Master Thesis in Statistics and Data Mining

Inferring user demographics from reading habits

by

Uriel Chareca

Division of Statistics
Department of Computer and Information Science
Linköping University

Supervisor
Prof. Mattias Villani

Examiner
Prof. Oleg Sysoev

“It is what you read when you don’t have to, that
determines what you will be when you can’t help it.”
Oscar Wilde

Contents

Abstract

This thesis investigates models and methods to explore the connection between users' reading habits and their demographics. The models predict user demographics, such as gender and age, based on textual information from a collection of publications read by a user and the time spent by him or her in reading each of the documents. Two approaches are introduced. First, we propose a nearest-neighbor type of algorithm using the relationship between the topics from the documents read in a defined topic dimensional space. Second, a Bayesian probabilistic model based on the Dirichlet distribution where the proportion of reading time spent on a number of topics is modeled as a function of the user demographics.

Acknowledgments

I would like to thank my supervisor Mattias Villani at LIU, which without his guidance and help, this thesis would not have been possible.

Second, I want to acknowledge my classmates, whose support and friendship were key into any decision I made during the last years.

Last, but not least, the motivation and data were originated from Issuu and their interest in knowing more about their customers. This project was carried out on site of Issuu offices in Copenhagen, between January 2014 to June 2014. I would like to thank all the amazing staff that made my stay an enriching experience.

1 Introduction

1.1 Objective

In a world where internet traffic information is vast and in some cases overwhelming, it is possible to track down most of the records of activity of an individual. Data about which sites a user access, what they buy, with whom they connect, where they connect from and what they read, etc. is often readily available. Much of this information is easily trackable by the sites by reviewing the user interaction with it. Demographic user information such as gender, age and location is key in marketing campaigns, user interfaces and ad placements; complementing the user records with its demographics is therefore an important objective for many firms. A major obstacle in reaching this aim is that many users are unwilling to release this type of information, and not many sites request it for the same reason. Social media sites such as Facebook, Twitter, Google+ or even work related ones like Linkedin leave much of this information request as optional when an account is created for log in, while others even allow site use without a requiring user registration. For instance, YouTube allows any user to use most of its service and stream videos without logging in, and many of their users therefore remain anonymous. Privacy restrictions are a good and valuable policy to retain confidential or sensitive data. The use of some demographics indicators, nevertheless, could be used without personally individualize users in a way that trespass their privacy. In conclusion, the majority of internet users have unknown, or at least partially incomplete demographic information. The aim of this thesis is to propose statistical data mining methods for predicting the missing demographics based on recorded user activity patterns, such as tracking of viewed content.

Our methods are applied and evaluated on a dataset from the web publication service Issuu[1]. As a digital newsstand with over 15 million magazines and 80 million active readers, Issuu features leading and emerging titles in fashion, culture, arts, and hyper local content, all of which are accessible on multiple devices as computers, tablets or cellphones. Issuu features a recommendation service that helps users discover new content in a shared social environment. Advertising placement by demographic is a basis for targeting the desired customers, and it is therefore of key importance to summarize by content its users age and gender. Our dataset includes information per user basis about the read publications' text, its related topic and time spent on each

[1] www.issuu.com

publication. This information is used to build and estimate predictive models for the users' demographics. We also propose models for the additional aim of predicting users' reading habits from information about their demographics. Such information is for example important for publishers when deciding on directed advertisement campaigns for a new publication. Both algorithmic and probabilistic models are investigated and we contrast the models in terms of predictive accuracy.

1.2 Related Work

Profiling an individual based on the characteristic of the text is not a new approach. People tend to write differently, both in terms of topics of interest and of word choices that define their writing style. The considerable size of available literary works provide the best scenario for this type of analysis. Many papers have been written relating the analysis of the linguistic characteristics in literary corpus to the author's age or gender (e.g. Pennebaker and Stone 2003 and Pennebaker and Graybeal 2001). Koppel et al. (2002) reached a gender misclassification error of only 20% from an analysis of the of fictions and non-fictions works from the British National Corpus, based on a mix of part-of-speech features and function word distribution. Santosh et al. (2013) introduce a machine learning approach to predict unknown author's demographic. Their approach considers different details of the text such as content, style and topics. They start by calculating the marginal frequencies of different N-grams (combination of words), punctuation count and topic based features written by a particular gender. This provides a high-dimensional space where classification algorithm such as decision trees and support vector machine are tested.

Many different types of data were considered in different studies, blogs (Schler et al. 2006 and Rosenthal and McKeown 2011), telephone conversation (Garera and Yarowsky 2009) or social media platforms as Twitter or Facebook (Schwartz et al. 2013 and Rao et al. 2010) were used as explanatory variable to model age. Nguyen et al. (2011) uses three different genres of data simultaneously: blogs, telephone conversations and online forum posts into a linear regression platform to predict age, as a classification algorithm into segments or as a continuous variable. This type of corpora based on social communications like blogs or posts share the similar problem of being based on short texts and usually contain colloquial lexical and therefore becomes complex for a natural language processing analysis. Peersman et al. (2011) presents a study where they apply a text categorization approach for the prediction of age and gender based on the text included in social media platforms. They use a combination of n-grams and typical phrases that lead to classify users in age ranges. this helps them track down interactions between underage and adults for instance, but find non conclusive results as they acknowledge the problem of noisy data originating from fake information on the user profiles. Similarly, we also use an age range in certain parts of this work. Zhong et al. (2013) aim to discover user demographic based on mobile data usage, disregarding the content but analyzing the particular type of events, their length and timings into a sequence of marginal probabilities.

Much of this related work focuses on the identification of the author's demographic based on the lexicon used. Most work comes from a combination of lexical feature extraction and applying machine learning algorithms like support vector machines or decision trees. Hsu et al. (2003) present an introduction to SVM as a classification algorithm, while Joachims (1998) exemplify this methodology into text categorization. Brodley and Utgoff (1995) details the use of decision trees on a multivariate dimensional space as a classification or regression algorithm.

In contrast to previous work, we are not going to work directly on the corpus of the publications. Our aim is not profiling the authors demographic but the reader's age and gender. A reader may read multiple and diverse texts. If we consider the whole text read it will include a considerable large corpus from diverse authors and type of publications. We start our analysis acknowledging the text mining work done before at the company, and therefore work directly from the topic distribution that Issuu has already designed. A topic is defined as the abstract "theme" a text is generally talking about. Documents belonging to a defined topic will likely contain certain words more frequently than others. Section 2.2 describes this process based on the Blei et al. (2003) paper. Similarly to Hu et al. (2007) who based their demographic prediction of user's browsing behavior, we consider the set of browsing data and the topic related, and try to associate similar users as part of our algorithm. They achieve a performance of 80% on gender and 60% on age in terms of classification based on a grid of age-range sections. Camargo et al. (2012) introduces an estimation and model selection for compositional data vectors based on a Dirichlet regression; this focus lead to the probabilistic approach presented in detail in Section 3.2.3 using the topic proportion distribution as response for such regression model.

1.3 Structure

From Issuu's vast database of user behavior, the first stage is to compile a scalable set of users with reliable details of their fingerprints such as age and gender. It includes enhanced information of their reading behavior by session including time spent on each magazine and on each topics. Chapter 2 includes the analysis of the available information, including an initial exploratory review and the definition of the notation used. Chapter 3 focus on the methodology and models introduced and defines different algorithms along the presentation of their results. Chapter 4 compares the different models and discusses future steps and opportunities. This work finishes with the conclusion in Chapter 5.

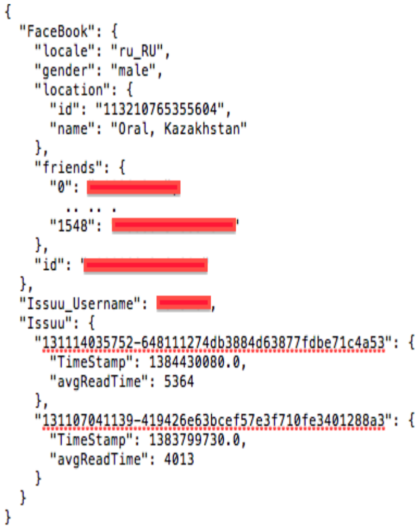
2 Data

2.1 Data sources

Our dataset comes from the internet publication service Issuu. It collects vast amounts of data, both from the 80 million people who read free publications every month (among which 9 million are registered readers) and from the 15 million uploaded document which at the moment add up to almost 200GB/day of traffic data from the website. Issuu users can review any material without logging in, with the exception of some specific functionality or restricted content where a formal access is required. Documents can contain text or just images. The document information includes such details as time of upload and by which user, cover, text, description, topics, etc. The user information is based on the login credentials completed by the user once they set up their account, which may or may not contain demographic information.

Users can link their Issuu profile to Facebook, Google+ or LinkedIn and in that case their credentials are copied from their public profile, or they can create a particular Issuu account. Issuu only ask at the moment for their email, display name, user name and age. After conducting an analysis of the available data, most of its information is stored in a static SQL database (known as "Issuu UserDB") that includes information from each document and each user. There is a clear problem as no consistent records are kept of the interaction, in order to find which documents are read by a user and for how long, we need to review each day raw traffic data. There is also a lack of consistency of the type of demographic information available. For many readers, as no login is required they are labeled as anonymous users. For the registered users, as their login credentials vary, demographic information for instance could include age, gender, both or none.

As mentioned, a fraction of the readers have registered via Facebook. This provides the opportunity of more reliable gender information, assuming users are more likely to provide truthful personal details on that social platform than on Issuu. For every user we can see their Facebook public profile including information about their gender, location, list of friends, education, etc. (see example of individual JSON file on Figure 2.1). We can also track down the list of documents read, their ID, the Unix time stamp (seconds since 1.1.1970) and their average read time. This last feature can be explained as the average time read on pages above a 3 second threshold, and is the indicator Issuu uses for how much a user likes a publication, and is the indicator we also use for aggregating the time spent by publications.



```
{
  "FaceBook": {
    "locale": "ru_RU",
    "gender": "male",
    "location": {
      "id": "113210765355604",
      "name": "Oral, Kazakhstan"
    },
    "friends": {
      "0": [redacted]
        .. .. .
      "1548": [redacted]'
    },
    "id": '[redacted]
  },
  "Issuu_Username": [redacted],
  "Issuu": {
    "131114035752-648111274db3884d63877fdbe71c4a53": {
      "TimeStamp": 1384430080.0,
      "avgReadTime": 5364
    },
    "131107041139-419426e63bcef57e3f710fe3401288a3": {
      "TimeStamp": 1383799730.0,
      "avgReadTime": 4013
    }
  }
}
```

Figure 2.1: Example .Json file.

The option of linking an Issuu account to Facebook was only made available recently; therefore we can see in some cases demographic information both from Facebook and from Issuu log in credential, and it is a perfect scenario to compare the reliability of the information in Issuu UserDB. From a small sample test of 30'000 users, 70% of users have both age and gender filled between Facebook and Issuu UserDB (Table 2.1). We note that most of the completed information in Issuu UserDB could be considered reliable (only 3% provided fake gender, using Facebook as valid reference).

Facebook Gender	Issuu UserDB Male	Issuu UserDB Female	Issuu UserDB Unknown
Male	14688	316	2200
Female	339	9411	1113

Table 2.1: Comparison Issuu UserDB vs Facebook gender information.

A bigger dataset was extracted from the individual analysis of the raw traffic data of the first month of 2014. This is the dataset used in the rest of this paper, and it provides information of 300'000 registered users and their activity in a similar JSON file as seen in Figure 2.1. Nevertheless once we trim the dataset for reliable data, selecting only users with both age and gender completed and at least 3 documents read, the sample is reduced to only 57'000 individual users. From an initial exploratory analysis we can observe that Issuu readers are mostly male (67%, consistent with Issuu previous analysis), and that the age distribution is similar between male and female concentrated in the 20-40 age range. While most users are male, we can see a clear fall in proportion of female users at elderly ages. For instance, at younger ages the proportion of female is close to 40%, while when we review users over 60 years the proportion falls to approximately 20% (Figure 2.2).

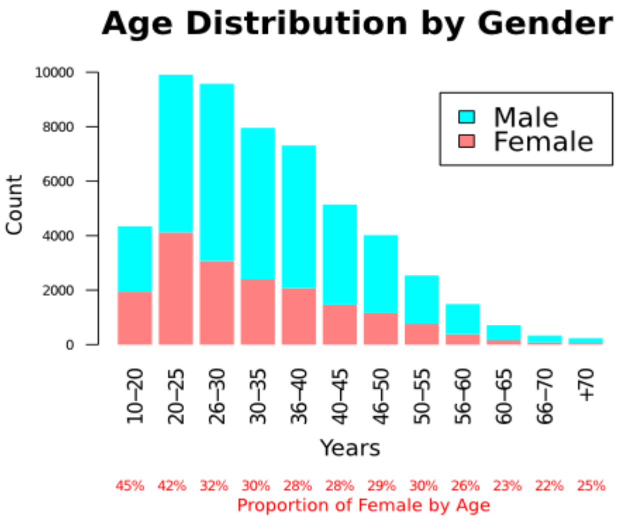


Figure 2.2: Age/Gender Distribution.

2.2 Topic data

Our dataset contains a number of added features for each document that were constructed by Issuu from separate statistical analysis of the publications' contents. For this project the topic distribution and label category are the most relevant.

Most of the publications are made of text (40% contain only images). A commonly used way to summarize textual information is to use topic models (Blei et al. 2003). A topic model summarizes the text in a document into a number of topics. Documents can then be compared by looking at differences in this low-dimensional topic space. This space works as a framework to find similar publication in terms of a topic. Representing a piece of text by a distribution of topics helps to separate different documents given its individualities, and also associate related content. The formal modeling framework is based on Latent Dirichlet Allocation (LDA) where a document can be described as being generated from a probabilistic distribution of topics (see Figure 2.3 for an illustration).

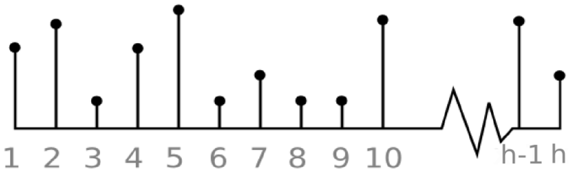


Figure 2.3: Topic Distribution, h topics.

LDA is a three-level hierarchical Bayesian model where each publication is seen as a random mixture over latent topics. Each topic is a unique probability distribution over a dictionary, a simple bag-of-words model. Therefore each word seen in a document can be explained as generated with a certain probability according to which topic distribution the document is identified with. The full generative LDA topic model is presented in Algorithm 2.1 (Blei et al., 2003).

Algorithm 2.1 LDA Generative model.

LDA works on the assumption of a generative model that create every word w_n in a set of documents $\mathbf{D}$ $(d = 1, \ldots, M)$ as:

- Each document d contains N words (N is generated from a $Poisson(\xi)$).
- For each document $(d = 1, \ldots, M)$:
  1. The topic proportions θ_d in a given document are a random draw from a Dirichlet distribution $Dir(\alpha)$ with parameters $\alpha = (\alpha_1, ..., \alpha_H)$.
     $$Dir(\theta|\alpha) \propto \prod_{i=1}^{H} \theta_i^{(\alpha_i - 1)}.$$
  2. For each word $(n = 1, \ldots, N)$:
     a) Generate a topic $z_{d,n} \sim \text{Multinomial}(\theta_d)$.
     b) Generate the word $w_{d,n}$ from $w_{d,n}|z_{d,n} \sim \text{Multinomial}(\beta_{z_{d,n}})$, where β_h is the probability distribution over the vocabulary for the hth. topic.

The parameters α and β can be found by maximizing the marginal log likelihood of the observed data. The posterior distribution can be approximated by variational Bayes techniques, see Blei et al. (2003) for details. A visual representation of this model can be made by the plate representation in Figure 2.4.

LDA Probability functions

- Document d is a sequence of N words from N topics $\mathbf{z}$, $\mathbf{w}$=$\{w_1, w_2, ..., w_N\}$.
- Corpus $\mathbf{C}$ of documents is a collection of M documents d_i, $\mathbf{C} = \{d_1, d_2, ..., d_M\}$.
- The joint distribution of the topic mixture can be described as:

$$p(\theta, \mathbf{z}, \mathbf{w}/\alpha, \beta) = p(\theta/\alpha) \prod_{n=1}^{N} p(z_n/\theta) p(w_n/z_n, \beta) \tag{2.1}$$

- Integrating over θ and summing over all probable topics z, we can reach the marginal distribution of a document **W:**

$$p(\mathbf{w}/\alpha, \beta) = \int p(\theta/\alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n/\theta) p(w_n/z_n, \beta) \right) d\theta \tag{2.2}$$

- Last, multiplying the marginal probabilities of every document we reach the probability of the corpus $\mathbf{C}$:

$$p(\mathbf{C}/\alpha, \beta) = \prod_{d=1}^{M} \int p(\theta/\alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n/\theta) p(w_n/z_n, \beta) \right) d\theta_d \tag{2.3}$$

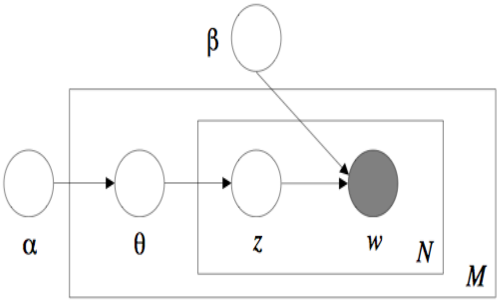
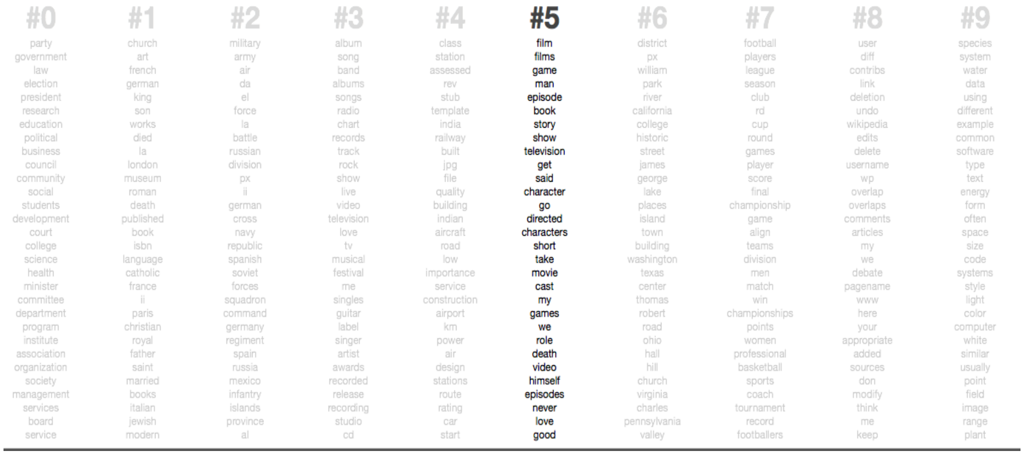


Figure 2.4: Plate representation of LDA.

On Issuu case, a bag of words was created from taking the 100'000 most frequent words from the full Wikipedia articles database, excluding common stop words. 150 topics were obtained using the gensim application (Rehurek et al. 2010); which aims to discover the topic distribution by examining the word frequency and patterns in a series of documents. Therefore for each wikipedia article, the probability to be generated from each topic can then be calculated. Figure 2.5 shows the most probable words in topic #5, which is led by words as film, films, game, episode, story, etc. Based on this word occurrences it seems sensible to label this topic as a movie-related topic. There are no real labels, in fact any topic label is an arbitrary interpretation based on the most commonly occurring words in the topic.



#0	#1	#2	#3	#4	#5	#6	#7	#8	#9
party	church	military	album	class	film	district	football	user	species
government	art	army	song	station	films	px	players	diff	system
law	french	air	band	assessed	game	william	league	contribs	water
election	german	da	albums	rev	man	park	season	link	data
president	king	el	songs	stub	episode	river	club	deletion	using
research	son	force	radio	template	book	california	rd	undo	different
education	works	la	chart	india	story	college	cup	wikipedia	example
political	died	battle	records	railway	show	historic	round	edits	common
business	la	russian	track	built	television	street	games	delete	software
council	london	division	rock	jpg	get	james	player	username	type
community	museum	px	show	file	said	george	score	wp	text
social	roman	ii	live	quality	character	lake	final	overlap	energy
students	death	german	video	building	go	places	championship	overlaps	form
development	published	cross	television	indian	directed	island	game	comments	often
court	book	navy	love	aircraft	characters	town	align	articles	space
college	isbn	republic	tv	road	short	building	teams	my	size
science	language	spanish	musical	low	take	washington	division	we	code
health	catholic	soviet	festival	importance	movie	texas	men	debate	systems
minister	france	forces	me	service	cast	center	match	pagename	style
committee	ii	squadron	singles	construction	my	thomas	win	www	light
department	paris	command	guitar	airport	games	robert	championships	here	color
program	christian	germany	label	km	we	road	points	your	computer
institute	royal	regiment	singer	power	role	ohio	women	appropriate	white
association	father	spain	artist	air	death	hall	professional	added	similar
organization	saint	russia	awards	design	video	hill	basketball	sources	usually
society	married	mexico	recorded	stations	himself	church	sports	don	point
management	books	infantry	release	route	episodes	virginia	coach	modify	field
services	italian	islands	recording	rating	never	charles	tournament	think	image
board	jewish	province	studio	car	love	pennsylvania	record	me	range
service	modern	al	cd	start	good	valley	footballers	keep	plant

Figure 2.5: The most probable words in the 5th topic shows that this topic is about movies.

Once this topic model is in place an individual topic vector can be calculated for every single Issuu publication. As the topic model is based on the analysis of a document text, no topic distribution can be created for publications that contain mainly images. The created vectors can be plotted as points in LDA space, the unit simplex of all possible topic proportions θ, see Figure 2.6 for a graphical representation. More importantly, distances between different documents can be calculated based on a chosen measure in the LDA space. This individual identification of a document in LDA space maybe called the “DNA of the publication”, similar to how DNA profiling helps to identify individuals based of their DNA profiles.

Based on these LDA distances, documents can be clustered in segments to find related content. Issuu used a modified K-means algorithm and experiments were run on segments sizes 3’000, 36’000 and 150’000. The number of topics was chosen to be 150 based on a rather arbitrary balance of visual interpretation, related content accuracy and computational scalability. For example, the number of topics was varied and it was checked how the content of supposedly similar documents was related, for example, checking if a Car magazine have similar Car magazines as their nearest neighbors.

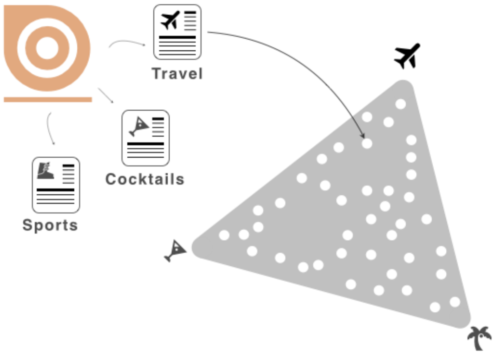


Figure 2.6: LDA Space.

The interpretation of the individual topics in an LDA model is based on the inferred word probability distribution of the topics; this can be a rather arbitrary and subjective choice. Also, in many cases there is no clear interpretation of the topic. To improve the interpretation of topics, Issuu included the original 4.5 million labelled documents from Wikipedia in LDA space. The main reason for this approach is that Wikipedia articles are typically concentrated into specific topics/labels, unlike a newspaper or magazine uploaded to Issuu. The often carefully labelled Wikipedia articles were therefore used to strengthen the topic interpretations for the non structured Issuu publications. Using 84 different Wikipedia labels, it was possible to create a vector of topic proportions for every Issuu document by considering its closest Wikipedia neighbor articles and weigh their respective Wikipedia labels. For instance a magazine could be classified as 50% travel, 40% sport and 10% dancing according to their distance to Wikipedia articles and their respective labels. Any publication that is considered to not be close enough to any Wikipedia article (using a 0.4 distance threshold, under scale 0-0.8 in the Jensen-Shannon divergence) is classified as unlabeled. Therefore, we can identify for every magazine a normalized vector of 150 topics, labels and their weight, and even to which segments they belong.

2.3 Exploratory analysis

The data to be analyzed consists of the reading habits of 57’000 users who read at least 3 documents. 1.2 million reads from 450’000 publications were found, belonging to 25’000 different segments; the average publication was read 2.6 times. Grouping the readers by magazine or segment allows us to perform an illuminating analysis by content.

A first observation is that there is a considerable fraction of publications with explicit content. Even if Issuu provides a wide range of publications and contents, porn is still the most popular content, with at least half of the most read documents belonging to the topic. As the Issuu label distribution model is based on Wikipedia labels and there is no Wikipedia articles related to porn, and moreover the explicit content is mostly pictures, this type of content is mostly classified as Unlabelable, or under labels as Photography, TV or similar media contents.

The LDA model is based on text mining of its content, for many magazines the content consists solely of images (only pictures or a scan of a mix of text and pictures). Therefore in those cases the label remains Unlabelable. If we add the number of magazines that remain too far away from any Wikipedia article, this Unlabelable portion adds up to 40% of the available content. A solution to this high proportion of Unlabelable publications is to consider the label proportions from the known segment, instead of the one of every particular document. A publication that might not be significantly labeled, belongs to a segment that does contain a majority of a certain specific topic related publications. Now the Unlabelable content is reduced to only 25%, with only 3.5% of the monitored users reading only Unlabelable documents (Figure 2.7). The remaining labels have a substantially smaller proportion, led by Literature with 10% and Art with 8.5%, and followed with a long tail with smaller numbers on the remaining labels.

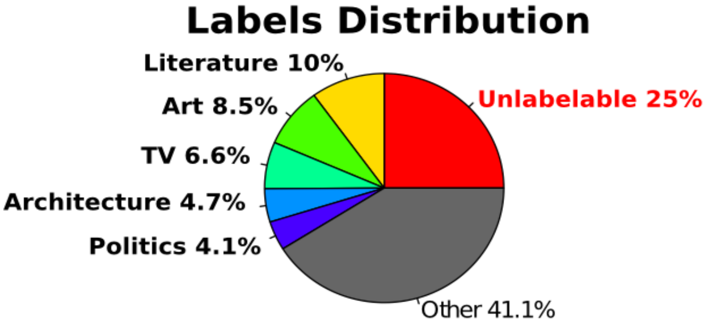


Figure 2.7: Label Distribution.

If we aggregate the information by the 25‘000 segments available, we can see that only 42% are read over 10 times and only 8’000 (33% of them) include at least 10 different publications. We can perform the exercise of classifying a segment as male or female if the proportion of readers of that gender is over 90%. 2000 segments can be labeled as male, but only 50 segments (1.5%) have a significant female concentration. These female segments have a smaller number of reads, an average of 20 reads versus 70 reads for the male segments. When we inspect the leading topics of the male segments, we can observe labels like Trains, Military or Games that might lead to the assumption that such labels are strongly correlated with male audience. Nevertheless the distribution of labels shares a similar order between genders. The chart in Figure 2.8 identifies labels like University and Arts as female related and Politics, Cars and Sports as male ones. When reviewing the top difference of gender by label, we can see that in fact labels as Audio, Airplanes and Motorcycles have a stronger masculine identification while labels as Food, Theatre and Fashion a feminine one. Given that most of the readers are male, a ruling based only on the labels read might be misleading, as seen with the still higher influence of male proportions in Figure 2.8; a normalization of the time observed by original gender proportions might help to solve this issue (see Future work section for details).

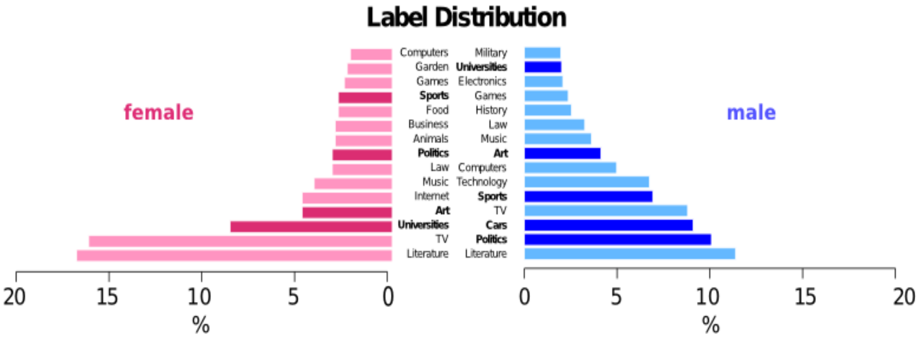


Figure 2.8: Label distribution by Gender.

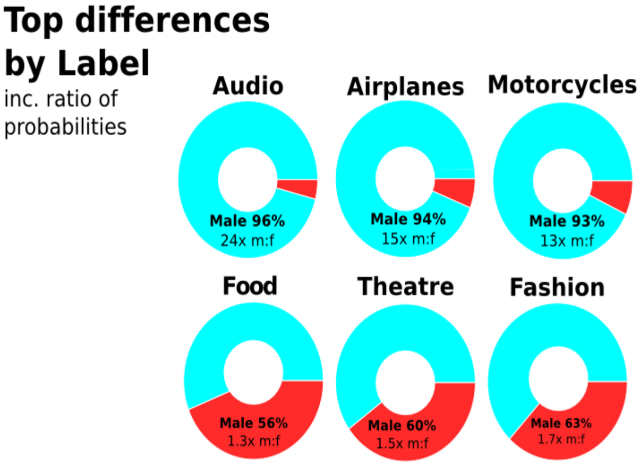


Figure 2.9: Top differences by Label.

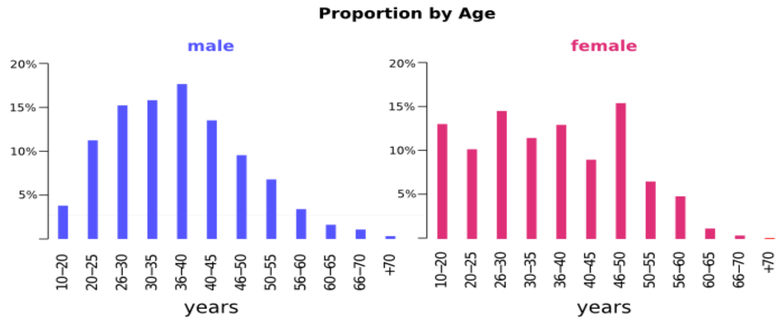


Figure 2.10: Age distribution of gender from related segments.

If we want to better analyze the influence of a label by gender, we can examine the relative frequency of a user reading a label, given his or hers gender (Figure 2.11). This analysis now excludes the significant influence from the high number of males in the sample set. Based on this analysis, calculating the proportion of reader of certain label by gender population; male labels are led by Technology and Cars, both with 10% higher male marginal proportion in comparison to the marginal proportion of the label in female readers. Higher difference of female labels proportion are led by Food and Art, with 7% and 5% respective of read proportion.

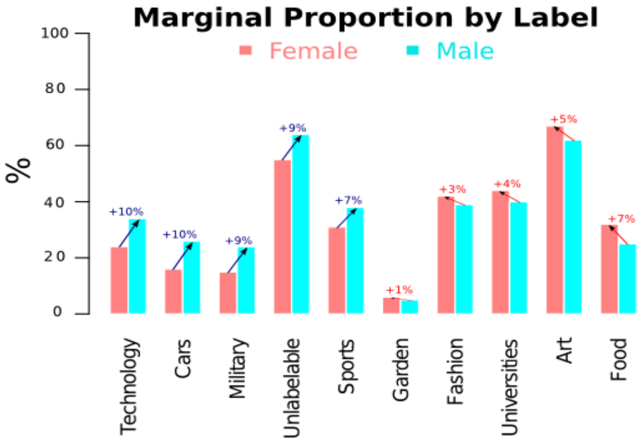


Figure 2.11: Marginal Proportion by Gender

The age of the readers present a similar density by gender (Figure 2.10). The time spent (identified by the logarithm of the total average read time of a user) also shows a similar shape, which rejects a possible idea that those variables might lead to a good classification criterion for gender. Nevertheless the total reading time might work as a good measurement of reliability, by weighting different observations if they come from a casual, average or big reader. We can label a user according to his reading time: if it is below the 33% percentile he can be considered a casual reader, while the 67% percentile splits between average and big reader respectively (see Figure 2.12).

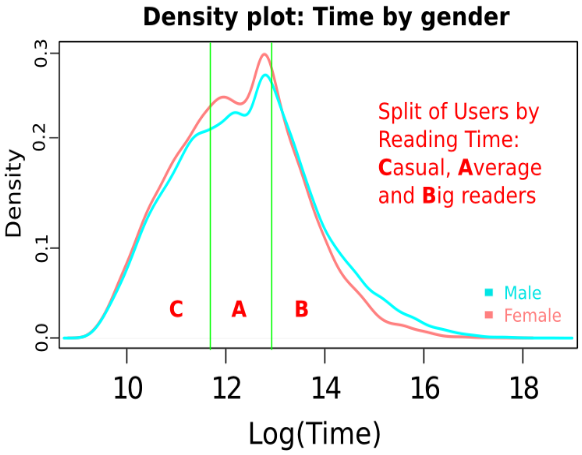


Figure 2.12: Reading Time Density.

2.4 User reading habits - the DNA of a user

We mentioned the concept of “DNA of a publication” as its individualized vector of topic proportions. Each document presents a different combination of topic weights, which helps to place it into an LDA space and consequently relate close ones into similar segments. The same concept could be used for the individual characteristic of a user (“DNA of User”) as his or hers corresponding proportion of reading time by topic, reading habits. A vector by user is created by aggregating his or hers distribution of total reading time over the 150 Topics or the 84 Labels. We can use any of this topic distribution vectors as explanatory variables in order to predict the demographic or vice versa.

Notation

- The LDA model has H topics. (taking in consideration the original 150 topics or the transformed version using the 84 **labels** from the near wikipedia labeled articles).
- The LDA model gives a **topic proportion** vector for each document (publication) $\theta_d, d = 1, ..., M$, where M is the total number of documents. θ_d is a H-dimensional vector on the unit simplex (the elements are between zero and one, and they sum to one).
- The created vectors can be plotted as points in an **LDA space**, the unit simplex of all possible topic proportions θ.
- **Segments** are defined as the cluster of documents, from a modified K-means algorithm, based on their distance in the LDA space.
- We have observed N readers over a period of time. Let t_{nd} be the average minutes spent on reading document d by reader n ($n = 1, ..., N$). Let $t_n = \sum_{d=1}^{M} t_{nd}$ be the **total reading time** of reader n.
- The **topic reading time** for reader n is defined as $\psi_n = \sum_{d=1}^{M} t_{nd}\theta_d$. This is a H-dimensional vector.
- The **reading habits** of reader n is defined as $\bar{\psi}_n = \psi_n/t_n$. Note that $\overline{\psi}_n$ is a H- dimensional vector which is now normalized to the unit simplex, and also referred as the "DNA of User".
- Performing the same procedure, we can create reading habits for the 150 Topics $\overline{\psi}_{n,150T}$ and for the 84 wikipedia Labels $\overline{\psi}_{n,84L}$.
- Let x_n be a vector with **demographic information** for reader n. For example $x_n = (Age_n, Gender_n)\prime$.
- All available user information can be consolidated into one user vector: $\mathbf{U_n} = (Age_n, Gender_n, \overline{\psi}_{n,150T}, \overline{\psi}_{n,84L})'$.

3 Methods and Results

3.1 Assessment of data quality

The goal of this project is to find optimal techniques to model user's demographics based on their reading habits. Initial exploratory techniques tested models that could clearly separate groups by age, gender, or even a mix of them as part of a regression framework. Techniques such as Logistic Regression, Decision Trees or Support Vector Machines were tested but every single example provided poor results. SVM produces a model that classifies all users as male, while using a decision tree shows a poor fit as well. Figure 3.1 shows an example of how the proportion spent on literature related content is distributed without a clear separation over age and gender. Males originally represent about 67% of the available users, therefore a slight male predicted bias effect is expected. Our sample data of 57'000 users seems like a reliable sample of the users distribution of Issuu, it also has over 60% of the male population; we therefore decided to continue with the full sample, instead of creating a subsample composed of an even gender split. Another option to solve this problem would be to use an asymmetric loss function that could use different weights according to the original gender proportion observed (Bach et al. 2006).

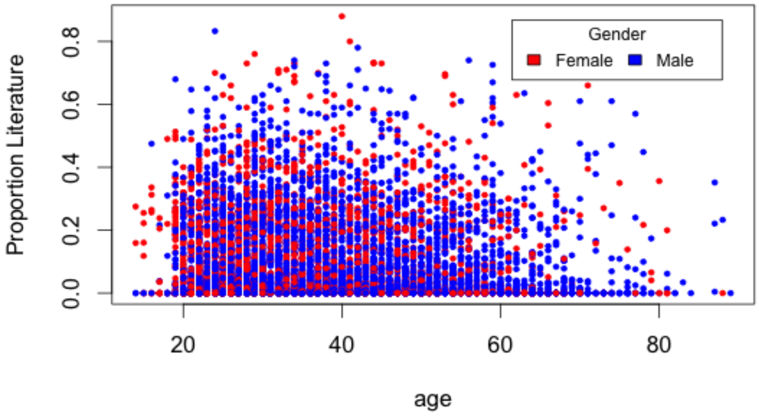


Figure 3.1: Example of concentrated data, difficult to separate.

3.2 Models

This project presents two new approaches to overcome the difficulties seen in this type of scenario where no linear separation models perform well. The first approach is based on projecting the data into a LDA space, and by performing a K nearest neighbor approach define different algorithm that use the reading history of a user as a weighted average for a predicted age and gender. The second approach is based on a bayesian framework, predicting the demographic distribution of a user based on the observed reading habits, using a Dirichlet regression model.

3.2.1 A nearest-neighbor predictor based on reading lists

Each user has a reading list, that includes every publication read in the Issuu system as defined in the Notation section. Similarly, each document d has a readers list that includes every user and their respective demographic details $x_n = (Age_n, Gender_n)$ and time dedicated to document d, t_{nd}. In the same way, a listing can be produced with all users reading a document that belong to a particular segment. We can use these statistics to weight the age and gender by the total reading time of the user and calculate the predicted age and predicted gender of the reader, based on those assigned to the document (or segment).

Let $t_d = \sum_{n=1}^{N} t_{nd}$ be the total reading time of document d, and the listed ages and gender for all users, **Age** $= \{A_1, A_2, ..., A_N\}$ and **Gender** $= \{G_1, G_2, ..., G_N\}$ respectively.

We can predict the document's age as an average of the ages of its readers weighed by their reading times:

$$Pred_D(Age) = \frac{\sum_{n=1}^{N} t_{nd} \cdot A_n}{t_d} \tag{3.1}$$

The predicted gender is based on the male probability $P_D(male)$, if the male probability is over 50% we can consider $Pred_D(Gender) = "male"$:

$$P_D(male) = \frac{\sum_{n=1,G_i=male}^{N} t_{nd}}{t_d} \tag{3.2}$$

As seen in the previous section, each document can be summarized by its topic proportions to describe its document DNA. Each document can be plotted in an LDA space of h dimensions, and a Euclidean distance can be calculated between a document (X) and a Candidate (C_j) by using Equation 3.3 below. Using this distance measure we can calculate as well the distance based similarity between two points in the range 0 to 1 by Equation 3.4. We can now, giving a particular document (X) , and a given list of Candidates, rank the K nearest neighbors based on the distance, or distance based similarity. The set of neighbors is declared as $\mathbf{KNN} = \{KNN_1, KNN_2, \ldots, KNN_K\}$ where KNN_1 is the neighbor with higher rank, higher similarity; KNN_K refers to the neighbor with the lowest similarity. Other distance scores (e.g. cosine angle) were tested with no particular improvements, so we continued with the simple approach of a Euclidean distance.

$$d(X, C_j) = \sqrt{\sum_{i=1}^{h} (X_i - C_{j,i})^2} \tag{3.3}$$

$$Sim(X, Cj) = \frac{1}{1 + d(X, C_j)} \tag{3.4}$$

We can now introduce Algorithm 3.1 to predict a user's demographics given the list of read documents. It uses both the particular documents read and the related neighbor documents, aiming to enhance the observed distribution of age and gender by compensating the shortage of information of single documents. We weigh each neighbor according to their rank order ($K - i + 1$), different weighing schemes and similarity measures were tested but no significant improvement was observed.

Algorithm 3.1 Model 1 - Reading list by Document.

A user u_n $(n = 1, .., N)$ with reading list of j publications $\mathbf{D_n}$ =$\{D_{n,1}, D_{n,2}, ..., D_{n,j}\}$, has a respective reading time t_{nd} for document d , and total reading time t_n.
The predicted age and gender of u_n, can be calculated based on its neighbors as follows:

1. For each element in $\mathbf{D_n}$, find the K nearest neighbors, **KNN**, in the LDA space using the similarity metric in Equation 3.4.

2. For each KNN_k calculate its $Pred_{KNN_k}(Age)$, and $P_{KNN_k}(male)$ by Equations 3.1 and 3.2.

3. Calculate the predicted demographics of each document as

$$Pred_{D_{n,i}}(Age) = \frac{\sum_{i=1}^{k}(K - i + 1) \cdot Pred_{KNN_i}(Age)}{\sum_{i=1}^{k}(K - i + 1)}$$

$$P_{D_{n,i}}(male) = \frac{\sum_{i=1}^{k}(K - i + 1) \cdot P_{KNN_i}(male)}{\sum_{i=1}^{k}(K - i + 1)}$$

4. Calculate the user predicted age and gender by a weighted average of the documents read's age and gender:

$$Pred_{u_n}(Age) = \frac{\sum_{i=1}^{j} t_{n,i} \cdot Pred_{D_{n,i}}(Age)}{t_n}$$

$$P_{u_n}(male) = \frac{\sum_{i=1}^{j} t_{n,i} \cdot P_{D_{n,i}}(male)}{t_n}$$

This algorithm provides a model to predict gender for a test user , according to if the probability of being male is over 50%, and can create a predictive interval using the predicted mean age and variance. We considered the option of assuming each Document age as coming from a $N_D(\mu_D, \sigma_D)$, where μ_D and σ_D are calculated from the mean and variance age of observed readers. Then the User predicted age and variance can be calculated from a $Multinormal(\mu, \sigma)$ weighting each document by the time read. This option provided non significant differences, therefore was discarded for this project but might provide better results when the data sample is bigger. As an example of the capabilities of the detailed predictive model (Algorithm 3.1), here is a possible output :

"*predicted sex is MALE with probability of 85%" "predicted age is 35, with a 95% predictive interval between 34.5 and 35.5.*"

As at each stage, the model searches for all K nearest neighbors for every document, which makes the algorithm computationally demanding. A solution could be to modify by take the predicted age and gender of the segments where this documents belongs (Algorithm 3.2), instead of taking in consideration the actual documents read and neighboring documents. One drawback expected is that by only looking at the segments's users distribution, we are not associating documents that are close to each other but not in the same segment, while we do link others that might not be in their nearest neighbors. Figure 3.2 provides an example of this situation: Document B is clearly closer to Document A, than Document C. So, Document B will have a large weight in the prediction of document A. Using the mentioned approach of by segment and not by document, Document B belongs to a different segment than Document A, and will not even be considered in the prediction of document A.

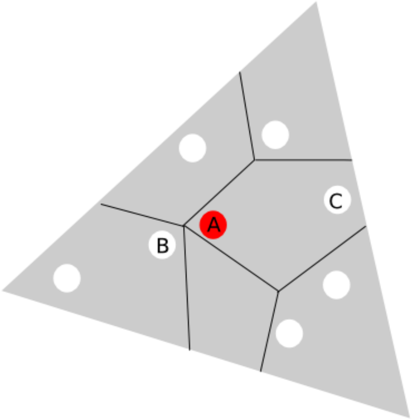


Figure 3.2: A belongs to same segment to C, but is closer to B.

Algorithm 3.2 Model 1 - Reading list by Segment.

A u_n with reading list of j segments $\mathbf{S_n}$ =$\{S_{n,1}, S_{n,2}, ..., S_{n,j}\}$, has a respective reading time t_{nd} for document d , and total reading time t_n.
The predicted age and ender of u_n, can be calculated by:

1. For each element in $\mathbf{S_n}$ calculate its $Pred_{S_{n,i}}(Age)$, and $P_{S_{n,i}}(male)$ by 3.1 and 3.2.

2. Calculate the user predicted age and gender by a weighted average of the read segments

$$Pred_{u_n}(Age) = \frac{\sum_{i=1}^{j} t_{n,i} \cdot Pred_{S_{n,i}}(Age)}{t_n}$$

$$P_{u_n}(male) = \frac{\sum_{i=1}^{j} t_{n,i} \cdot P_{S_{n,i}}(male)}{t_n}$$

For validating purpose we split our sample set into a test set of 10'000 users, versus the remaining part of the original 57'000 as training set. As the first algorithm is comparing each document read versus all the corpus of documents, it becomes a slow and computationally heavy process. As an approximation to improve the computational scalability, we can run the models for a series of random sub-samples of 100 users from the test set and consolidate the results. Figure 3.3 compares the misclassification error (see eq. 3.5, also defined as $1 - Accuracy$) for different numbers neighbors, K. Increasing K doesn't seem to have a considerable effect on the misclassification error, but a clear trend can be see in the predicted age RMSE (Figure 3.3). Considering the small differences, no significant conclusion can be made. The predictive performance is considerably improved when we focus directly on the segments. If we use the whole 10'000 test samples, we still observe a considerable improvement that confirms the initial finding. Below in Table 3.1 the summary of gender predictions and the actual gender running the segment based algorithm for the whole test set . The gender misclassification rate is almost 30%, clearly with a worse performance for female users, almost 2/3 of them being classified as male. Defining the precision as the fraction of relevant predictions, the precision of a predicted male user is 70%, while the female precision is considerably higher (84%). In other words, the female predictions are more reliable than the male predictions as the later include a higher number of misclassifications.

Real Gender	Predicted Male	Predicted Female
Male	6509	120
Female	2729	642
Misclassification Error	28%	
RMSE for Age	10.33	

Table 3.1: Performance of full test set of 10'000, by Segment based algorithm.

$$MisclassificationError = \frac{\sum_{n=1}^{N} I\{Pred_{u_n}(Gender) \neq Gender_n)\}}{N} \tag{3.5}$$

$$RMSE \text{ for Age} = \sqrt{\frac{\sum_{n=1}^{N}(Pred_{u_n}(Age) - Age_n)^2}{N}} \tag{3.6}$$

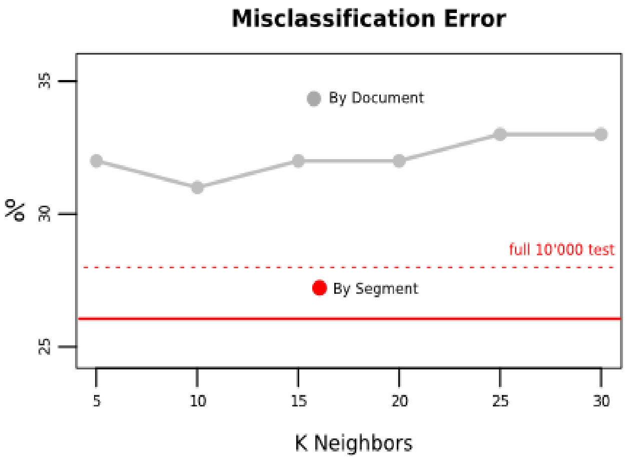


Figure 3.3: Comparison of Misclassification Rate from models based on reading times for documents or segments.

No real trend is seen when increasing the number of neighbors in the algorithm by document (in GREY). In RED the results when using the same random sub samples under the segment based algorithm. The misclassification error observed for the 10´000 test set (dashed line) is above the one of the consolidated random samples, but still with a considerably better performance than the results observed on the algorithm by document.

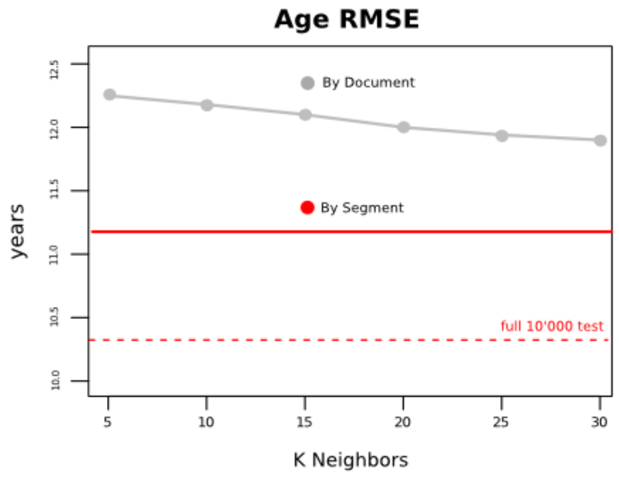


Figure 3.4: Comparison of Age RMSE between model based on documents or segments read.

Age RMSE performance presents a slight decline when the number of document neighbors is increased. The algorithm by Segment reaches the best observable performance of 10.33 when the whole test set is considered.

Figure 3.5 displays the predicted ages versus observed ages. Even if there is a clear positive trend, a sign of a good fit, the domain of the predicted ages is only between 20 years to 50 years while the real ages are distributed between early teens to late 80s (no significant different pattern in female or male sets). This means there is a clear concentration of mid aged prediction. If we analyze the distribution of the residuals we can in fact see a larger proportion of prediction under real values, but with a longer tail of prediction falling behind of the real values, in particular in elder ages (see Figure 3.10). The QQ plot (see Figure 3.11) shows a clear deviation from the normal distribution of residuals.

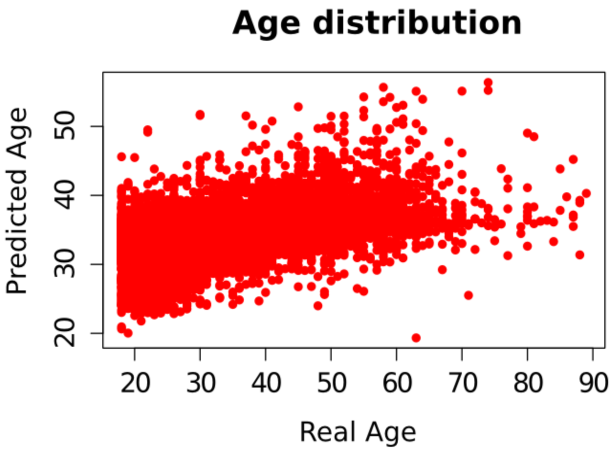


Figure 3.5: Predicted Age vs. Real Age (Model based on Segments read).

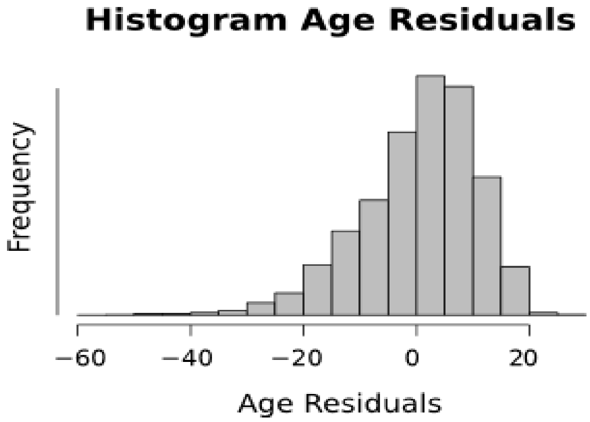


Figure 3.6: Histogram Age residuals.

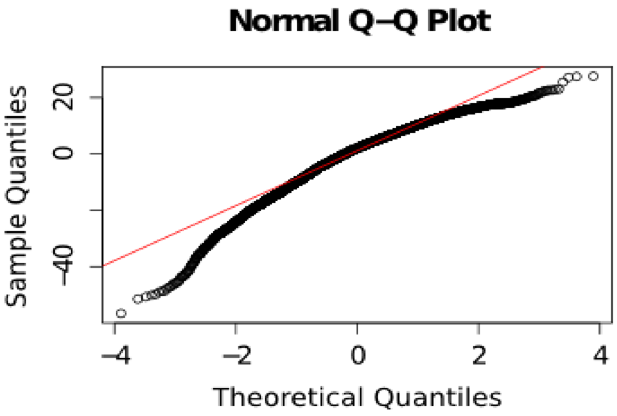


Figure 3.7: QQ plot Age Residuals.

3.2.2 A nearest-neighbor predictor based on user DNA

In the previous section we presented a set of variables describing the reading habits of a user. The DNA by user is a topic proportion vector, similar to the one seen by documents in the original LDA space. Therefore a distance based algorithm can be used (see equation 3.4), and by so find related users by their nearest user neighbors. For every user, K nearest neighbors are found using the aggregated 150 LDA space (a normalized vector based on the aggregated reading habits $\overline{\psi}_n$) and the distance between the user vector and the rest of the users in the training sample.

Algorithm 3.3 Model 2 - DNA of User.

Each u_n has particular topic proportion vector that defines its **reading habits** $\overline{\psi}_n$sec. 2.4

The predicted age and gender of user u_n , can be calculated by:

1. Find the K nearest neighbors, **KNN**, using the $\overline{\psi}_n$ reading habits in the respective LDA space (see equation 3.4).
2. Each KNN_i correspond to a $User_i$, and we know its own $KNN_i(Age) = Age_i$, $KNN_i(Gender) = Gender_i$ and total reading time $KNN_i(t) = t_i$
3. The User predicted age and male probability are calculated as:

$$Pred_{User_n}(Age) = \frac{\sum_{i=1}^{k}(K-i+1) \cdot KNN_i(Age) \cdot KNN_i(t)}{\sum_{i=1}^{k}(K-i+1) \cdot KNN_i(t)}$$

$$P_{User_n}(male) = \frac{\sum_{i=1}^{k}(K-i+1) \cdot KNN_i(t) \cdot I\{KNN_i(Gender) = male\}}{\sum_{i=1}^{k}(K-i+1) \cdot KNN_i(t)}$$

Algorithm 3.3 is applicable both for the 150 original topics weights, and for the 84 labels from the Wikipedia articles. A third topic proportion vector can be generated using the respective segments label weigh distribution, instead of using the one of each individual document read. The benefit of using the Wikipedia labels as input proportion vector is that it provides clearer understanding of the meaning of each proportion, because the labels have a clear meaning; topic Sports means really sports, and not an arbitrary meaning assigned from a rank of a bag of words (see Section 2.2). Nevertheless the performance of Model 2 by different type of input vectors (Figure 3.9) shows that a DNA by user based on the 150 topics performs better in comparison. The misclassification error does not provide a clear trend using the label vectors, but the model with 150 topics present a significant improvement as K increases. The Age RMSE as seen before still shows an improvement at higher number of neighbors considered.

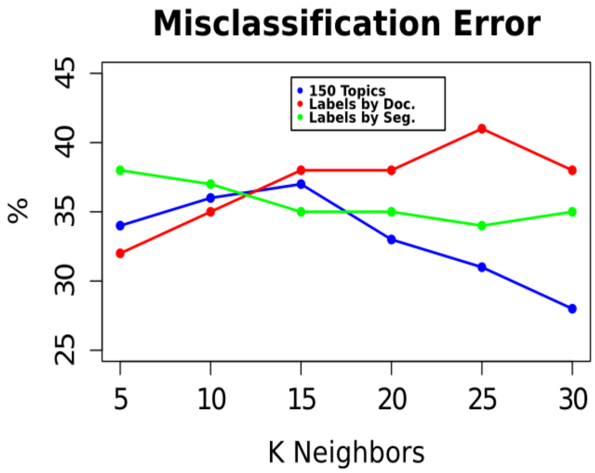


Figure 3.8: Performance of Model 2 (Misclassification Error) - 3 LDA inputs.

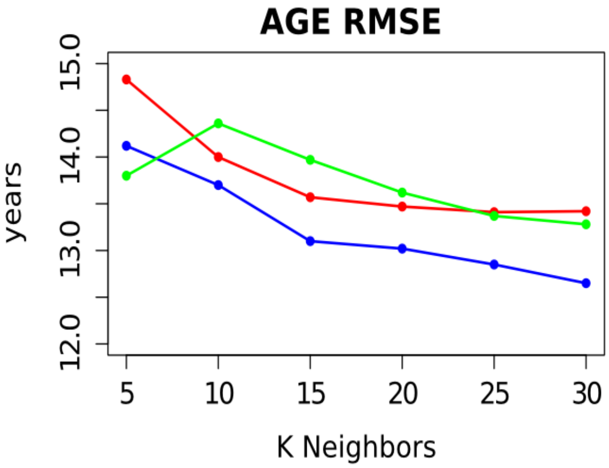


Figure 3.9: Performance of Model 2 (Age RMSE) - 3 LDA inputs.

3.2.3 A probabilistic model based on Dirichlet regression

The two models presented so far use algorithms based on weighted averages of the demographic of similar users based on their reading habits, with weights based on the distance between the particularities of the documents read or directly using the newly introduced concept of the DNA of the reader. Models of this type benefit from a clear understanding of their processes but produce point predictions without probabilistic assessment of their uncertainty. In this section we build a model for the predictive distribution $p(x_{new}|\theta_{new})$ of a new publication's demographics x_{new} given its topic proportion θ_{new}. In order to do this, we first propose a probabilistic model of the reading habits (time spent in different topics) conditional on readers' demographics. This model is subsequently used via Bayes' theorem to infer a reader's demographic from the reading habits.

3.2.3.1 Generative model for reading habits

Assume that the reading habits $\overline{\psi}_n$ are known for every user, and that they follow a Dirichlet distribution with parameters α, $Dir(\bar{\psi}_n|\alpha)$.

To incorporate demographic covariates we model reading habits by the following Dirichlet regression model

$$\overline{\psi}_n|x_n \backsim Dir[(\alpha_1(x_n), \alpha_2(x_n), ..., \alpha_H(x_n)]$$

where

$$ln(\alpha_1(x_n)) = \gamma + \beta_1' x_n$$

$$ln(\alpha_2(x_n)) = \gamma + \beta_2' x_n$$

$$\vdots$$

$$ln(\alpha_H(x_n)) = \gamma + \beta_H' x_n$$

Given the known topic proportions and observed reading habits in a sample we can learn the parameters of the above model, $\gamma, \beta_1, \beta_2, ..., \beta_H$ using maximum likelihood or Bayesian inference. The log-likelihood of this parametrized model is given by:

$$l_H(\bar{\psi}|\alpha) =$$
$$\textstyle\sum_{n=1}^{N} \ln\Gamma \left[\sum_{h=1}^{H} \alpha_h(x_n)\right] - \sum_{n=1}^{N} \sum_{h=1}^{H} \ln\Gamma\left[\alpha_h(x_n)\right] + \sum_{n=1}^{N} \sum_{h=1}^{H} \left[\alpha_h(x_n) - 1\right] \ln(\bar{\psi}_h)$$

3.2.3.2 Predicting demographics of a new user from observed reading habits

We can compute the predictive distribution of an unknown user's demographics given his or hers observed reading habits $\bar{\psi}_{new}$ using Bayes Theorem

$$p(x_{new}|\bar{\psi}_{new}) \propto p(\bar{\psi}_{new}|x_{new})p(x_{new})$$

Here $p(x_{new})$ is the marginal distribution of the demographics, which can be estimated directly from the observed data (e.g. proportion of females aged 25-30). Figure 3.10 displays the heatmap of the distribution of the grid of ages/gender combinations.

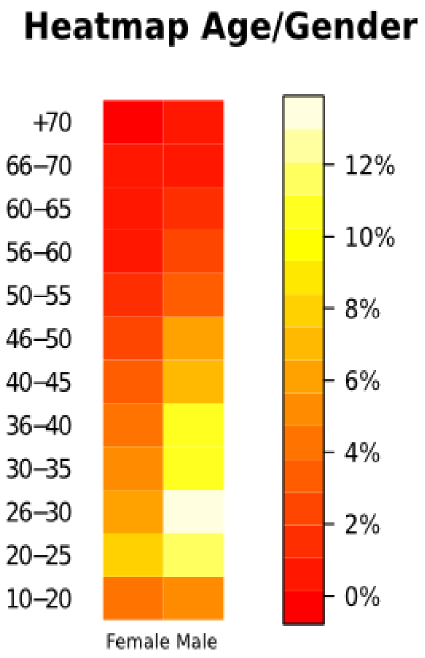


Figure 3.10: Heat map Age and Gender.

The R-package *DirichletReg* (Maier 2014) was used to estimate the model in sec. 3.2.3.1. To make predictions for a set of test users, we can, for example, use a likelihood-based approach where we first calculate the αs for each topic using the model's estimates. From a grid as seen in Figure 3.10, 24 combinations of gender and age range are available as demographic options $x_{new,i}$, $i = [1, \ldots, 24]$. e.g. $x_{new,1} = (male, "10 - 20")$, where the first 12 options belong to male gender, and the last 12 to female gender. Therefore we can generate 24 sets of different $\hat{\alpha}$s, one for each combination of gender and age range. We can now calculate the densities/likelihood of the observed proportion vector $\bar{\psi}_{new}$, given each demographic option/set of $\hat{\alpha}$s, $p(\bar{\psi}_{new}|\hat{\alpha}_i, x_{new,i})$. Those are then compared per observation and the relative frequencies adjusted with the prior knowledge $p(x_{new,i})$, in order to reach the predicted gender and age $p(x_{new,i}|\bar{\psi}_{new})$.

Given the marginal distribution of the gender, we can calculate the marginal probability of being male as

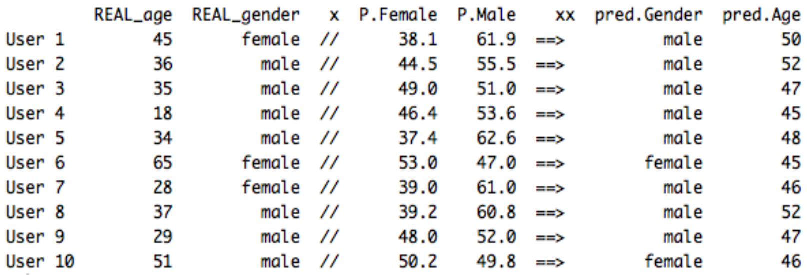
$p_{u_n}(male|\bar{\psi}_{new}) \propto \sum_{i=1}^{12} p(\bar{\psi}_{new}|x_{new,i})p(x_{new,i})$

and consequently predict its gender. Consolidating in a vector $\mathbf{p}(\mathbf{x_{new}}|\bar{\psi}_{\mathbf{new}})$, the 24 $p(x_{new,i}|\bar{\psi}_{new})$, we can now multiply by a vector $\mathbf{\Lambda}$, that includes the mean value of each age range twice:

$\mathbf{\Lambda} = [15, 22.5, 27.5, \ldots, 80, 15, 22.5, 27.5, \ldots, 80]$.

We can predict the user's age by $Pred_{u_n}(Age|\bar{\psi}_{new}) = \mathbf{p}(\mathbf{x_{new}}|\bar{\psi}_{\mathbf{new}}) \cdot \mathbf{\Lambda}'$

In Figure 3.11 you can see an example output of the real demographics and the predicted demographics.



```
         REAL_age  REAL_gender   x  P.Female  P.Male    xx  pred.Gender  pred.Age
User 1         45       female  //      38.1    61.9   ==>         male        50
User 2         36         male  //      44.5    55.5   ==>         male        52
User 3         35         male  //      49.0    51.0   ==>         male        47
User 4         18         male  //      46.4    53.6   ==>         male        45
User 5         34         male  //      37.4    62.6   ==>         male        48
User 6         65       female  //      53.0    47.0   ==>       female        45
User 7         28       female  //      39.0    61.0   ==>         male        46
User 8         37         male  //      39.2    60.8   ==>         male        52
User 9         29         male  //      48.0    52.0   ==>         male        47
User 10        51         male  //      50.2    49.8   ==>       female        46
```

Figure 3.11: Example output of Dirichlet regression prediction.

The R package presents several limitations when applied to such high-dimensional compositional data. Long processing times were experienced in every model fit, therefore a smaller training set was required. Instead of taking the original 57'000 training split, smaller sample sets of about 1'000 users were required to be considered for the fit in order to avoid extremely long calculations. Results are then averaged.

We already demonstrated the difficult separation as seen in Figure 3.1 where we just plot the proportion distribution of one label by gender. In the previous models we show topic proportions from an LDA space by 150 topics and one by 84 labels. A third kind of topic vector was now introduced in order to reduce dimensionality and by so increase computational efficiency. *Top Labels*, are introduced by aggregating related labels to high level meta topics. For instance, new label Society and Government group together individual labels of Law, Politics, Military, Crime, Police and Society. This new topic vector has only 20 topics, reducing to approximately a quarter of the original 84 labels, and almost 9 times the original 150 topics. Result show that a reduced topic vector provides a non significant performance difference (Figure 3.12).

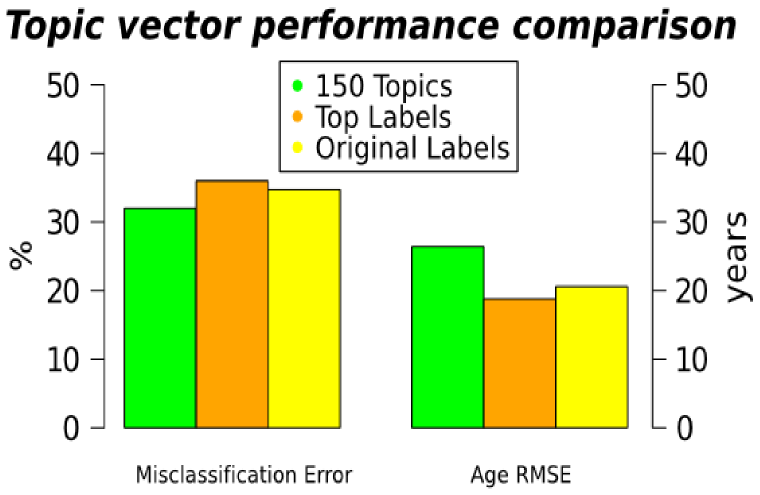


Figure 3.12: Performance comparison between probabilistic model based on different LDA spaces.

The model allows the inclusion of an observation weight variable, assigned according to frequency. This means than an observation whose weight value is 2 is counted twice in the fit in respect to one with weight value of 1. We tested the inclusion of a weight value based on the total reading time of an user, weighting big readers as 3, average users as 2 and casual users, as 1. No real improvement was observed, so the option was discarded.

A different classification can be performed based on the most likely age range segment instead of the Predicted age and gender. This option classifies a user as the segment with higher probability from the posterior distribution. After testing, this option provides a similar gender misclassification error with worse age RMSE results than the ones observed before.

Lastly, when we review the model fit we can observe that in most of the fitted models the coefficients for age and gender variables remain non significant (high p-value). This explains the poor results observed: most of users are classified as male, as the variable related to gender remain non significant in the regression. Similarly, the predicted age distribution does not present high variability.

3.2.3.3 Other uses

Predicting reading demographics of a new document

Given a new document with topic proportions θ_{new}, we might want to know its potential readers (demographics). One idea is to use the above presented probabilistic model to represent the new document as a new reader with habits $\bar{\psi}_{new} \simeq \theta_{new}$, with unknown demographics. We can now calculate the posterior distribution $p(x_{new}|\theta_{new})$, the predicted gender based on $p_{User_n}(male|\theta_{new})$ and the predicted age $Pred_{User_n}(Age|\theta_{new})$. Other analysis could be calculated, for instance the probability of a document being read by a user with certain gender and age range which can be used for targeting ads. Targeting an ad for a female between 40 and 50 years of age amounts to looking for documents with the desired predicted gender and expected age, or just set a minimum threshold for the sum of $p(x_{new} = (female, "40{-}45"|\theta_{new})$ $+p(x_{new} = (female, "45-50"|\theta_{new})$ and search all documents that comply with that rule.

Recommend content for a specific demographic

The model allows to predict relevant content for a specific gender and age range group $x_{new,i}$, using the generative model based on $p(\theta_{new}|\hat{\alpha}, x_{new,i})$ (sec. 3.2.3.1). Once we calculate the $\hat{\alpha}$ vector according to the model parameters, it is possible to generate different draws of a document topic proportions θ_{new} and find in the original LDA space K-nearest neighbors of real document whose topic proportions are closer to the ones generated. It is also possible to simply search for which segment these coordinates of the LDA space belong. For instance, recommended content for male aged 50-55 provides a specific topic proportion vector that could be used to find near publication neighbors and along with their respected segment. Once those segments have been singled out, more related content can be recommended.

4 Discussion

4.1 Analysis of results

The first two models presented are nearest neighbor algorithms based on publications read by the user. We propose distances that measure the proximities between documents or user by placing these objects in a space of topic proportions obtained from a Latent Dirichlet Allocation (LDA) analysis. These models depend on the number of neighbors to be considered when reviewing the similar content or users in order to make inferences about the user's age and gender. In both cases, increasing the number of neighbors increases the age prediction accuracy. The results about the gender classification are not conclusive, as no clear trend is seen in any of the models when the number of neighbors is varied.

A modification of the first model was then considered to reduce computational workload. Instead of predicting the demographics for a user based on a weighted average of the demographics of nearby users in LDA space, the modified model predicts the demographics of a user to be the average demographics within the user segment in LDA-space. In this model there is no need to compare each document every time or to rank neighbors. This simplified model, somewhat surprisingly, improves the quality of the prediction. This is most likely because of more stable estimates of the distribution of the demographic data to infer from. The best result reached by the model based on documents LDA positions and their neighbors is around 67% accuracy in gender prediction in a model with K=10 neighbors. Using the segments we can considerably reduce computer processing times, while at the same time reach up to 72% accuracy in gender prediction. The RMSE for age predictions is also reduced to only 10.33, which is considerably below the best result seen of 12 when working by individual documents and with $K = 30$ neighbors.

The second LDA space model is based on the individual topic proportion vector of a user. This model has the same drawback as the previous model since a user must be compared to all other users in order to find related users, resulting in long computation times. Nevertheless this is done only once by query and not by every publication in the reading list, and therefore it is faster than the former. Three types of LDA space were tested: the space based on the 150 topics provides better predictions than a variant of the LDA model that uses Wikipedia labels to improve the interpretation of topics. Under the 150 topics LDA space, best results were seen with higher number of neighbors considered, reaching an RMSE of 12.5 age, and about 74% accuracy for gender prediction.

The final approach compared three user vector proportions: 150th topics, 84 Labels and third using the meta labels, in a Dirichlet regression framework. This model can be used to predict user demographics from the distribution of a reader's reading time over a set of topics, and also generate content tailored to a specific age-gender combination. The performance of this approach falls behind the nearest neighbor algorithms, but as the model fit presented many difficulties in handling the size of the data, this evidence is not conclusive.

In conclusion, the best performance is observed when using the model that takes in consideration the reading list by Segments, predicting the user's age and male probability by a weighted average algorithm that considers the respective demographic of the readers of the related content. This approach combines in comparison good results along with the fastest computation times.

4.2 Future work

The higher proportion of male users versus female users resulted in a systematic misclassification into masculine labels. By using a gender balanced training sample this effect might be corrected. As an example, we tried re running the nearest neighbor model based on DNA of a user on a training sample of 30'000 with even proportion of male and female users. Using a random sample of 100 users for test (from the original unbalanced population, in order to make the results comparable to the ones described in the previous chapter) we can observe in Table 4.1 that the Misclassification error is now reduced to 30%, with a better Age prediction performance as well (vs. Table 4.2). If we compare the precision, in this case we can observe better results, both an increased precision in female prediction (60% vs 50%) and male prediction (78% vs 70%).

Still, further analysis and re run of all the presented models is required under balanced training and test samples before jumping to conclusions. The F-score is a measure that combines the precision of both predictions (Goutte and Gaussier 2005), from the tables below it is around 78% in both cases. As a comparison, on the model by Segment presented it was about 82%. Another option could be to use the full training sample but with a weight correction in the way the male probability is calculated (see Equation 3.2) where a different weight is assigned to the time of a woman than from a man; an usual adjustment would be the inverse of the specific gender proportion in the training sample.

Real Gender	Predicted Male	Predicted Female
Male	51	16
Female	14	19
Misclassification Error	30%	
RMSE for Age	11.48	

Table 4.1: Performance of random test set on **balanced** training sample. Model using LDA of User.

Real Gender	Predicted Male	Predicted Female
Male	59	8
Female	25	8
Misclassification Error	33%	
RMSE for Age	12.05	

Table 4.2: Performance of same random test set on **unbalanced** training sample. Model using LDA of User.

The predictive methods proposed in this thesis are illustrated on a small dataset collected over a few months of user reading behavior. Further analysis is required with a larger sample to more rigorously verify the usefulness of the methods. There could be a seasonal effect, which we might be missing by only taking in consideration January reads. Working with this data size already created many problems when fitting a model or running an algorithm; the use of more powerful machines is required when the training sample is increased, but also to re-run previous models to reach their full capabilities. Due to computational difficulties, the probabilistic model could not be trained on the full 57'000 user proportion vectors, and results vary considerably between different training samples. Also, once the data sample is increased the users and documents' LDA will be further populated and as a result new neighbors could be found. Each document and segment will also provide a larger distribution of age and gender readers, which will increase the reliability of any prediction. The availability of reliable demographic data is clearly crucial for the success of the proposed methods. Most of social media sites restrict sharing private information. Users also tend to choose not to populate these fields in many log in credentials forms. Last, information might still be fake or wrong. It is therefore important to increase both the quantity and quality of the data.

The topic distribution of a document is based on an LDA analysis of the entire document using 150 topics as basis. The provider of the dataset, Issuu, is currently working on an LDA analysis by page, where a better understanding could be achieved of documents with mixed content. For instance publications as newspapers, covering multiple and different material, have a more flat topic distribution as no single topic stands out. This LDA by page will also allow to track the real interest of users when analyzing the reading patterns by page and therefore have a better understanding of users likes and dislike by topic. This more detailed level of the data opens up the possibility of a more elaborate predictive model.

This thesis measures the similarity of users by vector proportions in the LDA space. Another option that could be analyzed is the option of collaborative filtering (Sarwar et al. 2001) by mapping users according to the publication and/or segments. Yet another option is to find users based on same or similar documents read. For instance, if we count how many similar publications in their reading list a user has in common with another one, we can rank their similarity. This approach could be based by close by documents in a given LDA space or by segments. Figure 4.1 shows the example of user A and B; user B shares similar publications with A on 4/6 (67%) of its reading list.

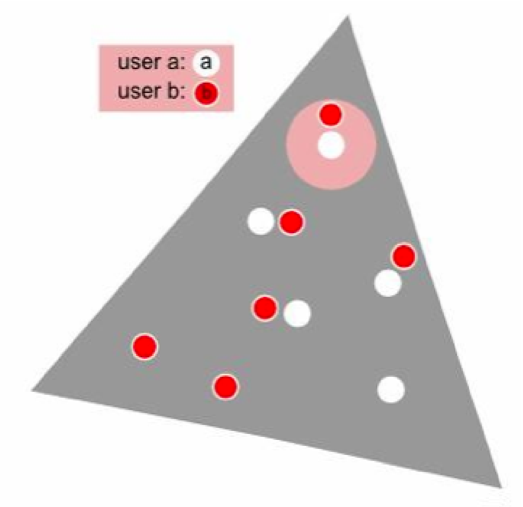


Figure 4.1: Example of an algorithm to match users based on proportion of similar reads.

Another approach considered was association analysis (Toivonen et al. 1996): mining the frequent sequences of segments or topics read with the respective user's age and gender. Frequent item sets are found from the set of all sequences of reading list with minimum support, to be defined. Our goal would be to identify all the rules $x \rightarrow y$ that have a minimum confidence. Confidence can be defined as the conditional probability that an item set having x, also has y is $P(x;y)|P(x)$. In our case we would be interested in the probability of e.g. $(read : cars) \rightarrow (male)$. A tree structure of rules (Figure 4.2) could be created and used to analyze how marginal probability evolution changes as long a reading list is extended. For instance $p(male|read : cars) = 0.6$, $p(male|read : cars, trains) = 0.8$, therefore reading a train related publication if read as well a cars related publication increase the probability of being male by 20% or a marginal 33%. Same scenario could be created for both age and gender, or only age from distinct age ranges. A set of rules can be defined and organized by an algorithm, where given a user's reading list, the assigned demographic (age and gender) is the one with higher probability.

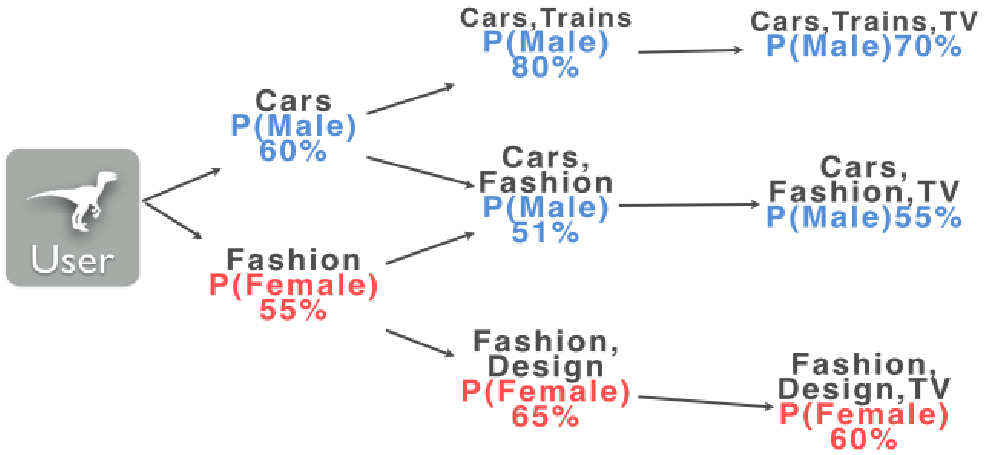


Figure 4.2: Probability tree based on association analysis.

From the exploratory analysis we observed that certain labels are more representative than others to identify the user's demographics. A model that considers feature selection of the topics, or labels, could provide a framework to sequentially add extra topics, or labels, according to their significance. We might reach a more efficient model with less parameters than the models introduced in this paper, and with a better predictive performance. This feature selection framework could be even more important when LDA is inferred on a per-page basis as the size of information would be significantly increased.

Finally, given a continuous (age) and discrete distribution (gender) as response, a joint distribution based on copula model could be evaluated. Single regression models could then be joined under a copula framework. Bayesian copula models provide an interesting approach that might lead to better results, see Murray et al. (2013).

5 Conclusions

This paper presents three different approaches to the problem of how to predict a user's demographics from his or her reading habits. We apply the methods on a complex dataset of recorded reading times on a web publication service. The data is challenging as no clear separation can be easily seen according to gender, time or topics read. Nevertheless the work presents evidence that the model can reach an acceptable level of prediction accuracy even on difficult and noisy data. These results can be used to assist in many objectives such as user profiling, recommendation systems or ad targeting.

If we compare the three presented methods, we introduce two models using nearest neighbors based on the reading habits and a third approach based on a Bayesian probabilistic model, the tracking of a user's reading list by segment combines computer scalability with faster and better results. Interesting results were obtained even though we used only a tiny sample of all Issuu traffic data. The model have a clear preference to label users as male, given the higher proportion of male users (an adjustment for this type of bias should be analyzed in future work). We obtained the best results for models that include the demographic information of many users in the prediction for the demographics of a given user, either by increasing the number of neighbors or by aggregating all users by Segment. Furthermore, the novelty of creating an individual description of a user's reading habits as the distribution of reading time over LDA topics provides a framework for interesting analysis. Clusters of similar users could be found, and therefore segments of users be set. Also, a probabilistic approach was developed using Dirichlet regression. Recommendations of content can be generated by matching the generated user's reading habits to the proportions of a document topic distribution. We can trace this proportion into the respective LDA space and find related publications.

Bibliography

Bach, F. R., Heckerman, D., and Horvitz, E. (2006). Considering cost asymmetry in learning classifiers. *The Journal of Machine Learning Research*, 7:1713–1741.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Brodley, C. E. and Utgoff, P. E. (1995). Multivariate decision trees. *Machine learning*, 19(1):45–77.

Camargo, A. P., Stern, J. M., Lauretto, M. S., Goyal, P., Giffin, A., Knuth, K. H., and Vrscay, E. (2012). Estimation and model selection in dirichlet regression. In *AIP Conference Proceedings-American Institute of Physics*, volume 1443, page 206.

Garera, N. and Yarowsky, D. (2009). Modeling latent biographic attributes in conversational genres. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 710–718. Association for Computational Linguistics.

Goutte, C. and Gaussier, E. (2005). A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In *Advances in information retrieval*, pages 345–359. Springer.

Hsu, C.-W., Chang, C.-C., Lin, C.-J., et al. (2003). A practical guide to support vector classification.

Hu, J., Zeng, H.-J., Li, H., Niu, C., and Chen, Z. (2007). Demographic prediction based on user's browsing behavior. In *Proceedings of the 16th international conference on World Wide Web*, pages 151–160. ACM.

Joachims, T. (1998). *Text categorization with support vector machines: Learning with many relevant features*. Springer.

Koppel, M., Argamon, S., and Shimoni, A. R. (2002). Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4):401–412.

Maier, M. J. (2014). Dirichletreg: Dirichlet regression for compositional data in r.

Murray, J. S., Dunson, D. B., Carin, L., and Lucas, J. E. (2013). Bayesian gaussian copula factor models for mixed data. *Journal of the American Statistical Association*, 108(502):656–665.

Nguyen, D., Smith, N. A., and Rosé, C. P. (2011). Author age prediction from text using linear regression. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 115–123. Association for Computational Linguistics.

Peersman, C., Daelemans, W., and Van Vaerenbergh, L. (2011). Predicting age and gender in online social networks. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, pages 37–44. ACM.

Pennebaker, J. W. and Graybeal, A. (2001). Patterns of natural language use: Disclosure, personality, and social integration. *Current Directions in Psychological Science*, 10(3):90–93.

Pennebaker, J. W. and Stone, L. D. (2003). Words of wisdom: language use over the life span. *Journal of personality and social psychology*, 85(2):291.

Rao, D., Yarowsky, D., Shreevats, A., and Gupta, M. (2010). Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pages 37–44. ACM.

Rehurek, R., Sojka, P., et al. (2010). Software framework for topic modelling with large corpora.

Rosenthal, S. and McKeown, K. (2011). Age prediction in blogs: A study of style, content, and online behavior in pre-and post-social media generations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 763–772. Association for Computational Linguistics.

Santosh, K., Bansal, R., Shekhar, M., and Varma, V. (2013). Author profiling: Predicting age and gender from blogs?notebook for pan at clef 2013. *Forner, et al.(eds.)[15]*.

Sarwar, B., Karypis, G., Konstan, J., and Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*, pages 285–295. ACM.

Schler, J., Koppel, M., Argamon, S., and Pennebaker, J. W. (2006). Effects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, volume 6, pages 199–205.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E., et al. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791.

Toivonen, H. et al. (1996). Sampling large databases for association rules. In *VLDB*, volume 96, pages 134–145.

Zhong, E., Tan, B., Mo, K., and Yang, Q. (2013). User demographics prediction based on mobile data. *Pervasive and Mobile Computing*, 9(6):823–837.

Upphovsrätt

Copyright

LIU-IDA/STAT-A--14/001—SE