entire set of web-sites. Following standard information retrieval practices, we used a stop words list to eliminate certain unimportant words and used 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.

A challenge associated with extracting the textual content of modern web-pages is that in addition to the web-page specific content, they also contain information that is irrelevant to its content. Such examples include headers, footers, navigation panels, and advertisements. Quite often, the portion of a web-page's text that is directly related to its

specific content is much smaller than that occupied by the irrelevant content. To address this problem, we developed a technique that identifies the web-page's specific content by analyzing the entire set of pages that were obtained from the same web-site. This technique is related to the template identification methods used by web search engines to determine the parts of the web-page that they index [2]. Specifically, given the set of web-pages that belong to the same web-site, our method constructs the DOM tree of all these pages and eliminates all the paths from the leaves to the root of the DOM tree that occur in at least a certain number of DOM trees (i.e., web-pages). The term vector of a web-page was generated from the unpruned leaf nodes text of its DOM tree. The motivation behind this approach is that elements of the web-page that are common across different pages will correspond to non web-page specific content and as such they can be eliminated. In our experiments, we used 15 as the leaf-node frequency above which a node will be pruned in order to account for the cases in which different web-page templates are used to generate the pages in a particular web-site.

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. This concatenation results in a doubling of the feature space. Since this concatenation operation involves vectors of unit length, the two sets of features contribute equally to the web-page's final representation. Note that the title and header tags were extracted only from the parts of the web-page's DOM tree that were not pruned. We will refer to this as the *TH* representation of the web-page.

2) Web-page's Structure: The set of features that we extracted were designed to capture the web-page's structure in terms of its style and organization. The main focus was on representing the structure of the web-page in terms of its complexity. As shown in [11], the visual appearance of a web-page greatly influences the way the user interacts with the web-page and the type of users that it attracts. As a result, the existence of certain structural elements can provide valuable clues as to the demographics of its users (or its intended users). Based on the various factors that influence the web-page's complexity given in [11], we extracted and used the following information from each web-page: (i) the number of different visual blocks, (ii) the number of hyper links, (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. These features were extracted from the entire web-page and not from the portions of the web-page that were used to derive textual features (Section III-C1). These structural features were used as additional features to augment the term-vectors extracted from the web-page's specific content. This was done by creating a new vector consisting of the structural features, normalizing it to be of unit length and then appending to the vector of the web-page's *TH* representation. 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. Besides its simplicity, this approach has the additional advantage of being fast, as the number of training set instances is equal to the number of web-sites. In the web-page level models, the training and prediction instances correspond to the unit length normalized feature vector of individual web-pages. During training, each web-page inherits the probability distribution of the demographic attribute of the web-site to which it belongs. During prediction, the ϵ-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. These web-page level predictions are then aggregated to obtained the prediction at the web-site level.

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. The second approach, uses the number of external inlinks of each web-page[1] 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

[1]An inlink is considered to be external if the citing web-page belongs to a different web-site.

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, whereas the second, assigns a weight that is logarithmic on the number of external inlinks. 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. Specifically, given a web-page with no external inlinks, its k nearest neighbors with external inlinks are determined and their inlinks counts are aggregated to get an external inlink count. The external inlink counts are aggregated using a smoothing factor α 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. Using this approach, the external inlink count of the ith web-page is $(\sum_{j=1}^{k}(\alpha m_j))/k$, where m_j is the inlinks count for jth nearest neighbor of ith web-page. Note that the similarity between web-pages was computed as the cosine of their *T* representation.

F. Converting Individual Predictions to Distributions

The prediction framework that we described so far builds an ϵ-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 ϵ-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 may fail to take into account certain correlations that exist among the different values of the attribute (i.e., user groups). For example, if a web-site has a large fraction of kids, then it will most likely have a somewhat larger fraction of teenagers as they share some common interests (at least among the users that are at the boundary of the age breakdown). To address this problem, we developed an approach that builds a second level model that uses as input the predictions obtained by the individual ϵ-SVR models. This approach was motivated by approaches used to build cascading classifiers that are used extensively in bioinformatics [8], [9].

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 III-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 ϵ-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 III-F.

Matrix W is estimated by using the Moore-Penrose method [12], [13] to obtain the pseudo-inverse P^{-1} of the non-square matrix P at which point $W = P^{-1}A$.

IV. Experimental Evaluation

A. Datasets

The performance of the methods were evaluated on a set of web-sites whose audience demographic information was obtained from Comscore and Quantcast. These two companies are the leading commercial providers of demographic information for web-sites and they collect their data by contacting user-surveys via online panels and by tracking users across different websites. The set of web-sites were selected as follows. First, the top 50,000 web-sites from Alexa's one million most visited web-sites [16] was selected and were crawled in a breadth-first fashion. The set of crawled pages was subsequently pruned to eliminate pages with less than 100 words. Furthermore, any web-sites with fewer than 100 pages remaining were also eliminated. Note that a web-site can have a small number of pages because either the crawler failed to fetch (e.g., pages generated by scripts that the crawler could not handle) or the pages fetched contained a small number of words. These steps reduced the total number of web-sites to 8,215. Figure 1 shows how Comscore's gender and age demographic information is distributed across the different web-sites, respectively. These plots show the percentage of web-servers for which the value of a demographic attribute falls within the corresponding probability bin indicated by the x-axis.

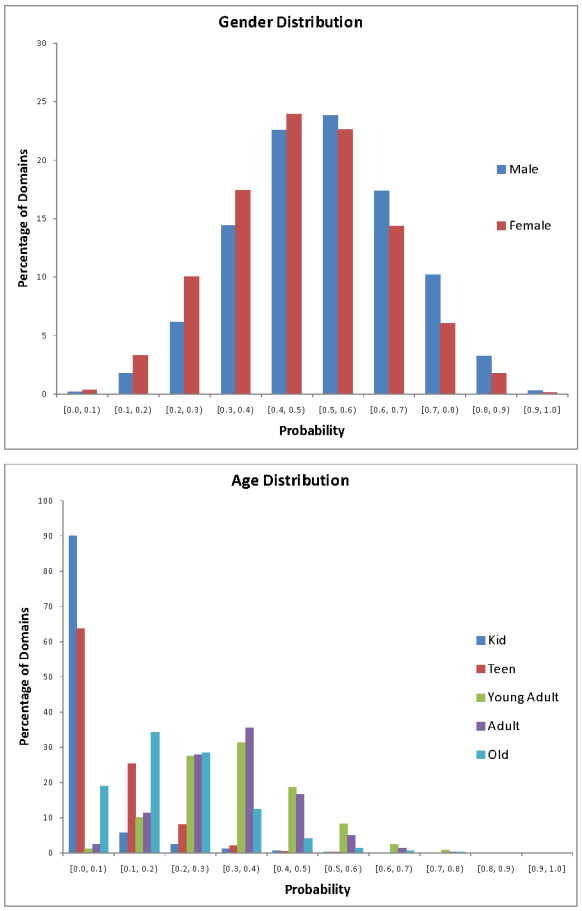

Figure 1. Distribution of Gender and Age Across the Web-sites

The above 8,215 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 first selecting the subset of web-sites and their associated web-pages that contained at least 100 web-pages in which the HTML title and sectioning tags contained at least 50 words, and then randomly selecting 100 web-pages from each of them. The third dataset, referred to as DS3, was generated by selecting the subset of web-sites that contained at least 100 pages whose length belonged in all the following intervals: 0-100, 100-200, 200-400, 400-800, 800-1600 words. Note that in order to get web-pages whose length falls in the first interval, the pages from the 8,215 web-sites that were initially eliminated due to small length, were included back in DS3. Finally, the fourth dataset, referred to as DS4, 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. Note that the DS1, DS2, and DS4 contain 100 web-pages per web-site, whereas the DS3 dataset contains 500 web-pages per web-site (100 for each interval). Various characteristics about these datasets are shown in Table I.

Table I
DATA SET STATISTICS

	DS1	DS2	DS3	DS4
No. of web-sites	8,215	2,475	3,602	7,912
Avg number of pages/site	100	100	500	100
Avg no. of words in the T vector	177	212	478	174
Avg no. of words in the TH vector	187	274	504	186
Avg % of pages/site with inlinks	11	10	10	10
Avg max number of inlinks/site	1,452	142	112	958

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. We used a small variation of standard k-fold cross validation. Instead of training on $k-1$ folds and then testing on the remaining fold, we trained on each of single fold and tested on remaining $k-1$ folds. This was done to speed the process of training as our data-set contains a large number of web-pages.

For the distribution prediction approaches based on the pseudo inverse method (Section III-F), matrix W was estimated from P by using a cross-validation approach during training. Specifically, the training set was itself split into five folds and each four-size subset of these folds was used to estimate the ϵ-SVR model and predict the left-out fold. The resulting set of predictions formed matrix P and was used to estimate W. During the actual prediction, a web-site was then predicted using the five different ϵ-SVR models that were estimated during the within-training five-fold cross validation, their predictions were averaged and then matrix W was used to predict the final distribution.

The SVMlight [3] implementation of ϵ-SVR was used to perform the learning and prediction. All the experiments were performed using a linear kernel function.

C. Evaluation Metrics

We used two different metrics to measure the performance of the predictions computed by the different methods. The first measures the accuracy of the overall predicted demographic attribute (i.e., distribution), whereas the second measures the accuracy of the individual values of the attribute (i.e., probability). The accuracy of the distribution was measured using the root mean squared error (RMSE) that is given by

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (\hat{x}_i - x_i)^2}{n}},$$

where $\hat{x}$ and x are the actual and predicted distributions of a web-site, respectively, and n is the length of these vectors

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

	Gender		Age	
Features	Web-Page	Web-Site	Web-Page	Web-Site
T	10.25	10.88	8.53	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.

(i.e., the number of values that the underlying discrete random variable takes). The accuracy of the prediction for a specific value of a demographic attribute was measured using the absolute error (AE) that is given by $AE = |\hat{\theta} - \theta|$, where $\hat{\theta}$ and θ are the actual and predicted values, respectively. For both these metrics, 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. Also, 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.

V. RESULTS

The experimental evaluation of the methods is done in two phases. First, we use the age and gender distributions obtained from Comscore to evaluate the different parameters of our methods (web-page features, granularity of the models and the use of second-level models). Second, we analyze the predictions obtained by our methods and those provided by both Comscore and Quantcast in order to assess their overall quality in light of the fact that there is an inherent error associated with the ground-truth data.

A. Performance of Different Features

Table II shows the performance achieved by our methods for the gender and age prediction tasks for the different features described in Section III-C. This table shows the average RMSE achieved by the T, TH, and THS features for both the web-page and web-site level models (Section III-D) as well as the average RMSE values obtained by the *baseline* model (Section IV-D).

From these results we can see 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.

Table III
AGE TENDENCY PREDICTION RESULTS AT WEB-PAGE LEVEL (DS1).

		Average Absolute Error				
Model Granularity	Features	Kid	Teen	Young Adult	Adult	Old
Web-Page	T	3.32	5.02	9.18	8.53	7.89
	TH	3.44	5.14	9.22	8.55	8.01
	THS	3.38	5.13	9.26	8.57	8.02
Web-Site	T	3.16	5.05	9.51	8.74	8.17
	TH	3.18	5.20	9.65	8.82	8.39
	THS	3.11	5.58	9.94	9.05	8.84
Baseline		4.50	7.30	10.86	9.86	10.11

Table IV
AVERAGE RMSE OF HTML AND TEXT FEATURES (DS2).

Features	Gender	Age
H	10.23	8.21
T	10.23	8.25
TH	10.15	8.22

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.

Table III further analyzes the prediction results for the age prediction task by showing the average AE for each of the five age groups. These results show that the errors achieved for each of the age groups does vary across the age groups, with the "Young Adults" achieving the worse AE of 9.94 and the "Kid" group achieving the lowest of 3.11. However, even in the case of the worst performing age group, the actual AE is low and better than that obtained by the baseline approach.

1) Performance of HTML Features: A potential reason as to why the TH representation does not lead to substantially better performance over the T representation is that the portion of a web-page's text that exists within the relevant HTML tags (i.e., title and section tags) may be too small. In order to test this hypothesis, we evaluated the T and TH representations on the DS2 dataset that was specifically created to contain web-pages with at least 50 HTML-derived words (Section IV-A). Table IV presents the results of this experiment for the T and TH representations and also a representation that used only the text from the HTML tags (H) and the web-page-based models. Comparing the average RMSE of these models we can see that when a sufficiently large number of HTML features are present in each web-page, then the TH representation leads to consistent improvements over the T representation. Moreover, for web-pages with a large number of HTML-derived words,

the performance achieved by the H representation alone is comparable or better to that achieved by the T representation.

B. Performance of Model Granularity

The results of Table II also provide information as to the performance of the two levels of granularity at which the models can be learned and applied. Comparing the relative performance of the web-site and web-page level models, 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 better capture the web-site's overall characteristics 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 that it contains) 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 prediction, we performed the following experiment. We used the DS3 dataset, which contains 100 pages whose length belonged in all the intervals of the set $\{0-100, 100-200, 200-400, 400-800, 800-1600\}$. For each interval we created a dataset that contained only the web-pages belonging to that interval. These datasets were used to evaluate the performance of the prediction method using a five-fold cross-validation approach. Note that for each length interval, the five-fold split was the same across the 3602 web-sites; thus, the different models were trained and tested on exactly the same subsets of web-sites. The results of these experiments for the T representation are shown in Table V and correspond to the table's "diagonal entries" (i.e., the entries in which the column length interval matches the row length interval). In addition, the table also shows the performance that is achieved when the model built on web-pages that belong to a certain length interval is used to predict the datasets containing web-pages from other length intervals.

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 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 the quality of the predictions improves as longer (testing) web-pages are used to predict the demographic attribute under consideration. For example, using the model trained on web-pages with 100–200 words, the RMSE of the gender predictions decreases from 8.50 when estimated using web-pages of length 0–100 to 8.12 when it is estimated by using web-pages of length 800–1600. A similar trend can be observed for the age attribute. A potential reason as to why longer testing documents are better may be 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.

Table VI
AVERAGE RMSE FOR DIFFERENT NUMBER OF WEB-PAGES IN THE TRAINING SET (DS3).

Number of web-pages	Testing Data					
	100		200		300	
Training Data	Gender	Age	Gender	Age	Gender	Age
100	10.18	8.21*	10.17	8.21*	10.18	8.21*
200	10.03*	8.16*	10.03*	8.16*	10.03*	8.17*
300	9.98	8.14	9.97	8.14	9.98	8.15

Underlined entries correspond to the best performing scheme along each column.
The entries marked with * correspond to schemes that perform statistically worse than the best scheme along each column at $p < 0.01$

D. Performance of Different Training and Testing Set Sizes

Table VI shows the performance achieved by the prediction models (using the T representation) when they are trained using different number of web-pages from each web-site. Specifically, these experiments were performed by generating three different datasets from DS3 by randomly selecting 100, 200, and 300 web-pages with more than 100 words from each web-site. The performance of these models were evaluated using testing sets derived from each of the three datasets. These evaluations were done using a five-fold cross validation framework in which the underlying set of web-sites were split into five folds in exactly the same way.

These results show that, as expected, the performance of the models improves as the number of training web-pages increases. This trend is consistent for both demographic attributes and for all three testing sets. However, an interesting observation is that for each model, the performance achieved by each one of the three testing sets is nearly identical. This suggests that when it comes time to compute a prediction for a new web-site, there is almost no benefit in using a large number of web-pages, which has positive performance implications.

E. Evaluation of Prediction Aggregation Methods

The web-page level predictions need to be aggregated in order to derive a prediction for the demographic attribute at

Table V
AVERAGE RMSE FOR DIFFERENT WEB-PAGE LENGTHS (DS3).

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 VII
AVERAGE RMSE OF AGGREGATING MECHANISMS USING INLINKS COUNTS (DS1).

Experiment	Gender	Age
Average	10.25	8.53
Log Scheme	10.21	8.51
Linear Scheme	10.00	8.42
Linear scheme with $\alpha = 0.10$ and $k = 15$	9.97	8.41

the web-site level. In all of the results reported thus far, this aggregation was performed by simply averaging the predictions across the web-pages of the testing web-site. However, as discussed in Section III-E, the inter-website inlink information can be used to emphasize some web-pages over others. Table VII shows the gains that can be achieved by using such methods. These results show that for both demographic attributes, the utilization of such information lead to prediction improvements, with the linear-weighting scheme outperforming both the simple averaging and the log-weighting schemes.

Figure 2 shows the results for using the k-nn smoothing technique for web-pages with no inlinks (Section III-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 α 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 VII). 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 α. Fine tuning the values of both k and α helps to achieve the best results.

F. Performance of the Second Level Model

The average RMSE achieved by the approach that uses the second level model (Section III-G) to couple the predictions obtained by the individual ϵ-SVR models for the age demographic attribute is 8.26 (Table VIII). 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.

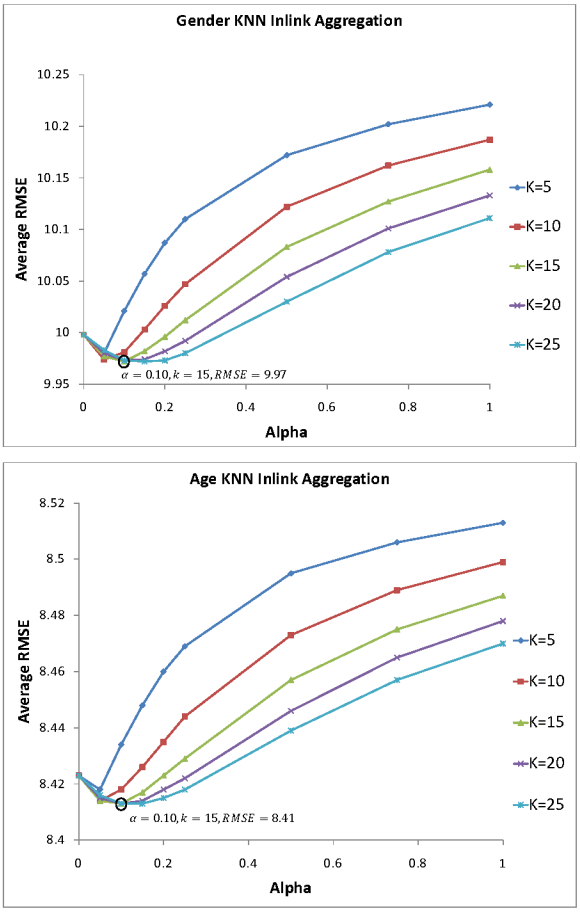

Figure 2. Average RMSE of Gender and Age for k-NN smoothing of web-pages with no inlinks.

Note that the second level model cannot be applied to the gender attribute as it only takes two values.

G. 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. In this section we compare the ground-truth information obtained from Comscore and Quantcast with each other other and also

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

Model	Gender	Age
Baseline	12.64	9.34
ϵ-SVR with prediction averaging	10.25	8.53
ϵ-SVR with inlinks & smoothing	9.97	8.41
ϵ-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 IX
AVERAGE RMSE OF COMPARISON WITH COMSCORE AND QUANTCAST DATA (DS4).

	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

against the predictions obtained by our best model that corresponds to the underlined entries in Table VIII. For ease of presentation, we will refer to our model as *Panopia*. These comparisons were performed on the DS4 dataset for which we have demographic information from both Comscore and Quantcast.

Table IX 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 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.

VI. 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 VIII) 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] Jian Hu , Hua-Jun Zeng , Hua Li , Cheng Niu , Zheng Chen, *Demographic prediction based on user's browsing behavior*, Proceedings of the 16th international conference on World Wide Web, May 08-12, 2007, Banff, Alberta, Canada.

[2] D Chakrabarti, R Kumar, K Punera *Page-level template detection via isotonic smoothing*, Proceedings of the 16th international conference on World Wide Web, 2007, pp61-70

[3] Joachims, T. *Text Categorization with Support Vector Machines: Learning with Many Relevant Features*, In Proceedings of the 10th European Conference on Machine Learning (ECML), Chemnitz, Germany, 137-142, 1998

[4] B Zhang, H Dai, HJ Zeng, L Qi, T Najm, TB Mah, V Shipunov, Y Li, Z Chen - Microsoft Corporation, *Predicting demographic attributes based on online behavior*. US Patent Publication number 2007/0208728 A1

[5] D. Murray and K. Durrell. *Inferring demographic attributes of anonymous internet users*. In Web Usage Analysis and User Profiling Workshop, volume 1836 of Lecture Notes in Computer Science, pages 7-20. Springer, 200

[6] E Adar, LA Adamic, FR Chen - Xerox Corporation, *User profile classification by web usage analysis*. US Patent Publication number 2007/0073682 A1

[7] Vladimir Vapnik. The Nature of Statistical Learning Theory. Springer-Verlag, 1995

[8] George Karypis. *YASSPP: better kernels and coding schemes lead to improvements in protein secondary structure prediction*, In Journal of Proteins, Aug 2006, Volume 64-3, pages 575-586

[9] Huzefa Rangwala and George Karypis, *Building multiclass classifiers for remote homology detection and fold recognition*, In Journal of BMC Bioinformatics, 2006, vol 7, page 455

[10] Ricardo Baeza-Yates and Berthier Ribeiro-Neto *Modern Information Retrieval*, Addison Wesley Longman Publishing Co. Inc.

[11] E. Michailidou, S. Harper, and S. Bechhofer. *Visual complexity and aesthetic perception of web pages*, In the 26th ACM International Conference on Design of Communication, SIGDOC08, Lisbon, Portugal on September 22-24, 2008

[12] Moore, E. H. (1920). *On the reciprocal of the general algebraic matrix*, Bulletin of the American Mathematical Society 26: 394-395.

[13] Penrose, Roger (1955). *A generalized inverse for matrices*, Proceedings of the Cambridge Philosophical Society 51: 406-413.
[14] Comscore - http://www.comscore.com/
[15] Quantcast - http://www.quantcast.com/
[16] Alexa Top Sites - http://www.alexa.com/topsites