Demographic Prediction Based on User's Browsing Behavior

Jian Hu, Hua-Jun Zeng, Hua Li, Cheng Niu, Zheng Chen
Microsoft Research Asia
49 Zhichun Road
Beijing 100080, P.R. China
{jianh, hjzeng, huli, chengniu, zhengc}@microsoft.com

ABSTRACT
Demographic information plays an important role in personalized web applications. However, it is usually not easy to obtain this kind of personal data such as age and gender. In this paper, we made a first approach to predict users' gender and age from their Web browsing behaviors, in which the Webpage view information is treated as a hidden variable to propagate demographic information between different users. There are three main steps in our approach: First, learning from the Webpage click-though data, Webpages are associated with users' (known) age and gender tendency through a discriminative model; Second, users' (unknown) age and gender are predicted from the demographic information of the associated Webpages through a Bayesian framework; Third, based on the fact that Webpages visited by similar users may be associated with similar demographic tendency, and users with similar demographic information would visit similar Webpages, a smoothing component is employed to overcome the data sparseness of web click-though log. Experiments are conducted on a real web click-through log to demonstrate the effectiveness of the proposed approach. The experimental results show that the proposed algorithm can achieve up to 30.4% improvements on gender prediction and 50.3% on age prediction in terms of macro F1, compared to baseline algorithms.

Categories and Subject Descriptors
I.5.2 [**Pattern Reorganization**]: Design Methodology-*Classifier design and evaluation*.

General Terms
Algorithms, Experimentation, Performance, Human Factors

Keywords
Demographic Prediction, Singular Value Decomposition, Supervised Regression, Browsing Behavior.

1. INTRODUCTION
Many general web services, such as search engines, websites and etc, start to pay more and more attentions to customized service for a better user experience. *My Yahoo!* [13] and *Google Personal* [7] are two good examples among these approaches. *My Yahoo!* allows users to build their preferences explicitly and only show sections and details which they may be interested in. Google Personal organizes users' search results according to their search histories including their previous search results and news headlines clicked. Accompany with the prosperous of general web service, online advertising is growing rapidly in recent years, in which behavioral targeting is becoming particularly popular [22]. Behavior targeting helps advertisers to target proper users upon their behaviors while surfing online. As reported in [18], companies like Tacoda Systems, Claria, Revenue Science and TM Advertising provide advertisers with behavioral targeting technologies. According to the recent studies of TM Advertising, compared to simple web ads, behavior-based ads gain 115% more business traffic a year, and the targeted consumers also scored 3% higher than the average viewers in brand awareness. User profile which includes prior search results, demographic information, geographic information and interested topics plays a key role in these systems to provide personalized targeting.

However, demographic information is usually not easy to obtain. Internet users are reluctant to expose this kind of personal data to public. The alternative way to predict users' demographic information is then of great interest to both industry and academia. In Koppel's work [11], blogers' writing styles are used to predict their actual gender and age information. However, only 8% internet users write blogs [29]. In contrast, the majority of users browse news, products, or other webpages through internet, which provides us a large number of web-page click-through log data.

Previous studies show that there is correlation between users' browsing behavior and their demographic attributes. As reported in "Computerworld" [3], 74% of women seek health or medical information online, while only 58% of men do so. 34% of women seek religious information from the Web versus 25% of men. Similar phenomena occur in movie domain, where demographic information correlates the genres of the movies the audiences appreciate. "Action for men", "love for women", or "cartoon for teenager" are common mappings between movie genre and audience demographic categories. So the diversity of the user's online browsing activities can be exploited to determine an unknown user's demographic attributes such as gender and age on the basis of user's online browsing activities.

In this paper we investigate the problem of predicting internet users' gender and age based on their browsing behaviors, in which the webpage view information is treated as a hidden variable to propagate demographic information between different users. The solution consists of three steps. Firstly, based on users' profiles and their browsing history, the user's age and gender information is propagated to the browsed pages, and then, a

WWW 2007, May 8–12, 2007, Banff, Alberta, Canada.
ACM 978-1-59593-654-7/07/0005.

supervised regression model is trained to predict a Webpage's gender and age tendency, i.e. the probability distribution of the ages and genders of a given Webpage's readers. Secondly, within Bayesian framework, an internet user's age and gender are predicted based on the age and gender tendency of the Webpages that he/she has browsed. Based on the error analysis, the prediction model resulted from the above two steps suffers from serious data sparseness, hence leaves much room for improvement. In the user browsing history data, many Webpages are browsed only by a few users and a significant portion of users are associated with a browsing history as short as a few pages. So both the Webpage demographic tendency prediction and the users' demographic prediction are not accurate. To deal with the noise and data sparseness problem, in Step 3, a smoothing approach is employed by making use of the fact that Webpages visited by similar users may be associated with similar demographic tendency, and users visiting similar Webpages may also share similar demographic attributes. First, latent semantic indexing is applied on user browsing data to derive the similarity among users and the similarity among Webpages. Then linear interpolation is used to combine content-category based demographic prediction and demographic attributes of similar Webpages and similar users.

Experiments are conducted on a real web click-through log to demonstrate the effectiveness of the proposed approach. The experimental results show that the proposed algorithm can achieve up to 30.4% improvements on gender prediction and 50.3% on age prediction in terms of macro F1, comparing with baseline algorithms.

The rest of the paper is organized as follows. In Section2, we present related work. In Section 3, we define the demographic prediction problem. In Section 4, we propose our solution for demographic prediction. The experimental results are shown in Section 5. Then we draw a conclusion and highlight future research directions in Section 6.

2. RELATED WORKS

In this section we briefly present some of the research literature related to demographic prediction.

Previous research on demographic prediction mainly focused on modeling the diversity of the linguistics writing and speaking styles associated with the demographic attributes. [4, 6] classified the user's gender by the spoken language difference including intentional, phonological and conversational cues. Mulac et al studied the gender difference in primary and secondary students' impromptu essays [13, 14]. Herring et al studied the gender difference in writing electronic communications [5]. Palander studied male and female styles in 17th century correspondence [15]. Biber investigated male and female difference in language structure using on correspondence corpus [2]. Berryman-Fink [1] and Simkins-Bullock[17] investigated the male and female writing styles in formal contexts such as books and articles, and asserted that no significant difference between male and female writing styles in such formal contexts. However, scant evidence indicated differences between male and female writing studied in these works are enough to be parlayed into an algorithm for categorizing an unseen text as being authored by a male or by a female [12].

Koppel [12] proposed to automatically categorize written texts by author gender. Based on a corpus from the British National Corpus, a simple Balanced Winnow algorithm is used with features including function words and parts-of-speech n-grams for author gender prediction. This model achieves classification accuracy of approximately 80%. After analysis of a corpus of tens of thousands of blogs, Koppel [11] found that there are significant differences in both writing style and content between male and female bloggers as well as among authors of different ages. Based on such difference on blog's content and style, they used the Multi-Class Real Winnow algorithm to learning models that classify blogs according to author gender and age, and obtain 80.1% accuracy on gender and 76.2% accuracy on age segmented in three categories (13-17, 23-27, 33-42).

These research works were mainly focused on classifying users' demographic attributes based on authorship. As far as we know, there is litter work on predicting users' gender or age according to what they browsed on the Web.

3. PROBLEM DEFINITION

Before introducing our technology for demographic prediction, we formulize the problem in this section.

The demographic attributes concerned in this paper include gender and age. We present a user's demographic attributes as two vectors *gender* and *age*. The gender prediction is defined as classifying users as male or female, while the age prediction is defined as classifying users into one of the following groups in Table 1.

Table 1. Age Group

Group	Age
Teenage	< 18
Youngster	18-24
Young	25-34
Mid-Age	35-49
Elder	>49

We define the browsing data as a set of records, where each record is a pair comprised of the user and the corresponding Webpages that the user viewed. So the browsing data can be modeled as a *weighted directed bipartite graph* $G=(V, E)$. A node in V represents a user or a Webpage, and each edge in E denotes that the user has clicked on the page. We can divide the nodes in V into two subsets, $U=\{u_1, u_2, \ldots, u_i\}$ and $W=\{w_1, w_2, \ldots, w_j\}$ where U represents the users and W represents the Webpages. A matrix R is used to represent the adjacency matrix, whose element r_{ij} in R is the weight from user u_i to Webpage w_j. In this paper, we simply deem the weight as the frequency of the Webpages being viewed by the user.

Given the webpage click-through log of some users with known demographic attributes, the problem is to find a general method to predict some users with unknown demographic attributes given their web-page click-through log.

4. DEMOGRAPHIC PREDICTION

One intuitive way for demographic prediction is to use Collaborative Filtering (CF) [8]. For a user with unknown gender/age, we could "recommend" the user's gender/age based on the users with similar online behavior. However, the webpage

click-through log is quite sparse (see in experiments part), while CF is quite sensitive to data sparseness [24]. Another simple way for demographic prediction is to train classifier in the user side directly. We can aggregate all the Webpages a user clicked as a document, and trained classifier in user side. Since different users have different tastes on different Webpages, the feature of users may contain much more non-discriminative features than that of Webpages. Directly training the classifier in user side will lead the poor performance of classification. In our experimental result, we also show that classifier on user side show lower performance.

In following subsections, we first predict a Webpage's gender and age tendency by training a supervised regression model based on user self reported gender, age and his/her browsing history. Then, based on the age and gender tendency of the Webpages that a user has browsed, we predict a user's gender and age within Bayesian framework. To solve the data-sparseness problem suffered in the above two steps, we propose an approach to make use of similarity relationship between users and Webpages.

4.1 Webpages' Demographic Tendency Prediction

4.1.1 Gender and age tendency of Webpages

Since Webpages don't have explicit demographic attributes, we can not simply label a Webpage as Male, Female or Teenage directly. Instead, we propose to predict the demographic distribution among the readers of a given Webpage, and here the demographic attributes of a Webpage are described as follows:

$$\Pr(c \mid w_j) = \sum_{i=1}^{I} r_{ij} u_i(c) / \sum_{i=1}^{I} r_{ij} \qquad (1)$$

Let $\Pr(c \mid w_j)$ be the probability of a demographic attribute c of the j^{th} Webpage, $u_i(c)$ be the value of the same attribute of the i^{th} user, and r_{ij} be the edges between the i^{th} user and the j^{th} Webpage. There are six demographic attributes for a Webpage w_j : *male*, *female*, *teenage*, *youngster*, *young*, *mid-age*, *elder*, and each have a real value $\Pr(c \mid w_j)$. For example, $\Pr(male \mid w_j)$ means male tendency of this Webpage. Obviously, the sum of $\Pr(male \mid w_j)$ and $\Pr(female \mid w_j)$ is 1, and the sum of $\Pr(teenage \mid w_j)$ to $\Pr(elder \mid w_j)$ is equal to 1.

4.1.2 Learning Gender and age Tendency of Webpages

To learn the gender and age tendency of Webpages, we need to select some pages for training. Since the gender and age tendency of a Webpage is based on the demographic distribution of the readers of this page, the demographic tendency of pages visited by few users is not reasonable. So we selected Webpages which are read by at least 10 users. Based on the demographic attributes of a Webpage computed by Equation 1, we use the linear form of Support Vector Machine (SVM) Regression [20] to learn the gender and age tendency of Webpages. For each attributes of gender and age, we learn a model separately. After we get the tendency value of each gender/age attributes learned from their models, we normalize their value within the range [0, 1] using max-min normalization [26], so that the sum of $\Pr(male \mid w_j)$ and $\Pr(female \mid w_j)$ is 1, and the sum of $\Pr(teenage \mid w_j)$ to $\Pr(elder \mid w_j)$ is equal to 1.

4.1.2.1 Support Vector Machine Regression

The Support Vector Machine (SVM) model is a powerful classification and regression method based on a solid theoretical foundation -- *structural risk minimization* [21]. The classification and regression performance is outstanding in practice.

In the linear kernel mode, an SVM constructs the hyper-plane that lies "close" to as many of the data points as possible. The decision function is $f(x)=<w \cdot x>+c$, where $<w \cdot x>$ is the dot product of the hyper-plane's normal vector w and the example's feature vector x and c is a constant vector. For an input vector x_i and its correct value y_i, the aim of SVM is to select a hyper-plane and threshold *(w, b)* so that we can get a hyper-plane w with small norm, while simultaneously minimizing the sum of the distances from our points to the hyper-plane, measured using Vapnik's ε -insensitive loss function:

$$|y_i - (w \cdot x_i + b)|_\varepsilon = \begin{cases} 0 & if\ |y_i - (w \cdot x_i + b)| \le \varepsilon \\ |y_i - (w \cdot x_i + b)| - \varepsilon & otherwise \end{cases} \qquad (2)$$

4.1.2.2 Features

For the purpose of learning age and gender tendency of Webpages, each selected document is represented as a numerical vector in which each entry represented the weight of a corresponding feature in some feature set. Two different kinds of potential distinguishing features can be considered: content-based features and category-based features.

Content-based features

We take the content words of the Webpages as the features. We first remove "stopwords" in the Webpages, and then do content words selection based on distribution grade (DG) of a Webpage on demographics attributes and Information Gain (IG) [19]. DG can be readily calculated on the basis of the variance coefficient, which normalizes the variance of a distribution by its mean. Taken the gender as an example, the calculation is as follows:

$$DG = \frac{1}{\bar{g}} * \sqrt{\frac{1}{2}((m-\bar{g})^2 + (f-\bar{g})^2)} \quad \text{where } \bar{g} = \frac{m+f}{2} \qquad (3)$$

The DG measures the variance on gender. The smaller the value, the more evenly the gender is distributed. The bigger the value, the more value the feature is for the training. In our work, we set the minimal DG to 1.3. On the pages selected by DG, we select the top 20000 terms sorted by their IG value as the feature set.

Category-Based features

As new content will emerge on the Web everyday which can not be covered by current model, we use a hierarchy of web concepts (or categories) to alleviate the problem. Base on Web concept hierarchy, we first use SVM to build a hierarchical classifier. Then, all the Webpages in the training data are classified into the concept hierarchy. Finally, based on the demographic attributes of Webpages in each category of the concept hierarchy, we can compute the demographic distribution of categories in the concept hierarchy. Since the first level of the concept hierarchy is too coarse for demographic prediction. For example, for the category "Health", the majority distribution of gender is female, but for the category "Health\Men", the

subcategory of "Health", the majority distribution of gender is male. We build the classifier at deeper category level. Based on the demographic distribution of categories in the concept hierarchy, for each Webpage, we can get the demographic distribution value of its top 3 classified categories, and use them as features.

4.2 Users' Demographic Prediction

Based on the age and gender tendency of the Webpages that a user has browsed, we use a Bayesian framework [30] to predict the user's demographic attributes. Suppose the pages a user clicked are independent, then

$$\begin{aligned}\Pr(c \mid u_i) &\propto \Pr(c \mid \{w\}) \\ &\propto \Pr(\{w\} \mid c)\Pr(c) \\ &\propto \prod_j^k \Pr(w_j \mid c)\Pr(c) \\ &= \frac{\prod_j^k \Pr(c \mid w_j)\Pr(w_j)}{\Pr(c)}\Pr(c) \propto \prod_j^k \Pr(c \mid w_j)\end{aligned} \quad (4)$$

Where $\{w\}$ is the collection of Webpage w_j that clicked by the user u_i, c is the attribute of gender (male or female) or age (teenage, youngster… elder), and $\Pr(c \mid w_j)$ can be got from the Webpages' gender/age tendency prediction..

4.3 Demographic Prediction by Leveraging Similarity among Users and Webpages

Since a user may click pages from different sites of different topics every day, to predict a user's gender and age by analyzing clicked pages history within a few days is not accurate enough. As people in the similar gender or age may have similar interests and preference, they might visit same or similar pages, thus we can assist the prediction of a user's gender/age through analyzing the gender/age of users with similar browsing behavior. Also in the Webpages side, through analyzing the gender and age tendency of Webpages visited by similar users, we can assist the prediction of a Webpage's gender/age tendency. However, a Web site may contain hundreds of thousands of Webpages, and the pages a user clicked are relatively few. Thus, finding the similar users or pages in this sparse data may bring much noise. As Latent Semantic Indexing (LSI) [28], which uses Singular Value Decomposition (SVD) as its underlying matrix factorization algorithm, has been proved useful to address the data sparseness problem in many recommender systems [24, 25]. The reduced orthogonal dimensions resulting from SVD are less noisy than the original data and capture the latent associations between the pages and users [24]. In our work, we also use SVD to produce a low-dimensional representation of original user-page space.

4.3.1 Singular Value Decomposition

SVD is a well-known matrix factorization technique that factors a $m \times n$ matrix R into three matrices as the following:

$$R = U \cdot S \cdot V' \quad (5)$$

Where, $U = (u_1, u_2, \ldots u_m)$ and $V = (v_1, v_2 \ldots v_n)$ are the matrices of the left and right singular vectors. The column vectors $u_i, 1 \le i \le m$ and $v_j, 1 \le j \le n$ are orthogonal. $S = diag(\sigma_1, \sigma_2, \ldots \sigma_{\min(m,n)})$ is the diagonal matrix of singular values which satisfy $\sigma_1 \ge \sigma_2 \ge \ldots \ge \sigma_{\min(m,n)} \ge 0$. By setting the smallest $\{\min(m,n) - k\}$ singular values in S to zero, the matrix R is approximated with a rank-k matrix and this approximation is best measured in reconstruction error. Theoretical details on matrix SVD can be found in [23].

We start with a user-page click matrix that is very sparse, we call this matrix *R*. To capture meaningful latent relationship, we first removed sparseness by filling out user-page click matrix. A constant based smoothing is used: For pages that a user does not visit, an intuitive and straightforward smoothing method is to replace the zero elements with a small constant c (*0<c<1*). That is, even a page *p* is not visited by user *u* in the data, and it is assumed that page *p* is in general visited by *u* with a small probability if u browses in the site. We also considered two normalization techniques: for the normalization in the user dimension, all the values corresponding with *u* are divided by a constant and the values sum to 1 after division for each user *u*; Normalization in the page dimension is similar. We found the formal approach to provide better results. After normalization, we get a filled and normalized matrix $\mathrm{R}_{\mathrm{norm}}$.

We factor the matrix $\mathrm{R}_{\mathrm{norm}}$ and obtain a low-rank approximation after applying the following steps:

1. Factor $\mathrm{R}_{\mathrm{norm}}$ using SVD to obtain *U, S* and *V*.
2. Reduce the matrix S to dimension *k*
3. Compute two resultant matrices: $U_k S_k^{1/2}$ and $S_k^{1/2} V_k'$, we denote $R_u = U_k S_k^{1/2}$ and $R_p = V_k S_k^{1/2'}$.

Based on the low-dimensional space of use and page sides (R_u and R_p), we compute the neighborhood of each user and page respectively, and then we use the demographic attributes of its neighbors to smooth the gender/age tendency learning in the page side and gender/age prediction in the user side.

4.3.2 Smooth Webpages' Demographic Tendency Prediction

There are two kinds of neighbors for a Webpage: one is the neighbors computed by the vector similarity (cosine similarity) in the reduce space R_p, and we denote this kind of neighbors as *nbr*; the other is the neighbors computed by the similarity of Webpage's content, and we denote this kind of neighbors as *nbp*. We use both of them to enhance Webpages' demographic tendency prediction.

Based on the top *N* most similar *nbr* neighbors of page w_i, we predict the gender/age tendency of page w_i using the Equation below:

$$\Pr(c \mid nbr(w_i)) = \frac{1}{N} \times (Pr(c \mid w_{nbi1}) + \ldots + Pr(c \mid w_{nbiN})) \quad (6)$$

Where $Pr(c \mid w_{nbi})$ is the gender/age tendency probability of the top i ($0 < i \le N$) neighbor.

Thus, we can smooth the gender/age tendency of page w_i by

$$\Pr(c \mid w_i)_{smoothed} = \alpha \times \Pr(c \mid w_i) + (1-\alpha) \times \Pr(c \mid nbr(w_i)) \quad (7)$$

Where, $\Pr(c \mid w_i)$ is the original gender/age tendency value learned by SVM regression, and α is the parameter to control the

influence of the page's gender/age tendency predicted by *nbr* neighbors.

Based on the top *M* most similar *nbp* neighbors of page w_i, we predict the gender/age tendency of page w_i using the Equation below:

$$\Pr(c \mid nbp(w_i)) = \frac{1}{M} \times (Pr(c \mid w_{nb1}) + \ldots + Pr(c \mid w_{nbM})) \quad (8)$$

Where $Pr(c \mid w_{nbj})$ is the gender/age tendency of the top j ($0 < j \le M$) neighbor.

Then, the Equation 7 can be extended into (9) as below:

$$\begin{aligned} \Pr(c \mid w_i)_{smoothed} &= (1-\alpha_1) \times \Pr(c \mid nbp(w_i)) \\ &+ \alpha_1 \times ((1-\alpha_0) \times \Pr(c \mid nbr(w_i)) \\ &+ \alpha_0 \times \Pr(c \mid w_i)) \end{aligned} \quad (9)$$

Where, α_0 and α_1 are used to balance influence of gender/age tendency probability based on *nbr* neighbors and influence of gender/age tendency probability based on *nbp* neighbors.

Obviously, the smoothing can be further changed into an iterative procedure where the smoothed Webpage demographic attributes will be used to update the neighborhood average, and then re-smooth the Webpage demographic attributes. In the later experiment, the iterative learning is processed until the demographic attributes of each page are stable.

4.3.3 Smooth Users' Demographic Prediction

Based on the smoothed gender/age tendency of Webpages, we can use the Equation 4 to predict the demographic attributes of users according to the Webpages they clicked. After that, similar as the gender/age prediction of Webpages, we compute the neighbors of each user u_i based on the reduced space R_u, and then select the top *T* most similar users. The equation used to predict users' gender/age is as below:

$$\Pr(c \mid nb(u_i)) = \frac{1}{T} \times (Pr(c \mid u_{nb1}) + \ldots + Pr(c \mid u_{nbT})) \quad (10)$$

Where, $nb(u_i)$ is the set of top *T* most similar neighbors of user u_i computed on matrix R_u, $Pr(c \mid u_{nbi})$ is the gender/age tendency of the top *i* ($0 < i \le T$) neighbor.

Thus, we can smooth the gender/age prediction of user u_i by

$$\Pr(c \mid u_i)_{smoothed} = \beta \times \Pr(c \mid u_i) + (1-\beta) \times \Pr(c \mid nb(u_i)) \quad (11)$$

Where, $\Pr(c \mid u_i)$ is value got from Equation 4 and β is the parameter to control the influence of the user's gender/age probability predicted by its neighbors.

Same reason as the gender/age tendency prediction of Webpages, the computation $\Pr(c \mid u_i)_{smoothed}$ may also extended into an iterative procedure, and need several iterations to be stable.

5. EXPERIEMENTS

In this section, we introduce the experiment data set, our evaluation metrics, and the experiment results.

5.1 Dataset

We select a set of web-page click-through log collected by a large scale Web site. Each row in the web-page click-through log is a data, which is consisted with user id, age/gender information of the user and the Webpages users clicked. The full corpus comprises one week of access to the site in the year of 2006. For the data in the corpus, unique items with same user, clicked Webpage URL are grouped into one entry and the frequency is summed up. We filtered Webpages which occurred in the click-through log but not crawled by our crawler. We also removed the users who do not provide gender or age information, or clicked less than 10 crawled pages in our corpus since it is not reasonable to predict a user's gender or age by analyzing less than 10 pages. After the processing step, we get 189,480 users and 223,786 pages in sum.

Table 2 shows the demographic distribution over users. Given the large number of users in our corpus, it is reasonable to assume that our corpus reflects the actual demographic distribution of the Web site.

Table 2. Distribution over Age and Gender

age	gender		
	female	male	Total
Teenage	8.2%	8.2%	16.4%
Youngster	14.6%	16.8%	31.3%
Young	11.5%	15.7%	27.2%
Mid-Age	6.6%	9.0%	15.6%
Elder	3.9%	5.5%	9.4%
Total	44.7%	55.3%	100%

For evaluation consideration, we divided the users into training data and test data equally, and each of them keeps similar gender and age distribution as the corpus. We learn the Webpage tendency model on training data, and evaluate the model and user gender/age prediction performance on test data.

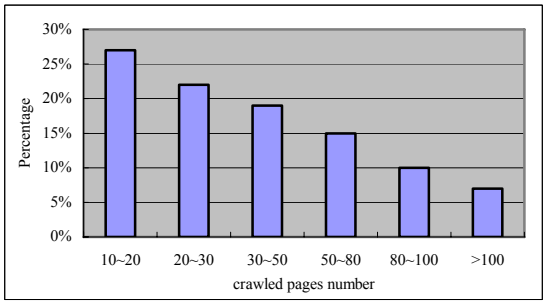

Figure1. User distribution within different groups in test set.

The web-page click-through log is very sparse, because the Web site contains tremendous of Webpages, but most users clicked only a few pages when browsing over the site. We segmented users of the corpus into six groups according the number of pages they clicked. As shown in Figure 1, most of users in the corpus clicked less than 30 pages.

5.2 Evaluation Metric

The performance of the presented methods was evaluated using the conventional precision (Prec), recall (Rec) and *F*1 measures. Precision *p* is defined as the proportion of correctly predicted examples in the set of all examples assigned to the target class.

Recall r is defined as the proportion of the correctly predicted examples out of all the examples having the target class. $F1$ is a combination of precision and recall defined as follows:

$$F1 = \frac{2pr}{p+r} \quad (12)$$

Furthermore, micro-averaged F1 (Micro F1) and macro-averaged F1 (Macro F1) [19] were applied to get single performance values over all prediction results.

5.3 Baseline Algorithm

For comparison purpose, we investigate training a classifier on user side directly. By assuming that the Webpages user visited should show the demographic attributes same to the user, we represent the user with the content of all the Webpages that the user visited. The linear form of Support Vector Machine (SVM) [20] classification model is used to learn model to classify user's demographic attributes. For gender and age, we train a model based on training data respectively. Besides, we also use Collaborative Filtering algorithm and LSI in the experiments. For CF, we apply memory-based algorithm with the vector similarity measure to form neighbors (Refer to Equation (1) and (3) in [8]). For LSI, we apply LSI on the <user, page> matrix and use the reduced rank approximation of original matrix for user gender/age prediction. The SVDPACK/las2 software package is used for SVD computation [31].

5.4 Experiments on Webpages' Demographic Tendency Prediction

As described in 4.1.2.2, the content words with high information gain associated with readers' demographic attributes are selected as features for Webpage demographic tendency prediction. The content words with top information gain associated with readers' gender is shown is Table 3.These words highlights the browsing differences between male and female readers. The female users prefer to visiting pages about movies, baby, kids, food, and family which are related to their personal life, while male users prefer to visiting Webpages about money, chat, girls, adult and cars. This is further borne out in Figure 2, where we show the similar statistics for categories of Webpages the male and female users visited.

Table3. Terms with highest information gain that appears more frequently among males and females respectively

Female	Male
download	sports
love	money
kids	car
food	search
movies	chat
baby	photo
music	news
life	software
animals	internet
family	girls

We also analyze the content words with top information gain associated with readers' age. Similar as Table 3, the content words show different browsing behavior of internet users with different age. Teenage concern more about sports and school; Youngster prefer to visiting Webpages about college, shopping, and movies; The Young are focused on entertainment, health, and their kids; Mid-age are interested in finance, privacy and their job; Elders show interests in news, market, and financial investment.

Except for the content words of Webpages, we also use the demographic distribution over concept hierarchy of Webpages as features. The taxonomy we used in the paper is from Open Directory Project (ODP, http://dmoz.org/). We crawled 1,546,441 Webpages from ODP which spanned over 172,565 categories. We can consider the hierarchy at different levels. In our work, we select the top 3 levels of ODP. After filtering the categories with less than 40 pages, we can get 1033 categories. Based on these hierarchical categories, we use the SVM to build the classifier. The linear form of SVM is used and the one-against-rest approach is applied for the multi-class case. The "document frequency selection (*DF*)" [19] and the information gain methods are used for feature selection. Then, we classified all the Webpages in the training data into the 1033 categories. Since the distribution of classified Webpages is unbalance, we only keep the categories which contain at least 10 Webpages. Thus, we can get 892 categories. Base on the demographic attributes of Webpages in each category, we can get the demographic distribution of each category. In Figure 2, the demographic distribution for some categories is shown. The Webpages under "Recreation\Autos" and "Sports" categories have a majority of male readers. Webpages about fitness and music are more likely read by females. Career development topics are more popular among younger generations than older generations.

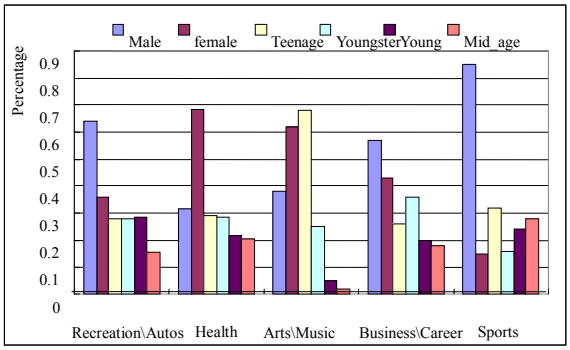

Figure 2. Demographic distribution of some categories

5.4.1 Evaluating the Webpages demographic tendency prediction

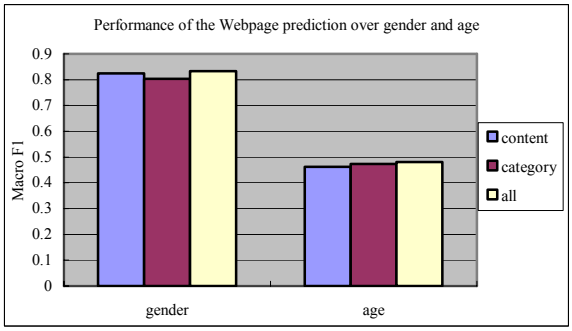

Figure 3. Results for predicting gender and age of Webpages

To evaluate the performance of Webpages' demographic tendency prediction, we selected the Webpages which are read by

at least 10 users in the test data, and use the gender and age tendency value as the answer. Since it is unnecessary to evaluate the prediction results by exact answer match, we divide the value into regions: [0, 0.2), [0.2, 0.4), [0.4, 0.6), [0.6, 0.8), [0.8, 1]. If the gender/age attribute with the highest prediction probability is the same as the answer, and the predicted probability is in the same region as that of answer, we denote this prediction is correct.

As shown in Figure 3, for gender tendency prediction, the content-based features are better than the category-based features, and their combination obtains the best result; while for age tendency prediction, the category-based features are slightly better than content-based features, and their combination is also the best.

5.5 Experiments on Users' Demographic Prediction

The experiment results on predicting users' gender and age on the test data measured in Macro F1 are shown on Figure 4. It is interesting to note that, although the category-based features are better than content-based feature in Webpage's gender tendency prediction, category differences are more telling than content difference in user gender prediction. Using all features we obtain 70.3% in terms of Macro F1. For age prediction, category features also proves to be slightly more useful than the content features, and the combination is also most useful. Table 4 shows the performance of the user prediction over gender and age.

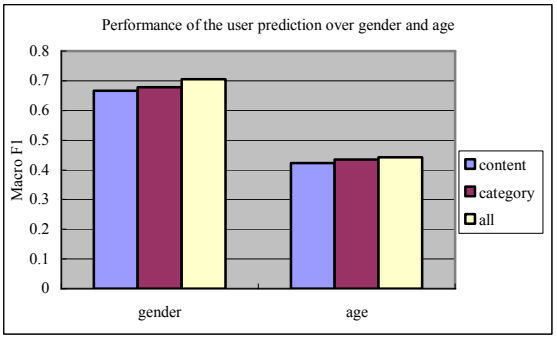

Figure 4. Results for predicting users' gender and age

Table 4.Results of users' demographic prediction

		Prec	Rec	Micro F1	Macro F1
Gender	Male	0.707	0.711	0.709	0.703
	Female	0.713	0.682	0.697	
Age	Teenage	0.361	0.323	0.341	0.461
	Youngster	0.498	0.503	0.5	
	Young	0.486	0.492	0.489	
	Mid-Age	0.457	0.44	0.448	
	Elder	0.403	0.297	0.342	

To investigate the influence of data sparseness, we evaluate users' gender and age prediction performance over different groups segmented by the number of pages the user clicked. As shown in Figure 5, the gender and age prediction performance of the users who clicked less than 30 pages is not good, and with the number of pages a user clicked, the user gender and age prediction performance becomes better, which indicates that it is necessary to deal with the data sparseness problem.

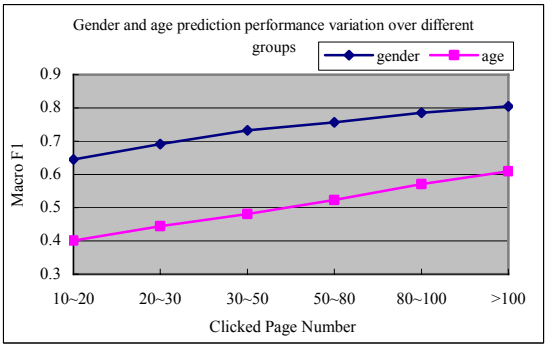

Figure 5. Gender and age prediction performance variation over different groups.

5.6 Experiments on Demographic Prediction by Leveraging Similarity among Users and Webpages

We first introduce the parameters we need to be estimated in the experiments, and the methodology to estimate them. Then, we evaluate the influence of different SVD dimension, normalization and smoothing methods.

5.6.1 Parameters Estimation

There are several parameters that need to be determined in our experiments. For the page's gender and age tendency prediction experiments, the parameter α_0, α_1 in Equation 9, the number of selected neighbors N in Equation 6 and M in Equation 8 are needed to be estimated. For the users' gender and age prediction experiments, the number of selected neighbors T in Equation 10 and parameter β in Equation 11 must be estimated. To tune the parameters for the page's gender/age tendency prediction, the optimization criterion we select is to maximize the gender/age tendency prediction results in terms of Macro F1. To tune the parameters of user's gender/age prediction, the optimization criterion we select is to maximize the gender/age prediction results in terms of Macro F1. At current stage of our work, the parameters are selected through exhaustive search.

We set the number of selected neighbors N = 10 in Equation 6, M = 10 in Equation 8, and T = 10 in Equation 10. Our experiments showed that page's gender/age tendency prediction and user's gender/age prediction are not very sensitive to the values of these parameters.

5.6.2 Influence of the SVD Dimensions

We conduct experiments to study the influence of dimensions on the performance of SVD. When we construct SVD matrix with different smoothing and normalization methods, all the results show the prediction performance depends on dimensions of the decomposed matrix. To find a good value of SVD dimension, we first use normalization from user dimension with constant smoothing (c = *0.05*), and fix the parameters $\alpha_0 = 0.6$, $\alpha_1 = 0.9$, $\beta = 0.6$. Then, we try several different values for dimension k from 50 to 200.

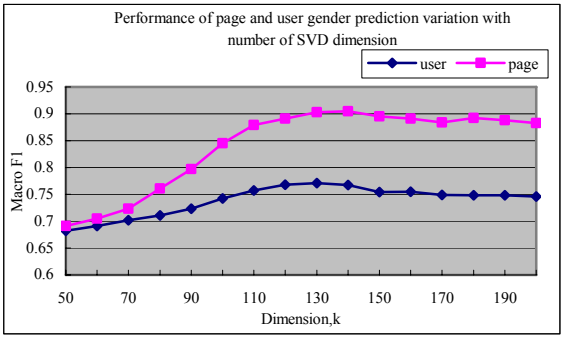

Figure 6. Determination of optimum value of k for gender in user and page side.

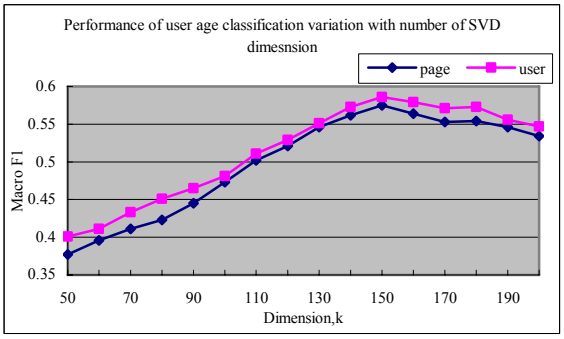

Figure 7. Determination of optimum value of k for age in user and page side.

As shown in Figure 6 and Figure 7, the performance of user gender/age prediction is correlated with the performance of page gender/age tendency learning. When the user gender/age prediction achieves the best performance - 77.1%/ 57.4% in terms of Macro F1 when k = 130/150, the page gender/age tendency prediction almost obtains the best performance. It means that if we can get better results in page gender/age tendency prediction, the results of user gender/age prediction will also be improved. Thus, we can first tune parameters in the page side, and then tune parameters in the user side.

5.6.3 Influence of neighborhood smoothing for page and user

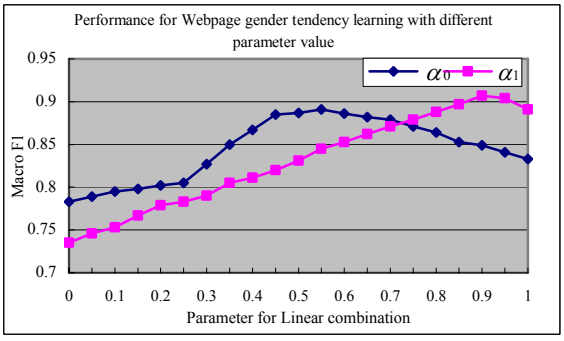

Figure 8. Performance for Webpage gender tendency learning with different value of α_0, α_1 .

We found that after introducing the neighbors, the best page gender tendency prediction result (90.7% in terms of Macro F1) is better than that of the one that doesn't use (83.3% in terms of Macro F1). The page gender tendency prediction achieves the best result when $\alpha_0 = 0.55$ and $\alpha_1 = 0.9$, which indicates that the *npr* neighbors are more important than the *nbp* neighbors. After that, we tune the parameter of β (the weight for user's gender/age probability predicted by neighborhood average). As shown in Figure 9, we can get the best user gender prediction result (78.2% in terms of Macro F1) when $\beta = 0.6$, which also indicates that the user gender prediction result is improved when introducing user's neighbors. We can get similar conclusion when tuning parameters $\alpha_0, \alpha_1, \beta$ on user age prediction. Through experiments, we can get 58.4% in terms of Macro F1 for user age prediction when $\alpha_0 = 0.6$, $\alpha_1 = 0.85$, and $\beta = 0.65$.

In order to measure the weight between the original gender tendency value learned by SVM regression, and the gender tendency value predicted by neighborhood, we first tune the parameter of α_0 (weight for gender tendency value predicted by SVM regression) from 0 to 1. Then, we tune the parameter of α_1 (weight for gender tendency value predicted by *nbp* neighbors) after selecting the best value of α_0 . The experiment results on test data are shown on Figure 8.

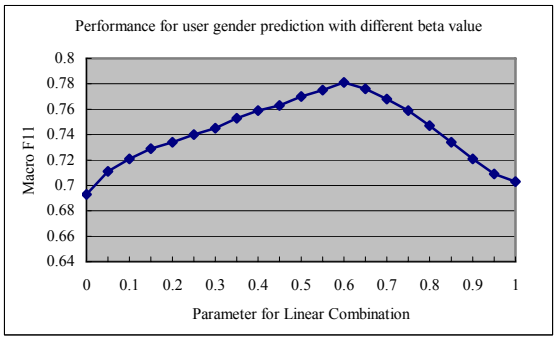

Figure 9. Performance for user gender prediction with different value of β .

5.6.4 Influence of iteration number for Webpage and user gender/age prediction

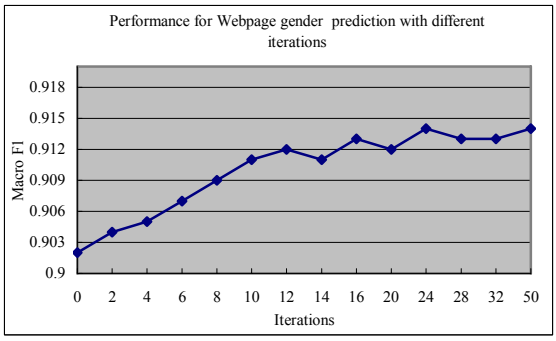

Figure 10. Performance for Webpage gender tendency prediction with different number of iterations.

In order to get a good iteration number that is effective for Webpage and user gender/age prediction, we try different iteration numbers to see how many iteration times are needed for our purposes. The results on gender are presented in Figure 10 and Figure 11, respectively. After 11 iterations, performance for

Webpage gender tendency prediction is quite stable; for user gender prediction, after 10 iterations, the performance becomes stable. So we use these values in the final gender classification experiments. We also tune the iteration number for age in the same way, and select 13 iterations for Webpage age tendency prediction and 11 iterations for user age prediction.

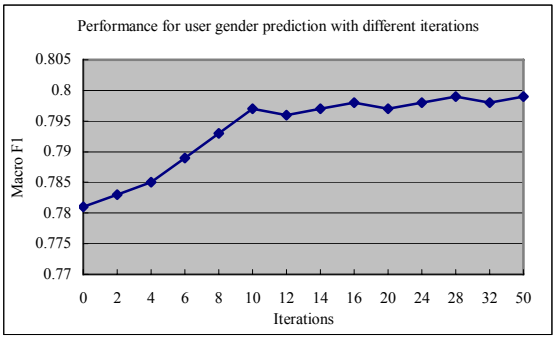

Figure 11. Performance for user gender prediction with different number of iterations.

5.6.5 User Gender and Age Prediction Results

As shown in Table 5, the user's gender prediction can archive 79.7% in terms of Macro F1 which is better than 70.3% of user gender prediction method without the smoothing of similar Webpages and similar users. Also, the user's age prediction can archive 60.3% in terms of Macro F1 which is better than 46.1% of our user age prediction method without the smoothing of similar Webpages and similar users.

Table 5. Results of user gender and age prediction on test data

		Prec	Rec	Micro F1	Macro F1
Gender	Male	0.791	0.81	0.8	0.797
	Female	0.805	0.782	0.793	
Age	Teenage	0.471	0.457	0.464	0.603
	Youngster	0.642	0.651	0.646	
	Young	0.632	0.642	0.637	
	Mid-Age	0.615	0.613	0.614	
	Elder	0.516	0.484	0.499	

To investigate the reasons for low performance of Teenage and Elder prediction, we also analyze the confusion matrix of them shown in Table 6. The confusion matrix indicates that many teenage are wrongly classified as youngster, and many elder are wrongly classified as mid-age.

Table 6. Confusion matrix for Teenage and Elder prediction

	Teenage	Youngster	Young	Mid-Age	Elder
Teenage	8,144	3,967	2,600	1,827	754
Elder	334	1,056	1,118	1,539	4,314

5.6.6 Comparison with Other Approaches

We also conduct experiment to compare our approach with CF, LSI and the approach to train classifier in user side. For CF, we vary the number of neighbors and report the best result. For LSI, the reduced dimension varies from 1 to highest possible dimension (the matrix rank) and tshe best result is reported. For the approach to train classifier in user side, we tried different feature selection methods and report the best result. As shown in Figure 12, our approach outperforms the other three approaches both on gender and age. Relatively, the improvement on gender/age over the CF is 29.6%/47.4%, LSI 16.7%/36.7%, the approach to train classifier in user side 30.4%/50.3% in terms of Macro F1.

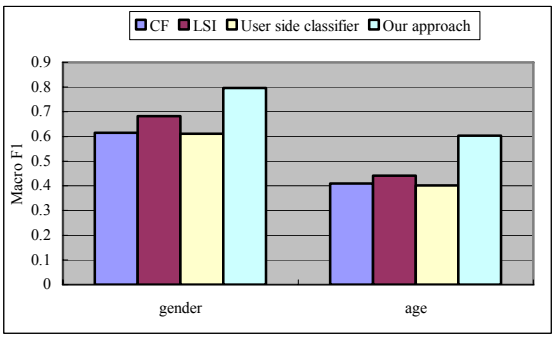

Figure 12. Performance for user gender/age prediction of CF, LSI, User side classifier, and our approach.

6. CONCLUSION AND FUTURE WORK

This paper focuses on demographic prediction based on people's internet browsing history. A novel solution is proposed to train a gender and age predictor based on web users' browsing behaviors. Experimental results on a real large page click-through log indicate that our proposed algorithm can achieve 79.7% on gender and 60.3% on age in terms of Macro F1, which achieves up to 30.4% improvements on gender prediction and 50.3% on age prediction in terms of macro F1, comparing with baseline algorithms.

There are also many areas for future research:

1) The profiles users may contain fake information. We believe that our proposed algorithm can be used to identify and refine the profiles which contain bogus demographic information.

2) In our current work, the demographic attributes we predict are only gender and age. We also plan to extend our research on other demographic attributes such as occupation, degree and location.

7. REFERENCES

[1] Berryman-Fink, C. L., J. R. Wilcox (1983). A multivariate investigation of perceptual attributions concerning gender appropriateness in language, Sex Roles 9, 1983.

[2] Biber, D., S. Conrad, R. Reppen (1998). Corpus Linguistics Investigating Language Structure and Use, Cambridge University Press, Cambridge, 1998.

[3] Computerworld Report: Men Want Facts, Women Seek Personal Connections on Web, http://www.computerworld.com/developmenttopics/websitemgmt/story/0,10801,107391p2,00.html.

[4] Eckert, P. (1997). Gender and sociolinguistic variation, in J. Coates ed., Readings in Language and Gender, Blackwell, Oxford 1997, pp. 64-75.

[5] Herring, S. (1996). Two variants of an electronic message schema, in S. Herring ed., Computer-Mediated

Communication: Linguistic, Social and Cross-Cultural Perspectives (John Benjamins, Amsterdam, 1996), pp. 81-106.

[6] Holmes, J. (1993). Women's talk: The question of sociolinguistic universals, Australian Journal of Communications 20, 3, 1993.

[7] Google Personal. http://labs.google.com/personalized.

[8] J. S. Breese, D. Heckerman, and C. Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In Proceedings of the Fourteenth Annual Conference on Uncertainty in Artificial Intelligence, pages 43–52. Morgan Kaufman, 1998.

[9] Lakoff, R. T. (1975). Language and Women's Place, Harper Colophon Books, New York, 1975.

[10] Lewis, D., R. Schapire, J. Callan, R. Papka (1996). Training algorithms for text classifiers, in Proc. 19th ACM/SIGIR Conf. on R&D in IR, 1996, pp 306-298.

[11] M. Koppel, J. Schler, S. Argamon, and J.W. Pennebaker. Effects of age. and gender on blogging. In AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs, 2006.

[12] M. Koppel, S. Argamon and A. R. Shimoni (2003). Automatically Categorizing Written Texts by Author Gender. In Literary and Linguistic Computing, 2003. Mulac, A., L. B. Studley, S. Blau (1990). The gender-linked language effect in primary and secondary students' impromptu essays, Sex Roles 23, 9/10, 1990.

[13] Mulac, A., L. B. Studley, S. Blau (1990). The gender-linked language effect in primary and secondary students' impromptu essays, Sex Roles 23, 9/10, 1990.

[14] Mulac, A., T. L. Lundell (1994). Effects of gender-linked language differences in adults' written discourse: Multivariate tests of language effects, Language & Communication 14, 3, 1994.

[15] Palander-Collin, M. (1999). Male and female styles in 17th century correspondence, Language Variation and Change 11, pp. 123-141.

[16] Manber U., Patel A., and Robison J. Experience with Personalization on Yahoo! Communication of the ACM, 43(8): 35-39, 2002.

[17] Simkins-Bullock, J. A., B. G. Wildman (1991). An investigation into the relationship between gender and language, Sex Roles 24, 1991.

[18] Search Engine Watch Journal, Behavioral Targeting and Contextual Advertising, http://www.searchenginejournal.com/?p=836.

[19] Yang, Y., Pedersen J.P. A Comparative Study on Feature Selection in Text Categorization Proceedings of the Fourteenth International Conference on Machine Learning (ICML'97), 1997, pp412-420.

[20] 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.

[21] Vapnik, V.N. The Nature of Statistical Learning Theory. Springer-Verlag, New York, NY, 2000.

[22] iMedia Connection: Behavioral Targeting Online Ad Spend, http://www.imediaconnection.com/content/9236.asp

[23] G. Golub and C. V. Loan. Matrix Computations, 2nd edition. The Johns Hopkins University Press,Baltimore, Maryland, 1989.

[24] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Application of dimensionality reduction in recommender systems-a case study, 2000.

[25] M. H. Pryor. The effects of singular value decomposition on collaborative filtering. Technical Report PCS-TR98-338, Dartmouth College, Computer Science, Hanover, NH, June 1998.

[26] J.H. Lee, “Combining Multiple Evidence from Different Properties of Weighting Schemes,” Proceedings of the 18th Annual ACM-SIGIR, pp. 180-188, 1995

[27] Pazzani M., Muramatsu J., and Billsus D. Syskill & Webert: Identifying Interesting Web Sites. In Proc. of the 13th National Conference on Artificial Intelligence, pages: 54—61, 1996.

[28] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. Journal of the American Society of Information Science, 41(6):391–407, 1990.

[29] Amanda Lenhart, Susannah Fox. Bloggers: A portrait of the internet's new storytellers. http://www.pewinternet.org/pdfs/PIP%20Bloggers%20Report%20July%2019%202006.pdf

[30] Finn V. Jensen. Bayesian Networks and Decision Graphs. Springer, 2001.

[31] M. Berry, T. Do, and S. Varadhan. Svdpackc (version 1.0) user’s guide. Technical Report CS-93-194, University of Tennessee, 1993.