USER ACQUISITION AND ENGAGEMENT IN DIGITAL NEWS MEDIA

HEIDAR DAVOUDI

A DISSERTATION SUBMITTED TO
THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN DEPARTMENT OF ELECTRICAL ENGINEERING
AND COMPUTER SCIENCE
YORK UNIVERSITY
TORONTO, ONTARIO

DECEMBER 2018

Abstract

Generating revenue has been a major issue for the news industry and journalism over the past decade. In fact, vast availability of free online news sources causes online news media agencies to face user *acquisition* and *engagement* as pressing issues more than before. Although digital news media agencies are seeking sustainable relationships with their users, their current business models do not satisfy this demand. As a matter of fact, they need to understand and predict how much an article can engage a reader as a crucial step in attracting readers, and then maximize the engagement using some strategies. Moreover, news media companies need effective algorithmic tools to identify users who are prone to subscription. Last but not least, online news agencies need to make smarter decisions in the way that they deliver articles to users to maximize the potential benefits.

In this dissertation, we take the first steps towards achieving these goals and investigate these challenges from data mining /machine learning perspectives. First, we investigate the problem of understanding and predicting article engagement in terms of dwell time as one of the most important factors in digital news media. In particular, we design

data exploratory models studying the textual elements (e.g., events, emotions) involved in article stories, and find their relationships with the engagement patterns. In the prediction task, we design a framework to predict the article dwell time based on a deep neural network architecture which exploits the interactions among important elements (i.e., augmented features) in the article content as well as the neural representation of the content to achieve the better performance.

In the second part of the dissertation, we address the problem of identifying valuable visitors who are likely to subscribe in the future. We suggest that the decision for subscription is not a sudden, instantaneous action, but it is the informed decision based on positive experience with the newspaper. As such, we propose effective engagement measures and show that they are effective in building the predictive model for subscription. We design a model that predicts not only the potential subscribers but also the time that a user would subscribe.

In the last part of this thesis, we consider the paywall problem in online newspapers. The traditional paywall method offers a non-subscribed reader a fixed number of free articles in a period of time (e.g., a month), and then directs the user to the subscription page for further reading. We argue that there is no direct relationship between the number of paywalls presented to readers and the number of subscriptions, and that this artificial barrier, if not used well, may disengage potential subscribers and thus may not well serve its purpose of increasing revenue. We propose an adaptive paywall mechanism to balance

the benefit of showing an article against that of displaying the paywall (i.e., terminating the session). We first define the notion of cost and utility that are used to define an objective function for optimal paywall decision making. Then, we model the problem as a stochastic sequential decision process. Finally, we propose an efficient policy function for paywall decision making.

All the proposed models are evaluated on real datasets from The Globe and Mail which is a major newspaper in Canada. However, the proposed techniques are not limited to any particular dataset or strict requirement. Alternatively, they are designed based on the datasets and settings which are available and common to most of newspapers. Therefore, the models are general and can be applied by any online newspaper to improve user engagement and acquisition.

Acknowledgements

I would like to express my deepest appreciation to my advisor Prof. Aijun An. I consider myself fortunate to have her as a supervisor in working toward the PhD. In fact, going along the path toward this dissertation was coincident with a lot of challenges in my life. I would like to say that achieving this goal was not possible without her support, advice and encouragement.

I would like to extend my sincere thanks to my committee members Prof. Natalija Vlajic and Steven Wang for their valuable advice. The members of data mining lab have been a source of friendship and inspiration. I had great pleasure of working with them and I would like to appreciate them. I am also thankful to the administrative and technical staff of our department at York university.

I would like to thank The Globe and Mail for providing the datasets used in this work and for providing me with an internship opportunity. This work is funded by Natural Sciences and Engineering Research Council of Canada (NSERC), The Globe and Mail, and the Big Data Research, Analytics and Information Network (BRAIN) Alliance

established by the Ontario Research Fund - Research Excellence Program (ORF-RE).

Last but not least, I am extremely grateful to my family, Zhamak, Kina and Ashkan for their support. The completion of my dissertation would not have been possible without their support.

Table of Contents

List of Tables

List of Figures

1 Introduction

1.1 Motivation

Users are traditionally the most important asset for digital newspapers. In fact, no prosperity is possible without successful user *acquisition*, and *engagement*. However, developing strong relationships with users is not a trivial task in the competitive world of digital media. Therefore, online newspapers need to build the effective strategies to identify potential subscribers, make sure that users have positive experiences, and keep the relationships with them strong. In order to achieve the aforementioned objectives, digital newspapers should gain insight into users needs and behaviors. This is where data mining/machine learning techniques come to play. In this dissertation, we focus on user acquisition and engagement as two challenging problems in the news domain. To that end, we frame the challenges into different research problems and proposed effective solutions for them accordingly.

Article Engagement: In the online news domain, building successful relationships with users without properly engaging them is not possible. User engagement reflects the qual-

ity of user experience when interacting with the media [40]. Among the engagement measure in web analytics *dwell time*, defined as the amount of time which a user spend on an article, is one of the most important ones [15, 26, 35]. In fact, understanding and predicting an article's dwell time is an interesting problem which has drawn little attention in the news domain from the data mining perspective. In other words, the interesting question is: *can we predict an article's dwell time based on its content?* An answer to this question can provide benefits for many applications in the news domain. For example, it can be used to organize advertisement on the article page (e.g., putting more advertisements on pages for more engaging articles to maximize the revenue). However, predicting the article engagement needs understanding the interplay between the article constituents such as events, people, emotions involved in the article story. We consider these elements and propose a deep neural network architecture to utilize them for dwell time prediction.

Subscription Prediction Model: Data mining techniques can identify and understand user behaviors, preferences and needs [48, 64]. Moreover, models in data mining can assess the value of customers and uncover the reasons behind their behaviors [11]. The successful achievements of data mining in many domains raise an interesting question: *can data mining methods help in customer acquisition by timely identification of valuable customers with high likelihood to subscribe*? Despite the importance of this question and potential benefits of data mining techniques in identifying potential subscribers, there is

no comprehensive research on using such techniques for user acquisition in the digital news domain. Note that identifying potential subscribers is important since it allows the management to launch an acquisition campaign in advance, and conducting a successful campaign results in significant profits and benefits. Most of studies to date have been focusing on the other domains (e.g., e-commerce) where the main revenue is based on individual purchases rather than subscriptions. Moreover, most of efforts have considered recommendation systems as the personalized one-to-one marketing solutions rather than subscription prediction models for the user acquisition problem.

Adaptive Paywall Mechanism: Many online newspapers across the world utilize a *paywall mechanism* to persuade users to subscription. In a traditional paywall solution, digital newspapers offer a fixed number of free articles in a period of time (e.g., a month) to visitors then direct them to a subscription page (i.e., paywall) for further reading. They use this mechanism as a tool to request subscriptions in order to achieve successful user acquisition, and increase revenue. However, in most cases there is no direct relationship between the number of paywalls presented to readers and the number of subscriptions. In fact, if this mechanism is not used smartly, it may disengage potential subscribers and reduce revenue. Clearly, this mechanism does not consider the articles that a user has already read nor the potential articles which the user may read in the future, and treats all users the same. However, users are different in terms of news consumption and subscription potentials. In fact, blocking a user from reading articles at a wrong time may

disengage the user too early or allow a non-potential subscriber to read too much content for free. As such, an interesting question is that: *can we estimate how many and what articles a particular user should be allowed to read before a paywall?* While this question is crucial to any online newspaper using the paywall mechanism, to date, it has not been investigated. In fact, answering this question is challenging since it needs an appropriate mathematical model on top of the proper criteria to make optimal decisions. We formulate finding the optimal paywall time point as a sequential decision problem, in which a decision on whether to show a paywall or not needs to be made at each time point during a reading session, and propose machine learning solutions to solve the problem. Moreover, we define the measures which can serve as the criteria in the decision process.

The rest of this chapter is organized as follows. We discuss the scope and the general problem tasks in §1.2. The proposed framework for user acquisition and engagement in digital news media is presented in §1.3. We provide the general picture of the main thesis components, their relationships and dependencies accordingly. Moreover, the proposed models and research contributions are summarized in that section. The Globe and Mail data collection platform and the datasets, which are utilized in the other chapters, are introduced in §1.4. Finally, we present the general structure of the dissertation, and introduce different parts of the thesis in §1.5.

1.2 Problem Statement and Scope

User acquisition and engagement are quite challenging and complicated issues. The terms of user acquisition and engagement are broad in the news domain as they can be studied from different discipline perspectives (e.g., journalism, marketing). In this research we consider them from data mining and machine learning point of views. Furthermore, applicability of approaches can be strictly restricted by the availability of data in the organization and domain (e.g., there are many restrictions on using users' data due to privacy and confidentiality issues). Therefore, we consider the problem of user acquisition and engagement in an environment/setting which is mostly common among all digital newspapers.

The *domain of interest* of this thesis is a digital news medium/newspaper where the users' interactions (for both subscribed and non-subscribed users) with the news portal are captured and collected. Moreover, the *data source* contains the articles published by the news agency, and clickstreams comprising sequence of articles which users have visited. A detailed description of clickstreams will be given in §1.4.

Given the domain of interest and data source, our goal is to bridge the gap between the journalism and data mining/machine learning communities. We aim to frame some current issues in digital newspapers as the feasible data mining/machine learning problems and proposed appropriate solutions to them accordingly. To that end, we focus on

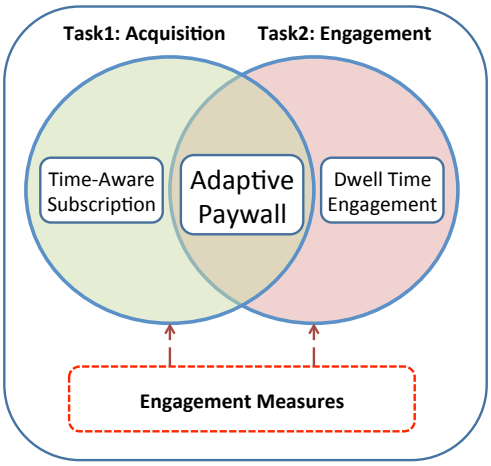


Figure 1.1: Relationship between defined tasks and the proposed models.

designing the cutting-edge data driven approaches for the following general tasks:

- **Task1 (Acquisition):** Developing algorithmic/machine learning tools to serve as a decision support model to increase the likelihood of users subscriptions.

- **Task 2 (Engagement):** Designing data exploratory and predictive models for predicting and understanding the article engagement.

It is worth to mention that the resulting models can be fully automated (e.g., proposed adaptive paywall mechanism) or work as a decision support system (e.g., time-aware subscription prediction model) in combination with other efforts (e.g., conducting user acquisition campaigns) to achieve the goals. Figure 1.1 shows the relation between these tasks and the proposed models. As can be seen, engagement measures can serve both tasks. Moreover, the adaptive paywall model can enhance both acquisition and engage-

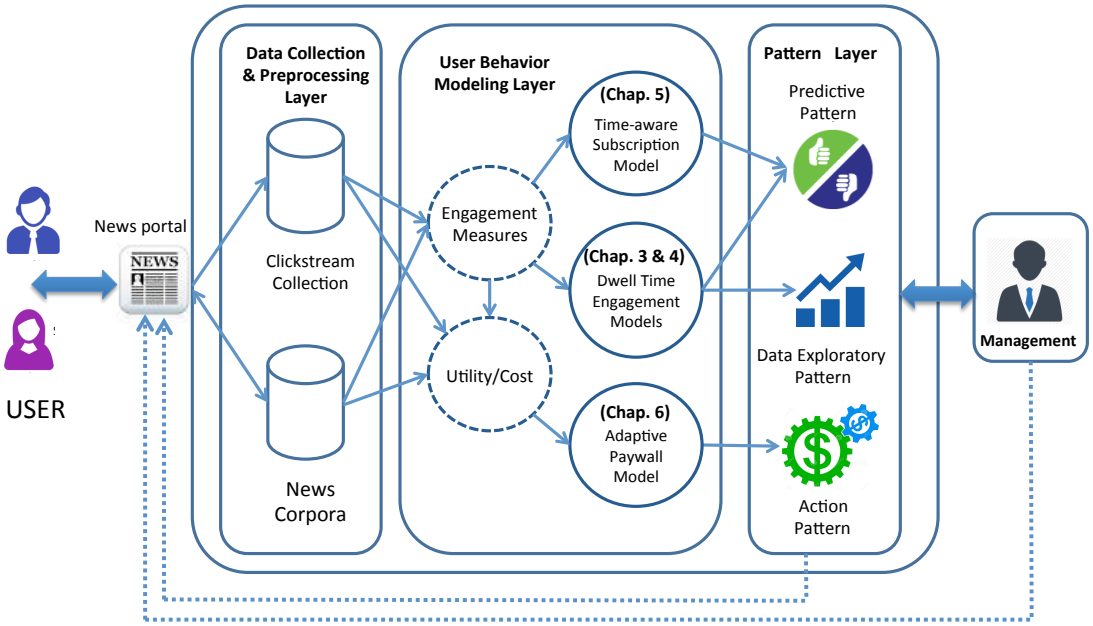


Figure 1.2: Framework for user acquisition and engagement.

ment tasks by using different criteria for making paywall decisions.

1.3 Proposed Framework and Research Contributions

Figure 1.2 shows the general overview of the proposed framework for the user acquisition and engagement as well as the structure of this dissertation. The general picture of the research components and their relationships are illustrated accordingly. The framework is categorized in three layers: *data collection and preprocessing*, *user behavior modeling*, and *pattern layer*. This categorization facilitates studying the relevant components and gives a clear picture of the proposed framework.

The data in the data collection and preprocessing layer consist of user clickstreams and news corpora. Most of digital newspapers have these two data components in their data analytics pipeline. The data collection platforms such as Omniture[1] or Google Analytics[2] usually operate in this layer. They usually record every single interaction of users with the online news portal, which forms user clickstreams. Moreover, there is a repository of articles in this layer usually managed by a text search engine such as solr[3]. We discuss the data collection and preprocessing procedure details in §1.4 briefly.

The pattern layer demonstrates the outcomes types of the proposed models. As can be seen, different models result in different outcomes. In case of the Time-aware Subscription Prediction (TASP) model, the outcome is the predictive pattern forecasting the potential subscribers. This outcome can be used by the management to facilitate the acquisition process (e.g., by conducting a marketing campaign). The dwell time engagement models are comprised of both data exploratory and predictive frameworks. The data exploratory outcomes uncover the relationships between the article story constituents (e.g., events) and the article dwell time distribution. Moreover, the predictive dwell time model results in the article dwell time forecast based on the article content. The adaptive paywall model is different from the others as it produces actionable patterns. More pre-

[1] https://my.omniture.com

[2] https://marketingplatform.google.com/about/analytics/

[3] http://lucene.apache.org/solr/

cisely, the result of this model is a set of decisions or the policy indicating the optimal paywall time points for different reading sessions.

The user behavior modeling layer is comprised of the proposed models. The utility and cost are the foundations for the the adaptive paywall component and are used exclusively by the adaptive paywall model while engagement measures are used by both dwell time engagement and the Time-aware Subscription Prediction (TASP) models. The notion of utility and cost are very general. However, it is possible to define utility based on the engagement measures.

The contributions of the proposed models fall into different categories which are summarized in the subsequent sections.

1.3.1 Data Exploratory Framework for Dwell Time Engagement

We design a novel content-based data exploratory framework to understand the dwell time engagement of the articles in news media. In particular, we study the main elements of article story (e.g., event), and combine them with the reliability analysis to provide insight on the role of these elements in article engagement. The contributions can be summarized as follows:

- We propose a new framework for analyzing user engagement of articles under the scope of different factors such as topics, events, and emotions.

- We design a novel model which combines the reliability analysis and text mining approaches to uncover the dwell time engaging patterns and the role of different factors in them.

- We conduct the experiments on a real dataset from The Globe and Mail articles and extract the patterns showing relationships between the article dwell times and main elements of the story.

1.3.2 Predictive Model for Dwell Time Prediction

We propose a new model based on the neural network architecture to predict the article dwell time engagement. The proposed model outperforms the state-of-the-art model in the prediction task. The contributions are as follows:

- We propose a novel model for article dwell time prediction based on the deep and wide neural network architecture and feature augmentation. While there are few studies on predicting Web pages dwell times, this is the first study that one leverages the deep neural network capabilities to build a content-based predictive model for this purpose.

- The extensive experiments are conducted to show the effectiveness of the proposed model against the state-of-the-art baseline approaches.

1.3.3 Time-aware User Subscription Model for Digital News Media

In order to target readers who are likely to subscribe we design a model which can effectively predict these users. Furthermore, the proposed model can predict the time that a user is likely to subscribe. The contributions are as follows:

- We define and frame the new problem of user subscription in digital newspapers and propose an effective data-driven approach to solve it. To best of our knowledge, this is the first study that defines this problem and tackles it.
- We propose user engagement measures which can serve as the predictors and design a novel probabilistic model to solve the prediction task.
- We conduct the extensive experiments on The Globe and Mail clickstream dataset and show that proposed method outperforms other baseline models.

1.3.4 Adaptive Paywall Mechanism for Digital News Media

To address the current paywall model drawbacks, we propose a adaptive paywall mechanism which makes a balance between delivering articles and presenting the paywall. The summary of contributions along this line is as follows:

- We define and formalize the new problem of adaptive paywall for digital newspapers. To best of our knowledge, this problem has not been investigated and this study is the first one that suggest a data-driven solution for it.

- We cast the problem into a sequential stochastic decision process and propose a novel and effective approach to solve it.

- The proposed framework is evaluated using a real dataset from The Globe and Mail dataset. The experimental results show significant improvement over traditional and other baseline models.

1.3.5 The Publications

The outcomes of this study have been published in the top tier conferences proceedings in the field of data mining such as the proceedings of the 24'th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [18], and proceedings of the 2017 SIAM International Conference on Data Mining [19]. Moreover, parts of this work are under review in 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2019), and the top journal in the field such as IEEE Transactions on knowledge and data engineering.

1.4 The Globe and Mail Dataset

Every time a user reads an article, watches a video or generally takes an action in a website, the action is recorded as a *hit* in the log file of the website. In data collection

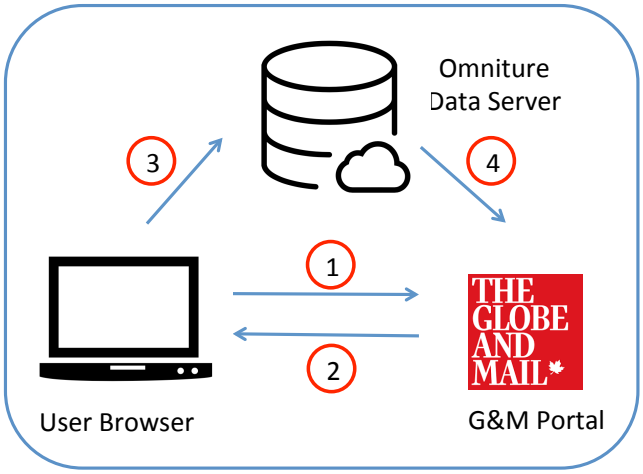


Figure 1.3: Data Collection Platform

frameworks (e.g., Omniture by Adobe which is used by The Globe and Mail[4]), a hit is simply represented as a row (record) in the log file and contains rich information about the visitor and her/his actions. Typically, a hit contains information like, date, time, user id, user environment variables (e.g., browser type), id of article, and special events of interest like subscription, sign in and etc.

Figure 1.3 illustrates the collection procedure for The Globe and Mail dataset. The sequence of interactions are as follows: (1) the user browser requests an article from The Globe and Mail news portal, (2) The Globe and Mail Web server responds by sending the requested article including a small javascript code, (3) each user interaction with the article pages will be sent by the javascript code to the 3'rd party server (i.e., Omniture data server), (4) the data collection servers send the collected data periodically to The

[4]https://www.theglobeandmail.com

Table 1.1: Attributes in The Globe and Mail Dataset.

Attributes	Descriptions
Traffic Variables:	• Count the instances of specific events on a web page Example: Number of clicks on advertisements • Group the pages (logically) based on a variable set in hits. Example: User logged in/not logged in
Conversion Variables:	• Conversion variables are persistent and hold the values for a longer period time. Example: Number of articles viewed
Events:	• A point on the website which a successful event occurs. Example: subscription

Globe and Mail to be used in the data analytics pipeline.

The Globe and Mail clickstream dataset used in this thesis is composed of 264,735,412 hits over period 2014-01 to 2014-07. It contains 246 attributes which capture different aspects of interactions with users. Table 1.1 illustrates different types of attributes in The Globe and Mail clickstream dataset. Although clickstream data can provide a lot of insights into users behaviors, they are usually noisy, and contain a lot of irrelevant information. As such, we need careful data preprocessing and cleaning steps before exploring, extracting and summarizing any useful pattern. Moreover, such a data repository contains low-level interactions between the users and the portal, which makes the aggregation steps necessary before any further processing. Table 1.2 shows the general preprocessing steps which are done on The Globe and Mail clickstream dataset.

In particular, given that data are organized as hits, we roll-up the data from page view

Table 1.2: Preprocessing on The Globe and Mail Dataset.

Preprocessing	**Descriptions**
Filtering out irrelevant hits:	• We only keep the hits related to articles and important events such as the subscription.
Extracting user types:	• Users fall into two broad categories of subscribed and non-subscribed users.
Extracting events of interest:	• We extract events such as login or subscription.
Computing dwell time:	• Dwell time is calculated using two consecutive page views timestamps.
Roll-up form hit to visit and user:	• The aggregation is done using the user unique id, visit, and page id.
Converting to expressive format:	• Each user activity record is stored in a json format to facilitate the analysis in later stages.

hits to visits and then to visitors. We refer to a *visit* as a set of page views in one “session" (a session is terminated if the data collection server does not hear from the same user for 30 minutes). We use cookie and the device’s IP information which is anonymized and encoded in the data warehouse to detect the unique *visitors*. Moreover, The data collection platform records a timestamp whenever an article page is requested, so the difference between two consecutive page click timestamps is used to calculate the articles dwell times. As usual in web analytics the last article in a visit is ignored as we cannot estimate the dwell time for it. For dwell time data exploratory and predictive models (chapter 3 and 4), the articles with less than 10 views are removed resulting in 28502 articles. For the Time-aware subscription prediction model (chapter 5) , the dataset used

in the experiments contains 17,009 subscribers and 71,639 non-subscribers. Finally, for the paywall model (chapter 6), after preprocessing (i.e., aggregation), the dataset contains 4,913,423 sessions from subscribed users. We use sessions with minimum 10 articles as the test and the rest as the training set (e.g., to build the navigational graph). The test data set provide real sequences of article requests for use to evaluate the proposed method. A minimum of length 10 gives us more possible time points for making paywall decisions in a session. The numbers of sessions in the training and test sets are 4,806,204 and 107,219, respectively.

1.5 Thesis Organization

This dissertation is organized as follows. In chapter 2, we categorize and discuss the literature related to the different proposed models. User engagement measures serve as an important component in our framework since our proposed models are built on them. As such, we introduce the concept of user engagement and discuss the relevant background in chapter 3. Moreover, we focus on data exploratory analysis on articles dwell time engagement in that chapter. In chapter 4, we present the article dwell time prediction model. This model aims to predict the articles dwell times based on the main constituents involved in the stories. The goal of Time-aware Subscription Prediction (TASP) model is to predict the users who are prone to subscription in a timely manner. We explain this model in detail in chapter 5. Chapter 6 discusses the proposed adaptive

paywall mechanism for making optimal paywall decisions based on the notion of utility and cost. Finally, in Chapter 7, we review the contributions, conclude the dissertation, and summarize the future research directions.

2 Literature Review

User acquisition and engagement are generally studied under the umbrella of Customer Relation Management (CRM), which focuses on four dimensions: customer *identification*, customer *attraction*, customer *retention* and customer *development* [59]. The common goal of all is to understand the customer behaviors and maximize the customer value to the organization in the long-term. In particular, customer identification refers to task of targeting the population who are very likely to become customers. As we identify the potential customers, we need some sort of customer attraction (e.g., using direct marketing through incentives and promotions). The concept of customer retention is closely related to the user satisfaction issue whereas the most prominent approaches try to address the problem based on one-to-one marketing (e.g., personalized marketing campaigns supported by user behavior modeling, user profiling, and recommender systems). Finally, customer development refers to the set of methods that maximize customers profitability (e.g., customer lifetime value). This line of research mostly focused on the first (i.e, user acquisition) and the third (i.e., user retention through engagement) goals the in digital

news media domain.

This dissertation designs a set of novel machine learning/data mining techniques as the algorithmic tools to help digital newspapers in user acquisition and engagement tasks. The variety of areas in machine learning and data mining ranging from survival analysis, and generalized linear models to approximate dynamic programming, and deep learning have been considered in designing the proposed approaches. In order to facilitate studying the related work, we categorize it based on the main components of the dissertation. Therefore, we study user engagement measures used in the literature in §2.1. The popularity and dwell time prediction studies are presented in §2.2. We discuss user acquisition and subscription prediction models related work in §2.3. Finally, we outline the research literature relevant to paywall problem in §2.4.

2.1 Engagement Measurements and Models

One important component in the proposed model is user engagement. Given the clickstreams collection as a low-level user-system interactions repository, user engagement modeling can be seen as a high-level feature extraction and quantification, which facilitates user behavior modeling. In web analytics, two types of user engagement measures are usually used in the literature: within a session (i.e., intra-session) and across sessions (i.e., inter-session). While inter-session measures take into account multiple sessions and as a result long-term user behaviors, intra-session measures only consider one session,

Table 2.1: Web-analytics user engagement measures.

Types	Measures and Descriptions
Intra-session	• **Dwell Time:** Contiguous time that users spend on a page (or a website) [5, 35, 83]. • **Revisit to site:** The number of times that user leaves a web site of interest and then returns to it (within a session) [43] • **Click through rate:** The number of time that an advertisement is clicked divided by the number of time it was shown [65]. • **Number of page view:** Number of pages visited within a site during a session or the percentage of session where a given number of pages were visited within a site [30].
Inter-session	• **Direct value:** The user value perceived by the web site operator (e.g., customer life time, ads clicked) [40] • **Total use:** Calculated based on observing the user behavior over long periods of time and across multiple sessions [24]. • **Return rate:** The fraction of users return to the site after the last visit or the time between two consecutive visits of a user [22].

so they are based on observations in a short period of time. Table 2.1 shows the summary of the most important engagement measures used in the literature.

There are several intra-session measures commonly used as the engagement. Among them *dwell time* is often used as a proxy of post-click user satisfaction [5]. Recently, this measure is used as an indirect indicator of users interests in items for the recommendation purpose [83]. This measure is one of the most important ones in digital newspaper domain, and we will study it in the subsequent chapters in details. Another measure is

the *revisit to site* defined as the number of times that a user leaves a web site of interest and then returns to it (within a session). While goal-oriented sites (e.g., shopping) have a lower within session revisit rate, sites belong to social media category have the highest number of revisit rates. The Click-Through Rate (CTR) is widely used for advertisement optimization and defined as the number of times that an advertisement is clicked divided by the number of times which it is shown (i.e., impressions). In fact, CTR can serve as an indirect proxy of conversion (e.g., purchase or subscription). This measure has been extensively used in the context of search engine optimization. The *Number of page viewed* is defined as the number of pages visited within a site during a session, or the percentage of sessions where a given number of pages are visited. This measure is appropriate for the sites that provide the content for consumption such as news portal sites. It could be a complementary measure to "revisit to site" as it only considers the number of on-site visits. For news portal sites, the number of articles read by a user might be the better measure as a poorly designed web site might result in navigating many pages. Intra-session measures can be easily misleading (one example is the case that a search engine provides irrelevant results and the user keeps looking for the relevant ones) [36].

The inter-session measures are the other class of user engagement measurements. The *direct value measurement* is calculated based on the user value perceived by the web site operator. The common measures are: total customer life time, the number of ads clicked, etc. The *total use* measurement is calculated based on users behaviors

observations over a long period of times across multiple sessions. The total usage time and the number of sessions per unit of time are examples of this category. Alternatively, measures like the total number of friends on a social media site could be an indicator of engagement [24]. Digital newspapers whose the major goal is to engage users by visiting site on regular basis can directly measure the success of the goal based on the fraction of users returning to the site after the last visit. Alternatively, they can calculate the time between two consecutive visits of a user (e.g., absence time) as the measure of engagement. These measures are refereed as the *return rate* measure [22].

User engagement depends on the domain of interest, attributes of users (i.e., visitors) and the targeted task [40]. For example, a news portal might have different engagement patterns in comparison to an on-line shopping site. For example in [84], it is shown that the average dwell time for 50 major Yahoo sites varies significantly. The lowest average dwell time belonged to sports sites while the highest average dwell time was recorded for leisure sites. Moreover, search sites have much shorter dwell time in comparison with entertainment sites.

Similarly, Lehmann et al. [44] studied 80 websites ranging from media to shopping sites. They clustered the users based on the number of days that they used a site and showed that different sites resembled different users types portions (e.g., site about special events like Oscar has the highest portion of users who visit the site only one day a month). Moreover, users loyalty is different among the sites (e.g., users of news por-

tal are more loyal than temporarily interesting sites like sites for buying a car). In fact, different users exhibit different types of engagement patterns.

The engagement pattern also depends on the task. For example, goal-specific tasks like checking e-mail have different engagement patterns from net surfing for leisure. That might be a reason for different engagement patterns in different times of a day [22].

2.2 Popularity and Dwell Time Prediction in Digital Media

Popularity has been widely studied in different domains using different metrics. For example, in Twitter, and Facebook popularity usually refers to the final number of re-shares [85] or likes [71] of a given post. However, in the news domain, most of the time popularity refers to the number of clicks/views which an article receives. There is a large body of research investigating and predicting the articles polarity. For example, Stratis et al. [29] considered the problem of predicting news traffics from different sources such as social media and search engines. They formulated the problem as a parameterized multidimensional time series fitting and utilized ADMM (Alternating Directions Method of Multipliers) to solve it. Kim et al. [32] showed that each individual word in a headline had its own Click-Through Rate (CTR), then, proposed a model to find the weight of each word. They developed a probabilistic generative model as an extension of Latent Dirichlet Allocation (LDA) [6] to generate the most engaging headlines for the news articles based on the historical logs of users clicks. Guo et al. [28] considered the problem

of CTR prediction and suggested a deep architecture for this purpose. The proposed model could capture the complex interactions between the features which are crucial in the CTR prediction task. Lehmann et al. [42] studied the effect of story-focused reading (i.e., a phenomenon of reading several articles related to a particular news development) on user engagement in the news domain. They showed that this phenomenon existed and promotion of it (e.g., through embedding some links in the news contents) could increase news consumptions in digital newspapers.

In the online news domain, dwell time is among the most important criteria for the assessment of user engagement and satisfaction. However, most of studies focus on utilizing dwell time in different applications (e.g., recommendation [83]), rather than studying effects of textual constituents on the dwell time. Chao et al. [45] modeled the distribution of dwell time on Web pages, and extracted different user browsing behaviors based on different Web page features. However, they did not focus on the textual content of the article which is the main reason of the engagement in the news domain and focus of our study. Dwell time has been used in many studies to quantify user satisfaction. For example, Yi et al. [83] argued that "click" was not reliable implicit feedback in recommendation systems. As such, they proposed dwell time as a user satisfaction measure to build the recommender model. They showed that dwell time was a batter inductor for user sanctification in the recommendation domain. Kim et al. [35] utilized the distribution of dwell time on web pages along with the other features of pages to predict the level

of user satisfactions on the search results.

2.3 User Acquisition in Digital News Media

User acquisition is traditionally studied under area of Customer Relation Management (CRM) [59]. However, most of efforts have focused on *user attraction*, *retention* and *churn management* rather than user acquisition. Moreover, despite the importance of acquisition task, there is no comprehensive study on using data mining/machine learning techniques for this task in digital news domain. To date, most of researches have been focusing on the other domains (e.g., e-commerce) where the main revenue is based on individual purchases rather than subscriptions.

Ng et al. [58] performed one of the first attempts on using data mining techniques for user retention in an imaginative telecommunication domain (the real domain was unanimous due to privacy issues). They divided the user retention problem into several sub-problems/sub-goals. First, they identified objective indicators, which was actually a feature selection procedure (a list of features as predictors was selected from many possible ones). They used the decision tree induction method (using a wrapper model of feature selection) for this purpose. Then, concept drift was addressed simply by frequently updating these predictors. They used deviation analysis (i.e., identifying the significant deviation from the expected value of predictors) for detecting potential defectors. Moreover, they used association rule mining to detect the customers who are likely

to follow the previously identified defecting customers. However, the proposed method was not applied to any real dataset due to privacy issues (the concepts and the proposed methods were explained using UC Irvine machine learning data repository [21]).

In [3], authors proposed a rule-based evolutionary algorithm and applied it to predict churns in the telecommunication domain. They argued that interpretability was important in this problem, and suggested that the rule-based method could address this issue by uncovering interpretable churn patterns. Their method encoded a complete set of rules in a single *chromosome* and defined the *population* as a collection of such chromosomes. The *fitness* function was defined based on a performance measure (i.e., the probability that the value of an attribute of tuple can be correctly predicted by other attributes values based on the rules in the current population). At the beginning, they generated a set of first-order rules (only have one attribute in the antecedent) by using a dependency analysis approach and initialized the population accordingly. Then, the evolutionary process (i.e., reproducing a new population based on genetic operators and the fitness function) produced a better population iteratively. The process continued by assigning the population with the new higher order rules by randomly combining the low order rules antecedents (e.g., second-order rules from first-order ones) and repeating the evolutionary process. They used features such as location, payment method, service plan in order to predict the users who are likely to churn.

In another study, Mozer et al. [54] considered the problem of churn prediction for

a major carrier company. They utilized features (overall 134 variables) such as call details records (e.g., data, time, duration), billing (e.g., monthly fee, additional charges), application for service (e.g., contact details, handset type), market (e.g., rate plan) and demographics to predict the churn. Three predictive models including decision tree (i.e., C4.5) logit regression, and non linear neural network with one single hidden layer were exploited and compared for this task. Based on the predicted churn probability of a subscriber, they decided whether to offer the subscriber some incentives or not (to persuade her/him to stay with the carrier). The basic idea was to provide an offer to any subscriber whose churn probability was above a certain threshold. The threshold was selected to maximize the expected cost saving of the carrier. They analyzed the profit of plan based on parameters such as the cost of incentive, the cost of churn and the retention rate (i.e., the portion of customers staying after receiving incentives).

Kim et al. [33] estimated the attractiveness (i.e., click values) of individual words in news articles headlines. Assuming some words significantly induce more clicks than others, they proposed a generative model which jointly modeled headlines, contents of news articles as well as the clickstream data. The model was an extension to Latent Dirichlet Allocation (LDA) [6] whereas the topic-specific click values of each word and the clicked words were modeled using beta and binomial distributions respectively.

Customer Life Time Value (LTV) analysis is another related area to user acquisition. Customer LTV is usually defined as the total net income that a company expects from its

customers. For example, Rosset et al. [67] calculated the current customer LTV based on three factors: customer value over time, length of service, and discounting factor. However, they estimated the effects of retention campaigns on Lifetime Value and did not investigate how a visitor (e.g., non-subscribed) becomes a customer (subscribed).

In this thesis, we consider the problem of user acquisition in the digital news domain. To best of our knowledge, this is the first work that considers this problem in the digital newspapers, and provides an end-to-end decision support solution for it. In particular, our proposed predictive model considers subscription time as a main component in both learning and inference process which has not been addressed before.

2.4 Paywall Mechanism in Digital Media

Finding revenue sources for journalism has been a major issue for the news industry over the past decade [56]. While diminishing income due to decline in advertisement revenue makes newspapers to start implementing paywall mechanisms, there are few studies on the analytical side to make these mechanisms more effective. Most studies in the journalism community focus on the qualitative and quantitative investigation of features (e.g., age) influencing people to subscribe to digital newspapers [14, 19, 25, 27]. For example, Fletcher et. al [25] utilized survey data of six countries to study the factors deriving users attitudes to pay for online newspaper subscriptions. They investigated different hypotheses and concluded that younger people and those who had paid for

print newspapers were more prone to pay for the online news subscription. Although the paywall mechanism can serve as a mean for optimizing business objectives of online newspapers, to best of our knowledge, this problem has not been studied.

The paywall problem can be framed as a sequential decision process (e.g., whether show the paywall for an article request or not). The sequential decision making over time has been studied in many disciplines with different names such as: *reinforcement learning* in machine learning, and *approximate dynamic programming* in operation research. In reinforcement learning, Markov Decision Process (MDP) is widely used to model the dynamics of an environment under different actions (i.e., decisions). When the environment model is available, dynamic programming methods such as *value iteration* or *policy iteration* can be used to find the optimal policy [74]. For example, Cai et. al [9] considered the biding problem where the main goal of the advertiser is to bid for every ad impression in an auction. Given a bid request, in each timestamp the advertiser should make a decision based on the ad request contexts and its current state (e.g., amount of budget). They designed a method based on Markov Decision Process (MDP) to learn the optimal decision policy. However, the environment model is often not available. In such cases, Temporal Difference learning [73] techniques such as Q-learning [78] or SARSA [74] can be utilized to learn the policy from the environment. However, model free approaches need a lot of interactions (i.e., exploration) with the environment before the convergence and suffer from transition dynamics of an enormous state space and the

sparsity of reward signals in the highly stochastic environment. Shani et. al [69] considered the recommendation (in the bookstore domain) as a decision problem compared to the traditional prediction perspective. They designed a framework based on MDP and utilized the value iteration to make an optimal decision (i.e., whether to recommend an item to a user or not). However, due to the limitation on exploration they finally used some heuristic techniques to learn the policy.

Approximate dynamic programming studies the sequential decision problem in a more general setting and with broader policy classes such as *myopic policy* (i.e., decision is based on the current state without considering the potential future states) or *lookahead policies* (i.e., decision is based on the predicted decision in the future) [61]. One related problem in this area is American price optioning [7], which has been studied extensively. American option allows option holders to exercise the option before the maturity date. This problem has been studied in the the stock dynamics in the risk neutral world (i.e., a stochastic differential equation). Despite some similarities between this problem and the paywall problem, in the paywall problem we do not have such a dynamic instead we need to build the solution in a data-driven fashion.

In the proposed paywall model, we utilize the *navigational graph* (based on users article reading behaviors) to build the paywall decision model. The navigational graph (or co-visit graph) provides important insights into users behaviors. While to best of our knowledge, this graph has not been used in the paywall problem, it has been widely

exploited to model users interests and preferences in different contexts. For example, Baluja et. al [4] considered the navigational graph as the basis to build a recommender framework for YouTube videos. They proposed a recommendation approach based on the random walk on the navigation graph. In another research, Xiang et. al [81] utilized the co-visit information (encoded in a navigational graph, which was called Session-based Temporal Graph) to capture the long-term and the short-term users preferences for the recommendation purpose.

2.5 Summary

In this chapter, we reviewed the relevant work on user engagement and acquisition problems in the news domain. In particular, we categorized the studies based on the problems and models proposed in the subsequent chapters. First, we introduced the work related to user engagement, then we provided an overview on popularity and dwell time. Next, we presented some related studies on user acquisition. Finally, we considered some literature relevant to the paywall problem.

3 Dwell Time Engagement Data Exploratory Model

3.1 Introduction

User engagement is one of the important components in the proposed framework. It has a close relationship with user acquisition and plays an important role in user satisfaction. Consider the scenario in which we want to predict online newspaper users who are prone to conversion (i.e., subscription) based on the historical data stored in the clickstream dataset. One reasonable assumption is that a user's decision on subscription is based on long-term and short-term positive experiences rather than an instantaneous thought. This is exactly related to the area of "user engagement" modeling. In other words, user engagement can be used intuitively as the predictor for many tasks such as user acquisition, retention, or even churn management. Moreover, understanding of user engagement allows any digital newspaper to make smarter decisions in many contexts ranging from article promotion (e.g., submitting engaging articles to social media) to revenue management (organizing advertisements on the article pages based on their engagement).

However, understanding and quantifying user engagement have their own challenges

and obstacles. In this chapter, we first introduce user engagement modeling generally, and then focus on the dwell time as one of the most important engagement measures in the news domain. We design a novel data exploratory analysis framework based on the main elements of news articles stories (i.e., topics, emotions, and events), and study articles dwell time under the effects of these elements accordingly. The summary of contributions of this chapter is as follows:

- We propose and design a novel data exploratory analysis framework to explore different dwell time reading pattern behaviors.
- We study the relationships between different aspects (i.e., topics, events, emotions) of articles and dwell time distribution patterns by combining text mining techniques with the reliability analysis.
- We conduct the experiments on The Globe and Mail dataset to show how the proposed approach can reveal interesting insights into the time that users spend on articles.

3.2 What is user engagement?

The user engagement concept has been known and studied over the past decade. In fact, measuring, understanding, and optimizing user engagement are major targets for many e-commerce companies and in particular digital newspapers. However, lack of common

definition among practitioners is a main obstacle in investigating user engagement. One relatively well-known definition is based on "positive aspects" of the experience of users, who are interacting with the on-line application [40]. The positive aspect emphasizes on the fact that the better experience does not always result in the higher level of engagement and vice versa (e.g., accessing Tweeter more frequently in comparison to Facebook does not show essentially better user experience due to difference in their engagement patterns).

User engagement depends on many complex factors, which makes its quantifying and understanding difficult. The common approaches to user engagement measurement can be divided into three broad categories: *self-report*, *physiological* , and *web analytics methods* [40]. While self-reporting methods use questionnaires, surveys, and interviews to measure user engagement, physiological approaches utilize observational methods such as facial expression, and speech analysis to quantify actual user engagement. On the other hand, web analytics measurement approaches are based on the trace of users while interacting with the Web site. While the methods based on self-report and physiological approaches are limited to a small number of users assumed to be the representatives of the whole population, Web analytics engagement measures are based on the whole population of users. Moreover, the web analytics approaches do not need explicit user feedbacks, which are not easy to collect. Due to these facts, web analytics methods are commonly used to measure engagement in the news domain even though they are

only an indirect proxy of real engagement.

3.3 Dwell time engagement in digital news media

Among a variety of measures proposed in web analytics for engagement quantification (see §2.1), dwell time (i.e., the time a user spent on an article) is one of the most important ones. Bellow are examples of dwell time applications in the news domain:

Content promotion: Posting articles to social media is a common practice for digital newspapers to attract the audience attention. This is crucial to any digital newspaper as it can increase the penetration rates and impacts of articles. For example, two-third (67%) of Americans obtain some of their news on social media [70]. Therefore, submitting the articles on which users spend a significant amount of time maximizes the likelihood of user attention. This results in boost in user acquisition and increase in advertisement revenue.

Content management: One challenging problem for digital newspapers is to organize the content on an article page before delivering it to users. For example, one important question is that how many advertisements should we place in an article page? The answer to this question has important implications on both user experience and revenue improvement (given the fact that online newspapers have limited space for the advertisement purpose). Most of the time, the answer to this question (or similar ones) depends on the time that users spend on the articles (if a user spends more time on an article, it is

more likely to notice the advertisements or click on them).

Recommendation system: It is well-known that dwell time can serve as a user preference indicator in news recommendation systems [83]. In fact, this measure performs much better than "click" as the implicit feedback. Thus, it is most desirable to utilize article dwell time as an important factor in combination with other factors (e.g., level of difficulty) to model the level of satisfaction [35].

Paywall decision model: The adaptive paywall model proposed later in the thesis works based on the notions of utility and cost. These are factors which can serve as the criteria in making a paywall decision (i.e., whether to show an article or the paywall). As we will see, it is possible to use dwell time as the utility of article in the proposed paywall model.

3.4 Content-based Article Dwell Time Distribution Analysis

Figure 3.1 shows the general overview of the proposed dwell time data exploratory framework. The framework provides insights on relationships between the news article dwell time patterns and article *topics*, *emotions*, and *events* as the main constituents of the article story.

More precisely, assume that $\mathcal{D} = \{a_i\}_{i=1}^{N}$ is a set of articles, and $\mathcal{A}_i = \{t_j\}_{j=1}^{N_i}$, is a set of dwell times of article $a_i \in \mathcal{D}$ (based on different users visits), and N_i is the number of visits which article a_i has . The main goal is to investigate the relationship between

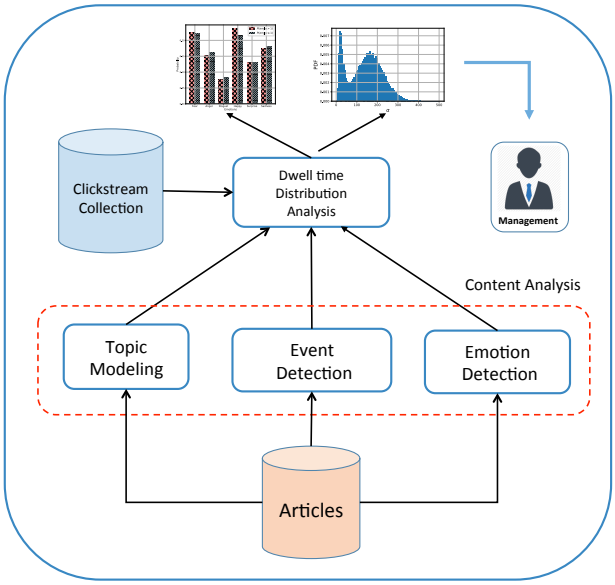


Figure 3.1: Framework for exploratory article dwell time analysis.

different article constituents (i.e., topics, emotions, and events) and different dwell time distribution patterns (see definition 3.4.1 and 3.4.2). The following subsections explain each of the proposed framework components in detail.

3.4.1 Article Dwell Time Distribution Fitting

The first question is which distribution is more appropriate for modeling article dwell time distributions. In order to answer this question, we fit the dwell times of all articles into different distributions and calculate the average log likelihood among all the articles as the fitness scores. Table 3.1 compares the negative log likelihood of Normal, Exponential and Weibull distributions for The Globe and Mail dataset (see § 1.4). As can be

Table 3.1: Negative log likelihood of different distributions for articles dwell time.

Score	**Normal**	**Exponential**	**Weibull**
Negative Log Likelihood	5100.57	4447.62	**4306.89**

seen, Weibull distribution is more appropriate for modeling dwell time of articles in our domain. Moreover, as we will see, Weibull distribution has a nice property in hazard and reliability analysis making it appropriate for our analysis. As such, we consider it as the dwell time distribution of the articles. The probability density function of Weibull distribution for the dwell times of article a_i can be written as follows:

$$f_{T_i}(t;\gamma_i,\alpha_i) = \frac{\gamma_i}{\alpha_i}(\frac{t}{\alpha_i})^{\gamma_i-1}\exp\{-(\frac{t}{\alpha_i})^{\gamma_i}\} \tag{3.1}$$

where α_i and γ_i are scale and shape parameters respectively, and T_i is the dwell time random variable for article a_i. The scale parameter α_i affects the stretch/ shrinkage of the distribution (e.g., the larger scale parameter spreads out the distribution more). The shape parameter γ_i affects the distribution form and has an important role in Weibull distribution analysis. Given the dwell time observation for article a_i (i.e., $\mathcal{A}_i = \{t_j\}_{j=1}^{N_i}$), parameters α_i and γ_i can be learned through the maximization of the following likelihood function:

$$LL(A_i;\alpha_i,\gamma_i) = \prod_{j=1}^{N_i}\frac{\gamma_i}{\alpha_i}(\frac{t_j}{\alpha_i})^{\gamma_i-1}\exp\{-(\frac{t_j}{\alpha_i})^{\gamma_i}\} \tag{3.2}$$

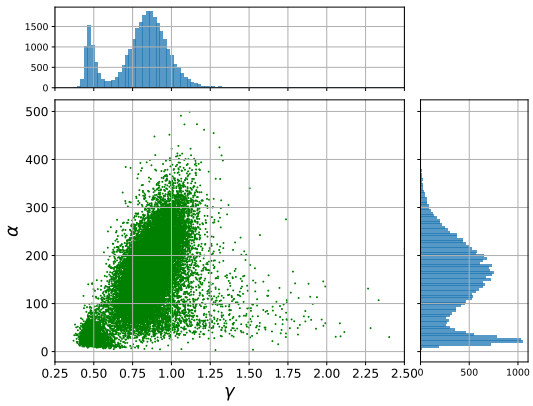


Figure 3.2: α_i and γ_i scatter plot and their respective histograms.

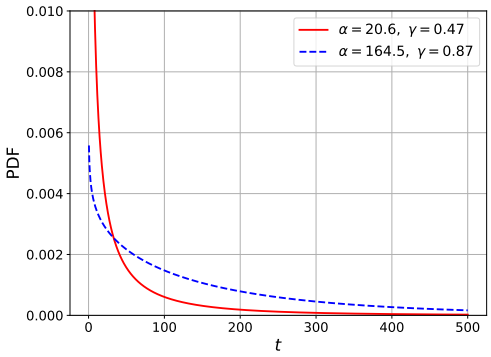


Figure 3.3: Two dominant distributions based on histograms of learned parameters.

where N_i is the number visits which article a_i has. However, there is no closed form formula for the parameters estimation. Therefore, we need an iterative approach to estimate the parameters [16].

Figure 3.2 illustrates the distribution of α_i and γ_i of The Globe and Mail articles dataset. As can be seen, there is two peaks for both γ (at 0.47, and 0.87) and α (at 20.6 and 164.5) distributions along the horizontal and vertical axes respectively. Figure 3.3 shows two dominant distributions of dwell time with respect to these peaks.

One important characteristic function in Weibull distribution analysis is *hazard rate*[5],

[5]In survival analysis, the hazard rate gives the rate of immediate failure of a survivor at time t, and shows the amount of risk which is associated with the survivor.

which is defined as:

$$h(t) = \lim_{\Delta \to 0} \frac{P(t \leq T < t + \Delta | T \geq t)}{\Delta} \tag{3.3}$$

where T is the dwell time random variable. The numerator is the conditional probability that the dwell time is in the interval $[t, t + \Delta)$ given that it is greater or equal to t, and the denominator is the width of the interval. Taking the limit as the width of interval goes to zero results in the instantaneous rate of leaving the article at time t (since the start of reading). It can be shown that [51]:

$$h(t) = \frac{f(t)}{1 - F(t)} \tag{3.4}$$

where $f(t)$ is the Probability Density Function (PDF) and $F(t)$ is the Cumulative Distribution Function (CDF) at time t. In case of the Weibull distribution, the hazard function can be written as follows:

$$h(t) = \frac{\gamma}{\alpha^{\gamma}} t^{\gamma - 1} \tag{3.5}$$

One important characteristic of hazard function is whether it is monotonically increasing or decreasing. The first order derivative of Equation 3.5 is:

$$h'(t) = \frac{\gamma (\gamma - 1)}{\alpha^2} (\frac{t}{\alpha})^{\gamma - 2} \tag{3.6}$$

As can be seen, when $\gamma \in (0, 1)$, Equation (3.6) is less than 0, so the hazard rate is monotonically decreasing, and when $\gamma > 1$, Equation (3.6) is greater than 0, and as

a result, the hazard rate is monotonically increasing. This makes the interpretation of hazard function for Weibull distribution straightforward. Note that when $\gamma = 1$, the hazard rate is constant over time.

Definition 3.4.1 (Positive aging reading pattern) *For the article with $\gamma > 1$, the rate of instantly leaving the article increases over time. This pattern is defined as the positive aging reading pattern.*

Definition 3.4.2 (Negative aging reading pattern) *For the article with $\gamma < 1$, the rate of instantly leaving the article decreases over time. This pattern is defined as the negative aging reading pattern.*

In another word, $\gamma > 1$ means that the longer a user read the article. the more likely she leaves instantly while $\gamma < 1$ indicates that leaving rate is higher at the beginning than the later stages of reading. In the subsequent sections, we study these two reading behaviors with respect to article topics, emotions and events.

3.4.2 Content-based Article Dwell Time Distributions Analysis

Figures 3.2 and 3.3 provide some insights on how people spend time on articles and reveal some reading pattern behaviors. However, it does not relate these behaviors with the articles contents. In this section, we design effective approaches which relate article contents (e.g., topics) to articles dwell time patterns.

3.4.2.1 Topic-based Article Dwell Time Distribution Analysis

Topic modeling based on Latent Dirichlet Allocation (LDA) [6] has been widely used in many text mining tasks. The basic idea behind topic models is that documents are mixtures of topics and each topic is a distribution over words. Given an article set $\mathcal{D} = \{a_i\}_{i=1}^{N}$ and vocabulary $\mathcal{V}$, a topic tp_k (where $k = 1 \ldots K$) extracted by LDA is a multinomial distribution over words. This distribution encodes the probability of word $w_j \in \mathcal{V}$ in topic tp_k as $P(w_j|tp_k)$. Usually top probable words are considered as the representative keywords for topic tp_k. Moreover, LDA assigns each article $a_i \in \mathcal{D}$ to a topic distribution $P(tp_k|a_i)$, which specifies the fraction of words in a_i discussing a particular topic tp_k. The parameters of both distributions are assumed to be drawn from Dirichlet distributions and estimated by Gibbs sampling [72] or variational methods [6].

Given topic tp_k of article a_i, we are interested in finding the relationship between the dwell time distribution characteristic of articles and their respective topics. In particular, our aim is to estimate the the occurrence of different topics in articles with monotonically increasing ($\gamma > 1$) and decreasing ($\gamma < 1$) hazard rate. Given the topic distribution of articles, $P(tp_k|\gamma > 1)$ can be estimated as follows:

$$P(tp_k|\gamma > 1) = \sum_i P(tp_k, a_i|\gamma > 1) \tag{3.7}$$

$$= \sum_i P(a_i|\gamma > 1)P(tp_k|a_i, \gamma > 1) \tag{3.8}$$

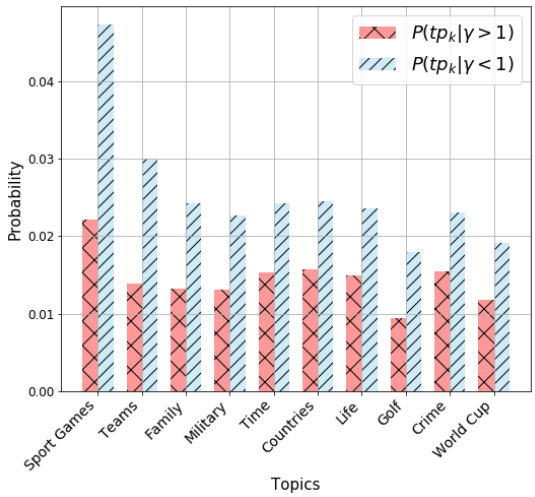


Figure 3.4: Topics presented mostly in articles with $\gamma < 1$

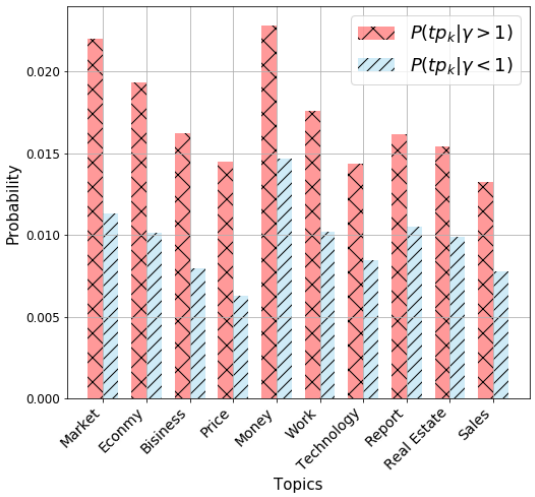


Figure 3.5: Topics presented mostly in articles with $\gamma > 1$

Assuming topics of article depends only on the article content a_i:

$$P(tp_k|\gamma > 1) = \sum_i P(a_i|\gamma > 1)P(tp_k|a_i). \tag{3.9}$$

$P(a_i|\gamma > 1)$ is estimated as:

$$P(a_i|\gamma > 1) = \frac{I_{\gamma>1}[a_i]}{\sum_j I_{\gamma>1}[a_j]} \tag{3.10}$$

where $I_{\gamma>1}$ is the indicator function for the articles with $\gamma > 1$.

Similarly, $P(tp_k|\gamma < 1)$ can be estimated:

$$P(tp_k|\gamma < 1) = \sum_i P(a_i|\gamma < 1)P(tp_k|a_i) \tag{3.11}$$

Figure 3.4 shows the top 10 topics[6] with the largest difference between $P(tp_k|\gamma < 1)$

[6]For each topic, we manually choose the best representative word in the distribution of topic words as

Figure 3.6: Word clouds representation of topics with respect to their dwell times.

and $P(tp_k|\gamma > 1)$. The figure highlights the difference between the topic distributions of articles with negative and positive aging reading behaviors. As can be seen, topics related to sport (i.e., *Sport Game*, *Teams*, *World cup* and *Golf*) appear more in articles with $\gamma < 1$. The topics such as *Family*, *Military*, and *Crime* are the other instances which their presence are significantly more in articles with $\gamma < 1$. This suggests the leaving rate of articles containing these topics is higher at the early stages of reading and decreases over time. In another word, users screening these kinds of topics at the beginning of the visit and many of them leaving the article vary soon.

Figure 3.4 illustrates the top 10 topics with the largest difference between $P(tp_k|\gamma > 1)$ and $P(tp_k|\gamma < 1)$. The topics related to *Market* and *Economy* are more likely to occur in the articles with $\gamma > 1$. As can be seen, most of these topics are about serious subjects such as *Business*, *Price*, and *Work*. In fact, for these topics the rate of leaving article at

---

the label. However, it is possible to use automatic topic labeling approaches [17].

the beginning is low and increases over time.

Moreover, $P(t|tp_k)$ can be estimated as follows.

$$P(t|tp_k) = \sum_i P(t, a_i|tp_k) \tag{3.12}$$

$$= \sum_i \frac{P(t, a_i, tp_k)}{P(tp_k)} \tag{3.13}$$

$$= \sum_i \frac{P(a_i)P(tp_k|a_i)P(t|a_i, tp_k)}{P(tp_k)} \tag{3.14}$$

Assuming that article dwell time is independent of the article topics given its content:

$$P(t|tp_k) = \sum_i \frac{P(a_i)P(tp_k|a_i)P(t|a_i)}{P(tp_k)} \tag{3.15}$$

$$= \frac{\sum_i P(a_i)P(tp_k|a_i)P(t|a_i)}{\sum_j P(a_j)P(tp_k|a_j)} \tag{3.16}$$

Given Equation (3.16), the expected value of dwell time for topic tp_k is:

$$\mathbf{E}[t|tp_k] = \frac{\sum_i P(a_i)P(tp_k|a_i)\mathbf{E}[t|a_i]}{\sum_j P(a_j)P(tp_k|a_j)} \tag{3.17}$$

Figure 3.6 illustrates the word clouds visualization of different topics according to their expected dwell time. The size of topics labels are proportional to the expected dwell time of the topics. As can be seen, topics related to political matters such as *Russia*, *Policy*, *Parties*, and *Government* have more expected dwell time in comparison to the other topics.

3.4.2.2 Emotion-based Article Dwell Time Distribution Analysis

In this section, we investigate the effect of different emotions involved in articles on the time which users spend on them. Emotion detection from text has been widely studied in different contexts [53]. However, it is not been investigated for the dwell time prediction task. We consider 6 basic emotions (i.e., *happiness*, *sadness*, *disgust*, *anger*, *surprise*, and *fear*) which have been widely used in the emotion detection literature [1, 23]. We utilize a publicly available emotion lexicon [52] as the seed words of different emotions. The basic idea is to label data from a pool of unlabeled data (e.g., articles) using existing labeled data (e.g., seed words of different emotions). Given an articles $a_i \in \mathcal{D}$, and the word $w \in a_i$, the emotion vector of word w can be defined as: $\boldsymbol{em}(w) = [emw_j]$, where emw_j is the average similarity (i.e., positive cosine similarity) between the pre-trained embedding vector word w [50] and the seed words belonging to the emotion j. The emotion vector for article a_i can be calculated as:

$$X^i_{EM} = \frac{1}{|\sum_{w \in a_i} \boldsymbol{em}(w)|_1} \sum_{w \in a_i} \boldsymbol{em}(w) \tag{3.18}$$

where the denominator is for the scaling purpose. Note that $X^i_{EM} = [P(em_j|a_i)]$ is a stochastic vector, where the $j'th$ component of the vector is an estimation of probability of emotion j in article a_i denoted as $P(em_j|a_i)$.

Similar to Equation (3.11), $P(em_j|\gamma > 1)$ and $P(em_j|\gamma < 1)$ can be estimated as

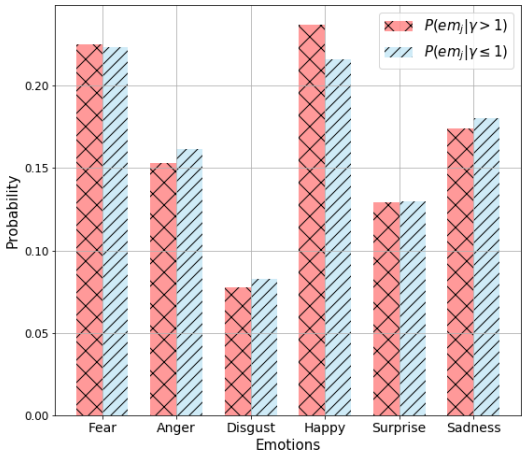


Figure 3.7: Presence of different emotions.

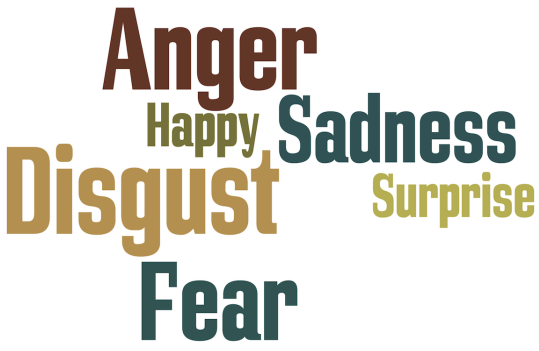


Figure 3.8: Word clouds representing of emotions with respect to their dwell times.

follows:

$$P(em_j|\gamma > 1) = \sum_i P(a_i|\gamma > 1)P(em_j|a_i) \tag{3.19}$$

$$P(em_j|\gamma < 1) = \sum_i P(a_i|\gamma < 1)P(em_j|a_i) \tag{3.20}$$

Figure 3.7 shows the probabilities of presence of different emotions in two categories of articles. We see that presence of *Happy* emotion is more in articles with positive aging reading patterns (i.e., $\gamma > 1$). This means that leaving rate is low in these articles at the beginning and increase over time. However, emotions such as *Anger*, *Disgust*, and *Sadness* appear in the article with the negative reading aging pattern (i.e., $\gamma < 1$), which suggests many users leave these articles at the beginning, but the leaving rate is decreasing over time. On the other hand, the probabilities of occurrence of *Fear* and *Surprise*

emotions in two types of articles are almost the same. Similar to Equation (3.16), we can estimate the expected dwell time for each emotion:

$$\mathbf{E}[t|em_j] = \frac{\sum_i P(a_i)P(em_j|a_i)\mathbf{E}[t|a_i]}{\sum_i P(a_i)P(em_j|a_i)} \tag{3.21}$$

Figure 3.8 illustrates the word clouds visualization of different emotions according to their dwell time. As can be seen, *Disgust*, *Anger*, *Sadness*, and *Fear* emotions, have the higher expected dwell time in comparison to *Happy* and *Surprise* emotions.

3.4.2.3 Event-based Article Dwell Time Distribution Analysis

News and events are closely related to each other. Most of the time, a news article reports one central event and a mixture of associated subsidiary events [10]. The central and subsidiary events manifest themselves in the article content through the event trigger words. Despite the importance of events in different news analytics applications [2], to best of our knowledge, no study has considered them in article dwell time analysis.

In this section, we study the relationship between article events and article dwell time patterns. We adapt the method proposed in [82] to extract the events at the article level. The basic idea is to identify the events and entities at the same time. Moreover, this method considers the relation between events in the document context. In particular, we are interested in extracting the entity mentions (i..e, noun phrase ot pronoun reference to an entity), event triggers (i.e., word or phrase expressing the event occurrence), and event arguments (i.e., entity having an specific role in the event).

For completeness, we briefly outline the approach. The model is composed of three sub-models: *within-event structure*, *event-event relation*, and *entity extraction*. More precisely, given an article a_i, for each event trigger candidate $w \in a_i$, a discrete variable tr_w is defined which takes the value from the set of event types (or None). Moreover, for each candidate argument $c_w \in \mathcal{P}_w$ (set of potential argument candidates for the trigger candidate w), we associate a discrete variables r_{wc_w} taking the value from the semantic roles set (or None), and a discrete variable en_{c_w} which defines the entity type of argument candidate c_w (or None). The within-event structure component models *event type*, *entity types* and their *semantic roles* in the event using a factor graph. The factor graph models the $p_\theta(tr_w, \{r_{wc_w}\}, \{en_{c_w}\}|w, \mathcal{P}_w, a_i)$. Similarly, the event-event relation component models the join probability distribution of entity types of two trigger candidates w and w', $P_\phi(tr_w, tr_{w'}|w, w', a_i)$, and the entity extraction models the probability of entity type of an argument candidate w'', $P_\psi(en_{w''}|w'', a_i)$. The parameters (i.e., θ, ϕ, and ψ) of these models are estimated accordingly using the L-BFGS method [46] . Finally, to assign the globally-optimum assignments of all variables (i.e., $tr_w, r_{w\ c_w}, en_{c_w}$), first, Conditional Random Field (CRF) models [38] are built on the training dataset to generate a set of event trigger $\{w\}$ and candidates argument $\{w''\}$ on the test set. Then, the AD^3 approach [49] is utilized to find the best assignments to the variables, which maximize the sum of confidence of all models [82].

The model is trained on the ACE 2005 corpus[7] [77]. It contains text documents from variety of sources such as newswire, broadcast conversation, and weblogs. The corpus defines 8 event types and 33 event subtypes. The event triggers and entity mentions are annotated in the document sentences accordingly. We follow the same setting as [82], and build the model on the ACE 2005 corpus, then, we extract the events in The Globe and Mail articles. We define the event vector for each word w in article a_i as follows: $ev(w) = [evw_j]$, where evw_j is 1 if tr_w is assigned to the $j'th$ event. The article level event vector X^i_{EV} for article a_i is defined as follows:

$$X^i_{EV} = \frac{1}{|\sum_{w \in a_i} \boldsymbol{ev}(w)|_1} \sum_{w \in a_i} \boldsymbol{ev}(w) \tag{3.22}$$

$X^i_{EV} = [P(ev_j|a_i)]$ is a stochastic vector, where the $j'th$ component of the vector is an estimation of probability of event j in article a_i denoted as $P(ev_j|a_i)$.

Similar to Equation (3.11), $P(ev_j|\gamma > 1)$ and $P(ev_j|\gamma < 1)$ can be estimated as follows:

$$P(ev_j|\gamma > 1) = \sum_i P(a_i|\gamma > 1)P(ev_j|a_i) \tag{3.23}$$

$$P(ev_j|\gamma < 1) = \sum_i P(a_i|\gamma < 1)P(ev_j|a_i) \tag{3.24}$$

where $j'th$ component of the vector is an estimation of $P(ev_j|a_i)$.

Figure 3.9 shows the presence of top 10 events (i.e., with highest probability of occurrence) extracted for The Globe and Mail articles. As can be seen, the events such

[7] https://catalog.ldc.upenn.edu/LDC2006T06

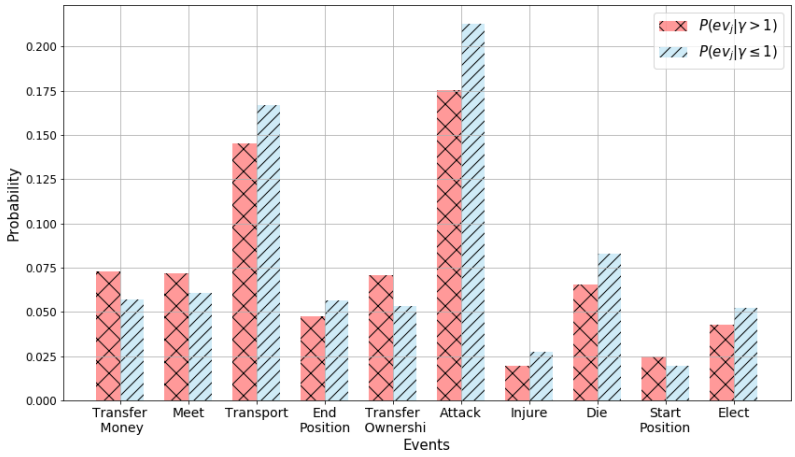


Figure 3.9: Presence of events in different articles.

as *Transfer Money*, *Meet*, and *Transfer Ownership* appear more in the articles with the positive aging reading pattern (i.e., $\gamma > 1$). However, events such as *Transport*, *Attack*, *Die* and *Elect* are occurred more in the article with the negative aging reading patterns (i.e., $\gamma < 1$). This means that the rate of leaving articles containing these events at the beginning is high for these events and decreases over time. In other words, these events are monitored harshly by users at the beginning.

Similar to Equation (3.16):

$$\mathbf{E}[t|ev_j] = \frac{\sum_i P(a_i)P(ev_j|a_i)\mathbf{E}[t|a_i]}{\sum_i P(a_i)P(ev_j|a_i)} \tag{3.25}$$

Figure 3.6 illustrates the word clouds visualization of different events according to their dwell time. As can be seen, events such as *Elect*, *Attack*, *Acquit*, and *Demonstrate* have higher expected dwell time engagement in comparison to others.

Figure 3.10: Word clouds representing of events with respect to their dwell times.

Note that we can compute the entity vector for word w in a similar fashion: $\boldsymbol{en_k}(w) = [enw_j]$, where enw_j is 1 if w is the $j'th$ instance of entity k (where k is a type of entity, i.e., person or organization), otherwise 0. For article a_i, the article level entity vector $X^i_{EN_k}$ is defined as:

$$X^i_{EN_k} = \frac{1}{|\sum_{w \in a_i} \boldsymbol{en_k}(w)|_1} \sum_{w \in a_i} \boldsymbol{en_k}(w) \tag{3.26}$$

We extract 31 events, 87083 people, and 79143 organizations from The Globe and Mail dataset. We use Equation (3.26) in subsequent sections for the prediction task.

3.5 Summary

In this chapter, we designed a data exploratory framework by combining the reliability analysis with text mining techniques to analyze news articles dwell time. We defined different reading patterns and investigated the presence of different article constituents

(i.e., topics, emotions, and events) on these patterns. The model outcomes revealed interesting patterns about article engagement and can be used in different tasks. For example, one interesting application is to use a mixture of articles with different reading patterns (i.e., positive and negative aging) in the recommendation task to maximize user engagement.

4 Dwell Time Engagement Prediction

4.1 Introduction

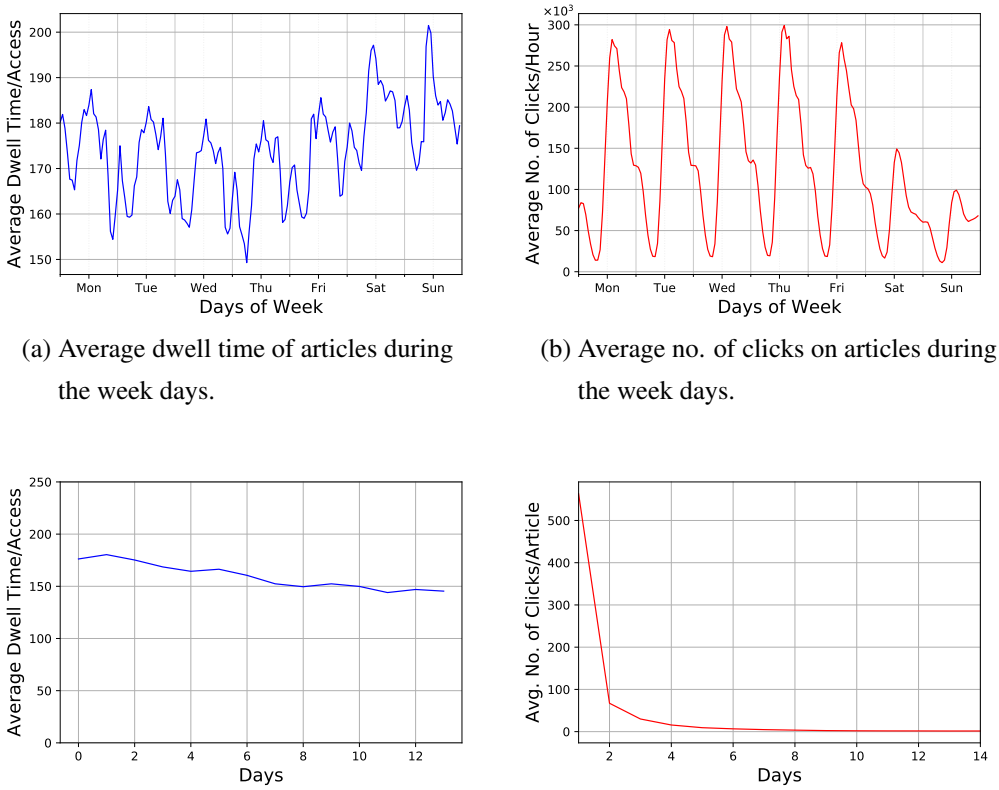
In the previous chapter, we explored articles dwell time engagement distribution under the scope of topics, emotions, and events involved in the news articles. However, it is most desirable to predict an article dwell time engagement[8] before publishing it as this allows editors to have an idea about its prosperity. This not only helps online newspapers in making plan to attract and persuade people to subscription in a long-term, but also leads them to make smarter decisions to increase revenue in a short-term (see §3.3). For example, if a user spends more time on reading an article, it is more likely she/he notices an advertisement on the article page. Therefore, we can place more advertisements on such a page to maximize the revenue.

Most of the previous studies focus on predicting the page views (i.e., user clicks) as the sole indicator of user engagement and article success. In fact, they try to predict

[8]We refer to the expected value of distribution of article dwell times as the article dwell time engagement and define it formally in §4.2.

the traffic which an article may receive in the future [29, 32]. However, click-based engagement modeling could be quite noisy (e.g., click on a wrong article) and may not show the actual user engagement nor satisfaction [83]. Alternatively, it is shown that the time that a user spends on a page, known as the *dwell time*, is one of the most significant indicators of engagement [15, 26, 35]. Nevertheless, there are few studies on predicting articles dwell times based on articles contents in the news domain. While there are some work on analyzing Web pages dwell times [45], most of them do not focus on the textual contents of the pages as the main factor of the page dwell time. However, the content is important since in the news domain the pages layouts are almost the same for all articles, and users are mostly engaged with the articles contents. Moreover, "dwell time" and "click" demonstrate completely different patterns in the news domain and patterns in one cannot be generalized to another one (see Figure 4.1). In this chapter, we consider dwell time as an engagement measure and design an effective model to predict the dwell time of an article based on its content.

The problem of predicting article dwell time is a challenging task. The main challenge is to extract and select the features having impact on the articles dwell time. In fact, finding the consistent indicators correlated with the article dwell time is not a trivial task as dwell time may be under influence of different factors. In order to predict an article's dwell time, we need to consider the interplay between the article content constituents (e.g., people involved, or events) and their roles in engaging readers in terms



(a) Average dwell time of articles during the week days.

(b) Average no. of clicks on articles during the week days.

(c) Average dwell time of an article since release.

(d) Average no. of click on an article since release.

Figure 4.1: Dwell time of articles and number of clicks to articles in The Globe and Mail dataset. Dwell time and number of clicks demonstrate different patterns.

of dwell time. Moreover, these factors may interact with each other to affect the article dwell time. For example, an article about two celebrities participating in one event may be more engaging (i.e., have high dwell time) in comparison to individual articles about each of them. The prediction model needs to consider these interactions carefully.

In this work, we propose a framework to address the aforementioned challenges. First, we define the expected value of an article dwell time distribution, which is inves-

tigated in §3.4.1, as the article engagement. Then, we consider *events*, *emotions* as well as *people* and *organizations* as the main contributors to the news article dwell time. In particular, we show that augmenting the article contents with the main factors of the story (i.e., events, people, organizations, emotions) improves the prediction task and leads to a method that outperforms the existing state-of-the-arts models. We propose a model based on the wide and deep neural network architecture [12] which *memorizes* the low order interactions between the augmented sparse features (e.g., people in articles), and at the same time *generalizes* the article contents through the deep component. The idea is while the deep component can effectively capture the main abstract features of text, these architectures with embeddings can over-generalize and produce less accurate prediction when the input dimension interactions are sparse and high-rank [12]. In order to learn the low-order interactions between the augmented features, we adopt the factorization machine [28], which extracts feature interactions automatically, as the wide component in the proposed article dwell time prediction model. In the proposed model, each input dimension can be represented by multiple vectors providing more flexible representation than traditional factorization machine. Our main contribution are as follows:

- We design an end-to-end framework which predicts articles dwell times based on their contents.
- We study the effect of different article constituents (e.g., events, emotions) on arti-

cles dwell times.

- We design an effective deep neural network model to predict an article dwell time using its content.
- We conduct the experiments on a real dataset from The Globe and Mail and show the effectiveness of the proposed model.

4.2 Content-based Correlation Analysis

In this section, we study how different factors of an article (i.e., entities, emotions, and events) impact the dwell time of the article.

Definition 4.2.1 (Article dwell time engagement) *The dwell time engagement of article a_i denoted by y_i is defined as the expected value of the distribution defined in Equation* (3.1) *as follows:*

$$y_i \triangleq \mathbf{E}[T_i] = \alpha_i \Gamma(1 + \frac{1}{\gamma_i}) \tag{4.1}$$

where Γ is the Gamma function. We utilize y_i as the target value and build a model to predict it.

Definition 4.2.2 (Factor level engagement) *The engagement score of factor c (i.e., entity mention, emotion, or event) is defined as:*

$$Score(c) = \frac{1}{df(c)} \sum_i \mathbb{I}[c \in a_i] \times y_i \tag{4.2}$$

where $\mathbb{I}$ is the indicator function and $df(c)$ is the number of articles containing c. The intuition is that if a factor c appears exclusively in some articles with high dwell time (i.e., y_i), it should have a high engagement score. For example, if *Barak Obama* appears in articles with high dwell time, it should receive a high engagement score.

To investigate the extend to which the engagement score of each article factor could explain the variability of the dwell time of articles, we do a Pearson's correlation analysis. In particular, we estimate the predicted dwell time of article a_i by averaging the engagement scores of all individual factors in the article a_i, and then calculate the Pearson's correlation coefficient between an article's actual dwell time and its predicted value. Figure 4.2, shows the correlation scores between the actual dwell times and the predicted ones for each factor. As illustrated, the emotions involved in the articles show the most correlation (weakly positive) with the article dwell times. Moreover, location (LOC) and time (TIME) have the least correlation to article dwell times. This observation motivates us to use emotion (EMO), event (EVENT), person (PER), and organization (ORG) as the *augmented features* in building the dwell time prediction model.

4.3 Deep Dwell-time Prediction Model

In the previous section, we studied the factors in articles contents affecting the articles dwell time distributions. In this section, we consider the problem of article dwell time prediction.

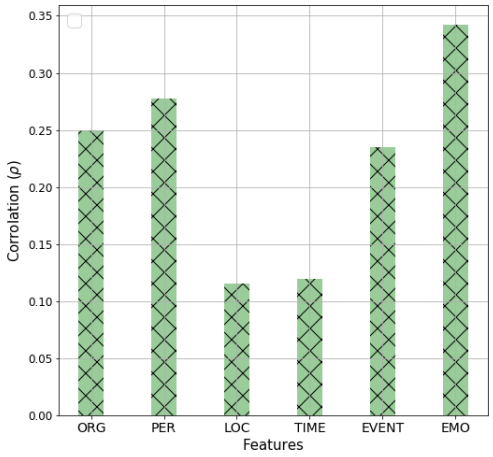


Figure 4.2: Pearson correlation coefficient between the true dwell time and predicted dwell time based on different factors.

PROBLEM STATEMENT (**Article dwell time prediction**): Given a set of article $\mathcal{D} = \{a_i\}_{i=1}^{N}$, the goal is to learn a model so that it can be used to predict the dwell time of a new article based on definition 4.2.1.

To learn a dwell time prediction model, we represent an article using both the words in the article and its augmented features (i.e., emotions, events, people and organizations). However, the event and entity features have a high dimensionality and sparse values. Thus, special attention should be paid to deal with such input features. *Deep neural networks* can learn feature representations and alleviate need for feature engineering by embedding sparse features into a low-dimensional dense space. However, the embedding space may be over-generalized and produce poor results in the prediction tasks, when the interactions between high-dimensional features are sparse [12]. How-

ever, such interactions are important for predicting dwell time. For example, an article about two celebrities attending the same event is more likely attracting more readers. Thus, we propose a deep neural network architecture which leverages the augmented features and their interactions in combination with the document (i.e., article) representation to predict the article dwell time.

4.3.1 Proposed Model Overview

We consider people, organizations, events and emotions in article contents as the *augmented features*. Moreover, we argue that extracting, and then augmenting an article content with these features highlights these factors (i.e., augmented features) in the article content and helps the article dwell time prediction task. Most of the time, these factors are sparse and have different levels of interactions. However, in the news domain, such interactions are hidden in data and are extremely hard to identify. Inspired by [28], we utilized the factorization machine [63] to capture the augmented feature interactions. However, the proposed model is different from [28] in the following aspects: (1) we augment the article content with emotion, event and entity features (2) our model allows multiple factorization machines (each feature is represented with multiple embedding vectors in the factorization layer). In particular, we use a wide component for the augmented features (deep and wide components have different inputs). Moreover, the proposed model produces a vector as the factorization layer output, which encodes and

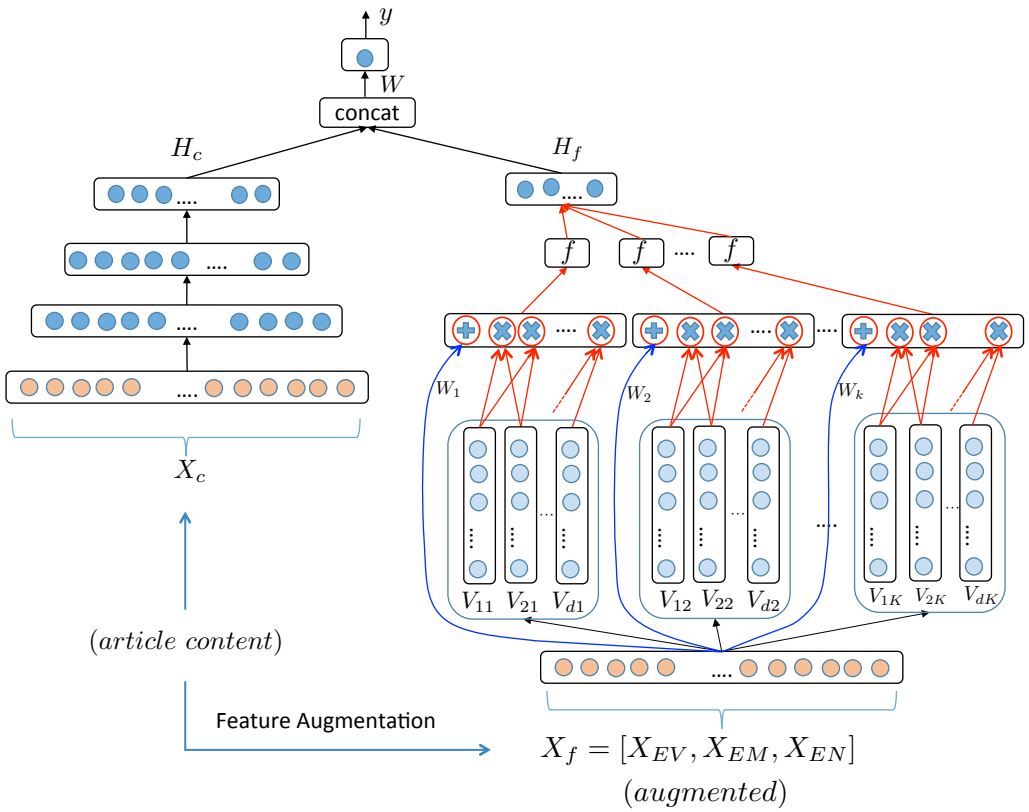


Figure 4.3: The architecture for article dwell time engagement prediction (the left side is the deep component and the right side is the factorization machine component).

represents the pairwise interactions between the augmented features.

4.3.2 The Architecture

Figure 4.3 shows the proposed architecture for the article dwell time engagement prediction task. The architecture consists of two main components: the *deep* and the *factorization machine* components. While the deep competent learns the high-order feature

interactions and generalizes the article content through a multilayer encoder, the factorization machine captures the low order interactions among the highly sparse augmented features. In particular, suppose that each article is represented by the *TFIDF* [68] vector X_c, which is fed into the deep component, and augmented vector $X_f = [X_{EV}; X_{EM}; X_{EN}]$, which goes to the factorization machine component, where X_{EV}, X_{EM}, and $X_{EN} = [X_{EN_{PER}}; X_{EN_{ORG}}]$ are event, emotion, and entity vectors respectively (see chapter 3 for their definitions). The whole model is specified by the following equation:

$$H = Concat(Hc, H_f) \tag{4.3}$$

$$y = g(WH + b) \tag{4.4}$$

where H_c, H_f are the latent vectors learned by deep and the factorization machine components, H is the concatenation of these two vectors, and W, and b are weight and bias parameters respectively. The g is the activation function which is Rectified Linear Unit (ReLU) in the proposed model.

4.3.2.1 Factorization Machine Component

A simple strategy to capture the interactions between features is to learn a weight for each combination of two features. However, this naive approach does not work when the input feature space is sparse. Factorization machine solves the problem by modeling the pairwise feature interactions as the inner product of low dimensional vectors.

The first layer in the factorization machine component is the embedding layer. Given the sparse (augmented) input vector $X_f = [x_i]_{d\times 1}$, it learns multiple vectors $V_{ik} = [v_{ikl}]_{M\times 1}$ for each input dimension i ($k = 1 \ldots K$), where V_{ik} is the $k'th$ vector for dimension i, v_{ikl} is the $l'th$ elements of V_{ik}, and M is the size of embedding vectors. Then, these factors are fed into the interaction layer to capture the first order and the second order interactions. The interaction layer operation along with the $k'th$ dimension can be formalized as follows:

$$hf_k = f(b_k + \underbrace{W_k \cdot X_f}_{\oplus} + \sum_{i=1}^{d}\sum_{j=i+1}^{d} \underbrace{V_{ik} \cdot V_{jk}\ x_i x_j}_{\otimes}) \tag{4.5}$$

where hf_k is the $k'th$ elements of factorization machine component output $H_f = [hf_k]_{K\times 1}$, $W_k = [w_{km}]_{d\times 1}$ (w_{km} is the $m'th$ element of W_k) and b_k are the parameter vector and the bias to be learned and f is the activation function. The $\oplus$ and $\otimes$ symbols in Figure 4.3 refer to the first order and the second order interaction operations respectively. In fact, factorization machine replaces the interaction weights between feature x_i and x_j with the inner product of respective embedding vectors (i.e., $V_{ik} \cdot V_{jk}$). From modeling perspective, this is powerful since each feature ends up in an embedding space where similar features in this space are close to each other.

The 3'rd term in the right hand side of Equation (4.5) can be written as follows:

$$\sum_{i=1}^{d}\sum_{j=i+1}^{d} V_{ik}.V_{jk}x_i x_j = \frac{1}{2}\sum_{l=1}^{M}((\sum_{i=1}^{d} v_{ikl}x_i)^2 - \sum_{i=1}^{d} v_{ikl}^2 x_i^2) \tag{4.6}$$

Therefore, Equation (4.5) can be computed in $O(M \times d)$ rather than $O(d^2)$ in the naive modeling solution (M is the size of embedding vectors).

4.3.2.2 Deep Component

In the proposed architecture, the deep component is a dense feed-forward neural network. Each article is vectorized using the *TFIDF* approach, and then is fed into this component. The feed-forward layer converts this sparse vectors into a low-dimensional dense real-valued vector (i.e., embedding vector). In particular, each layer performs the following operation:

$$a^{(p+1)} = f(\boldsymbol{W}_{\boldsymbol{deep}}^{(\boldsymbol{p})} a^{(p)} + b^{(p)}) \tag{4.7}$$

where p is the layer number, $a^{(p+1)}$, $\boldsymbol{W}_{\boldsymbol{deep}}^{(\boldsymbol{p})}$, and $b^{(p)}$ are the activation, model weights, and bias at layer p. The f is the activation function which is Rectified Linear Unit (ReLU) in the proposed model.

4.4 Empirical Evaluation

In this section, we evaluate the proposed architecture for the article dwell time prediction. All the experiments are conducted on The Globe and Mail dataset described in §1.4. Moreover, all the experiments in this section are based on the 10-fold cross validation. We set M and K to 100 and 10. Moreover, we set the number of layers to 3 in the deep

component. We use the code in [60] with default parameter setting for non-neural network models , and neural network models are implemented using Keras with tensorflow backend [13].

4.4.1 Baselines

We compare the proposed model with the following baselines:

Linear Regression (LR): This is a simple baseline used the topics and document vectors as the features and the linear regression method to predict articles dwell times. We extract the articles topics based on the LDA approach [6]. We set the number of topics to 70 based on the best coherence scores proposed in [66]. Moreover, We learn the vector representation of each article using the doc2vec method proposed in [41]. We set the vector size to 100 in all experiments.

Random Forest Regression (RF): Random Forest regression has been known to perform well in many industrial applications. The basic idea is to train an ensemble of uncorrelated weak learners (i.e., decision tree), then combine the average results. We used the topic and doc2vec vectors in combination with Random Forest regression to predict each article dwell time.

Word Embedding + CNN: We adapt the approach proposed in [34] for the dwell time prediction task. The architecture is comprised of one layer of convolution on top of word vectors pre-trained from an unsupervised neural language model. We use the word

vectors[9] trained on 100 billion words of Google News [50] to initialize the embedding vectors, then fine tuned them in the learning phase. We change the last layer of the architecture (i.e., softmax) to a fully connected (i.e., dense) layer for our task. The final architecture includes convolution, max pooling and fully connected layers.

Multilayer Perception (MLP): This is the multilayer feed-forward network with fully connected (dense) layers. The feed-forward network defines a mapping function and learns the value of the function parameters accordingly. In the model architecture we set 300, 200, and 100 as the hidden layer sizes respectively.

LSTM + Attention: This is the attention mechanism on top of Long-Term Short-Term Memory (LSTM) layer. The attention layer is designed according to [62]. The input of the LSTM are word vectors initialized to pre-trained vectors in [50]. We use a fully connected layer on top of the attention layer to produce the final output.

4.4.2 Evaluation Metrics

We utilize the following metrics to evaluate the performance of different models. Given the actual dwell time y_i and predicted dwell time $\hat{y}_i$ for article a_i ($i = 1, 2, \ldots N$). We calculate the *Mean Square Error* (*MSE*) as follows:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{4.8}$$

[9] Available at: https://code.google.com/archive/p/word2vec/

Table 4.1: Evaluation of different methods.

Method	**MSE**	**RAE (%)**
LR +LDA [6]	4835.74	90.75
LR + Doc2Vec [41]	4857.26	91.21
RF + LDA [6]	4750.10	87.96
RF + Doc2Vec [41]	4566.38	86.44
MLP	4122.35	80.79
Word2Vec+CNN [34]	4564.80	85.58
LSTM + Attention [62]	4553.85	90.66
Proposed Model	**3883.13**	**78.51**

Moreover, we calculate the *Relative Absolute Error* (*RAE*) as:

$$RAE = \frac{\sum_{i=1}^{N} |y_i - \hat{y}_i|}{\sum_{i=1}^{N} |y_i - \bar{y}_i|} \tag{4.9}$$

where $\bar{y}_i = \frac{1}{N} \sum_{i=1}^{N} y_i$. Note that *RAE* is between 0 and ∞.

4.4.3 Experimental Results

Table 4.1 shows the Mean Square Error (MSE) and the Relative Absolute Error (RAE) of different baseline approaches as well as the proposed model (with MSE and RAE equal to 3883.85 and 78.51% respectively). As can be seen, the proposed model consistently outperforms all models. For shallow models we learn the features using LDA and Doc2Vec

Table 4.2: Effect of different augmented Features.

Augmented Features	**MSE**	**RAE (%)**
PER	3966.15	79.49
PER+ORG	3963.55	79.36
PER+ORG+EVENT	3933.71	79.10
PER+ORG+EVENT+EMO	**3883.13**	**78.51**

approaches then train the model on Linear Regression (LR) and Random Forest (RF) accordingly. As it is illustrated, among the shallow models RF+Doc2Vec performs the best with MSE and RAE equal to 4566.38 and 86.44% accordingly. Moreover, although LDA has the advantage of interpretablity, our experiments show that Doc2Vec features (in combination with the Doc2Vec) work better for our prediction task. Furthermore, we observe that MLP (with MSE and RAE equal to 4122.35 and 80.79 accordingly) performs better than the other deep neural network based models, and among these baselines LSTM performs badly in comparison to other models in terms of RAE. Table 4.2 shows the effect of different augmented features on the overall performance. As we observe, augmenting all features, namely people (PER), organizations (ORG), events (EVN), and emotions (EMO), in the proposed model results in the best performance.

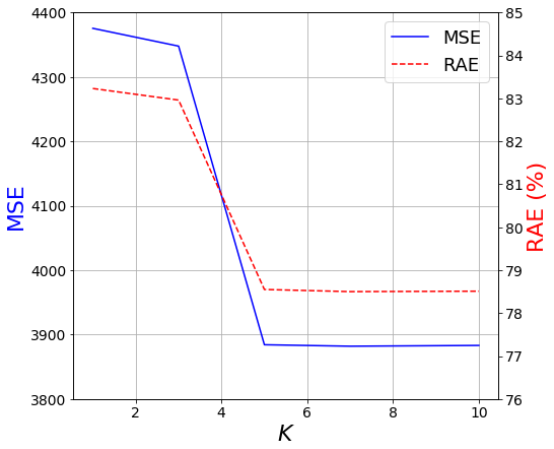


Figure 4.4: The number of hidden vectors.

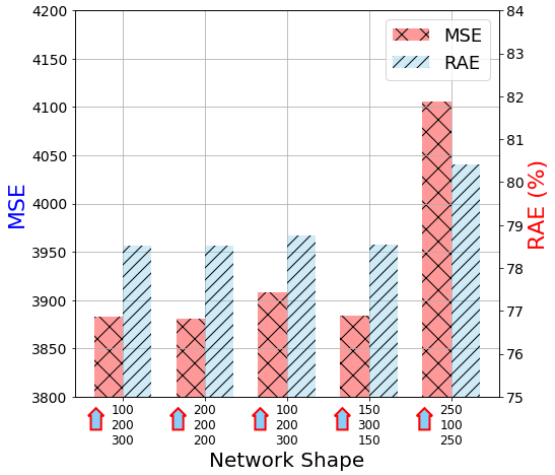


Figure 4.5: The architecture shapes.

4.4.4 Hyper parameter study

Figure 4.4 shows the model performance in terms of the number of hidden vectors per feature dimension. We increase the number of hidden vectors (i.e., K) in the factorization machine component and calculate the errors accordingly. As can be observed, the errors decrease significantly by increasing K from 1 to 5, then becomes stable. This suggests that a value between 5 to 10 would be a good choice for this parameter.

To see the effect of different deep component architecture shapes on the error measures, we keep the number of nodes constant (i.e., 600), and change the number of nodes in the hidden layers. Figure 4.5 shows the effect of selecting different architectures on the errors. As can be seen, the 250-100-250 is the worst among all architecture and 300-200-100 is slightly better than the others.

In order to study the effect of activation functions on the overall errors, we keep

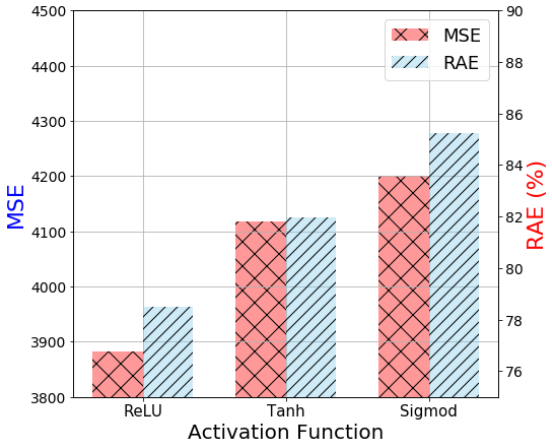


Figure 4.6: The activation functions.

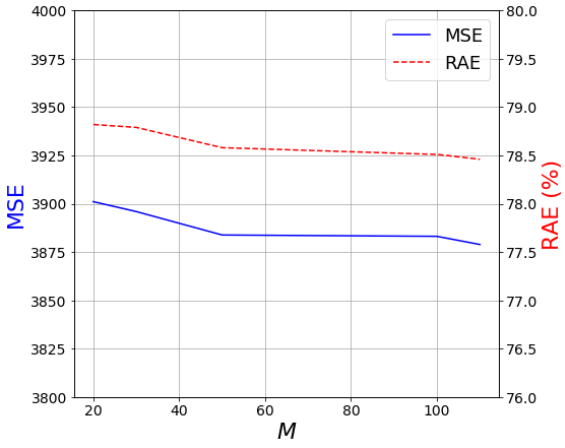


Figure 4.7: The hidden vector size.

the last layer activation function to ReLU (as it outputs a dwell time value which is always a positive real number) and change the other activation functions to $Tanh$ and $Sigmoid$, and then $ReLU$. Figure 4.6 shows the model errors for different activation functions. Among these activation functions, $ReLU$ gives the best performance and $Simod$ performs considerably worse than the others.

Figure 4.7 shows the effect of the hidden vector size (i.e., M) of factorization machine component on the overall errors. We observe that errors slightly decrease by increasing hidden vector size form 20 to 40, and then does not show any significant improvement for M between 40 to 100. As such, the proposed model is not sensitive to vector size and this parameter can be set with a value between 40 to 100.

Figure 4.8 shows the prediction errors for different numbers of layers of the deep component. As can be seen, the errors decrease as we increase the number of hidden

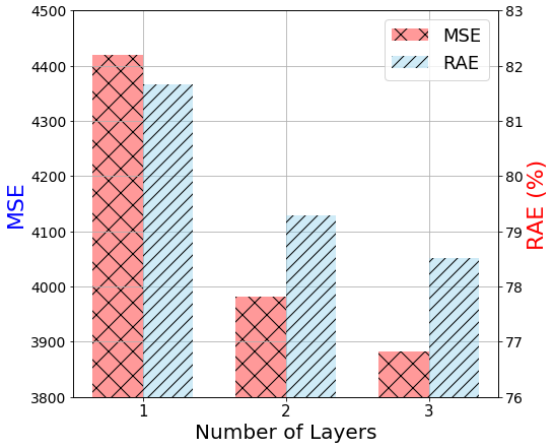


Figure 4.8: The number of layers.

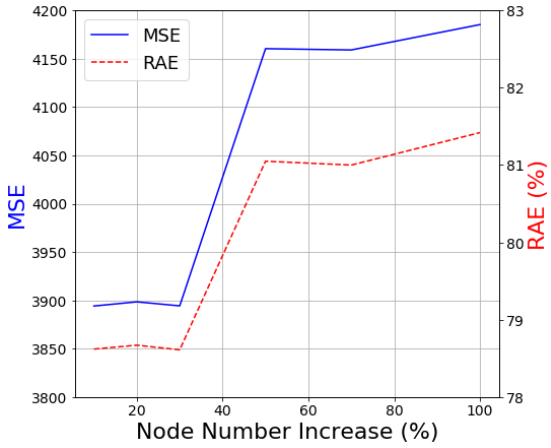


Figure 4.9: The number of nodes.

layers from 1 to 2 and is the best when it is 3.

In order to study the effect of neurons on prediction errors. We start from $300 - 200 - 100$ architecture and increase the hidden layer size by a certain percentage (i.e., 10%, 20%,...), then calculate the errors for each architecture. Figure 4.9 shows the performance of the model for different percentage of node number increase. We observe that the errors remain almost at the same levels when the node numbers in each layer increase by 30%, then starts to get worse from 30% to 100%. This could be due to the overfitting problem.

4.5 Summary

In this chapter, we proposed a novel model to predict the dwell time of an article based on its content. First, we extracted the main factors of article stories such as events,

emotions, peoples, and organizations, and then proposed a neural network architecture that learns the interactions among the extracted features. The final architecture combines the representations of these interactions (captured in a vector) with the article content representation to predict the article dwell time. The evaluation of the proposed model on a real dataset showed the superiority of it over the state-of-the-art baselines.As dwell time is a commonly used article engagement measure, the proposed method is of practical value for news agencies.

5 Time-aware Subscription Prediction Model

5.1 Introduction

Digital media and online news providers are facing the user acquisition challenge as the pressing issue more than before. In fact, from a business point of view, successful user acquisition can be directly translated to huge profits and values. However, whilst around 45% of people pay for a printed newspaper at least once a week, it has been much harder to persuade readers to pay for the online news subscription [57].

News recommender systems are widely exploited to improve the user experience, and consequently user acquisition indirectly. However, such systems mainly focus on recommending items that coincide with user's interests (to maximize the user's satisfaction) and do not identify potential subscribers and predict the subscription time. Identifying potential subscribers and predicting their subscription time are of paramount importance for news websites since it allows them to launch a targeted marketing campaign in advance. To the best of our knowledge, this problem has not been explored directly in the digital news media domain from data mining/machine learning perspectives, but rather

considered in marketing studies which need a lot of human efforts.

The problem of identifying potential subscribers for news media from the data mining/machine learning point of view is facing several challenges. First, a decision for subscription is under influence of many factors such as demographical, social, or cultural circumstances. For example, one might decide to subscribe as she/he was referred by her/his friend (e.g., word of mouth), or based on her/his good experience. Finding an appropriate set of predictors for identifying and recommending such users (i.e., potential subscribers) is a challenging problem. Second, domain knowledge is extremely limited for *"the decision to subscription"* process (i.e., the knowledge acquisition bottleneck). In other words, domain experts do not have a clear idea on who subscribes and why/when a subscription occurs. Third, subscription should be considered in combination with the time dimension. In fact, the predictive model should identify the potential subscribers in a right time (i.e., neither soon nor late) since targeting a user who is either not ready to subscribe yet or no longer interested in subscription (while was previously interested) by any marketing campaign results in no subscription.

In this paper, we propose an end-to-end solution to address the aforementioned challenges in the problem of identifying potential users prone to subscription in news portals. First, we argue that the subscription act is not an instantaneously sudden decision, but rather an informed decision based on previous positive experiences. Accordingly, we propose a set of engagement measures as subscription predictors. The engagement mea-

sures are quantified in fully data-driven fashion, so we do not rely on the domain expert knowledge for their calculation. Then, we propose a *Time-aware Subscription Prediction (TASP)* model that combines the time dimension with the suggested predictors. The proposed model not only identifies and recommends the users who are very likely to become subscribers but also is able to predict their subscription time. In the TASP model, we treat subscription time as a dependent random variable and utilize generalized linear model to combine all engagement measures (i.e., independent random variables). Then, we cast the problem into an optimization problem aiming to maximize the likelihood of the proposed model. The learning algorithm is designed and parameters of model are learned respectively. Our main contributions are as follows:

- We define the problem of *time-aware subscription prediction for user acquisition* in news portals and design an end-to-end data-driven solution based on the data which are usually available in news portals.

- We propose effective user engagement measures as the main component of the subscription prediction model and show that they have a good predictive power to model subscription occurrence/time.

- We argue that time is an important factor in user subscription prediction and develop a probabilistic model to recommend the trustworthy potential subscribers. The proposed model predicts the potential users prone to subscription before a

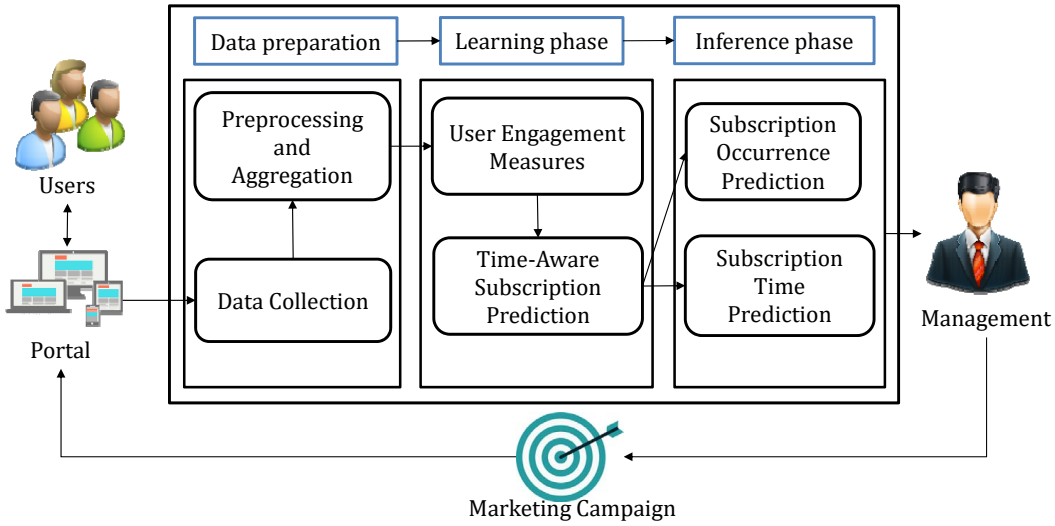


Figure 5.1: The proposed user acquisition framework.

given time. Moreover, it can predict when the subscription occurs.

- The conducted experiments on a real dataset show the effectiveness of the proposed framework and the developed model in solving the problem of time-aware subscription prediction for user acquisition.

The rest of chapter is organized as follows. Section 5.2 discusses the proposed framework for user acquisition. In particular, we present our Time-aware Subscription Prediction Model (TASP) in section 5.2.3. We outline the empirical evaluation in section 5.3. Section 5.4 concludes the chapter and present the future work.

5.2 Time-Aware User Acquisition in News Portals

Figure 6.3 shows an overview of the proposed framework for user acquisition in news portals. The framework consists of three main components: (1) *Data preparation*: most of news portals (e.g., The Globe and Mail[10]) use a data collection platform (e.g., Omniture by Adobe[11]) to capture the interactions with users. However, the captured data need to be preprocessed and aggregated before applying any learning algorithm (see §1.4 and §5.2.1). (2) *Learning phase*: given the preprocessed data, this component first finds a set of engagement measures (see §5.2.2) and then uses them to design the Time-aware Subscription Prediction (TASP) model (see §5.2.3). (3) *Inference phase:* as we learn the parameters of the proposed model, the interference models answer two type of questions: (i) time-aware subscription occurrence prediction: (i.e., what is the probability that a user becomes a subscriber by the given time t since the first visit?) (ii) subscription time prediction (i.e., when will a user become a subscriber since the first visit?). The inference outcomes can be utilized by the marketing campaign to boost user acquisition.

5.2.1 Preprocessing

We filter out the unnecessary attributes which are not needed in calculation of user engagement defined in the next section (§5.2.2). We perform data cleaning by removing

[10]www.theglobeandmail.com

[11]https://my.omniture.com

the outlier visitors whose engagement measures deviate more than 3 times the standard deviation from the mean of respective engagement measures in the data [37]. This helps us simply remove unreasonable values for the measures. Finally, all the engagement measures are normalized based on the z-score method.

5.2.2 User Engagement Measures

As we suggest that user engagement have a close relationship with user acquisition, one important task in the proposed framework is to measure user engagement. To understand the rationale behind the relationship, consider the scenario that we want to predict users prone to subscription based on the historical data stored as a clickstream collection. A reasonable assumption is that the user's decision on subscription is based on a long-term and short-term positive experiences rather than a sudden instantaneously thought. This is exactly related to the area of "user engagement" modeling. In fact, a well-known definition of engagement is based on "positive aspects" of user experience while interacting with an online application [40]. The positive aspects of experience are different among domains and applications and very hard to measure (e.g., visiting Twitter more frequently by a user in comparison to Facebook does not show essentially she/he has a better experience with Twitter due to differences in engagement patterns of these two social media). Moreover, other engagement measurement approaches such as self-reporting methods [47] (i.e., using questionnaires, surveys or interviews) and phys-

iological methods [31] (i.e., utilizing observational methods such as facial expression or speech analysis) are based on a small number of users while assuming to be the representative of the whole population.

Alternatively, as we aim to have a fully data-driven framework, we propose the following simple but effective web analytics measures, inspired by [40], to quantify the user engagement and show that they have predictive power for subscription prediction in digital news media domain.

Total Number of Paywall: In news portals which provide subscribed services, there is a restriction on the number of articles that a non-subscriber can read in a period of time. For example, in The Globe and Mail this period is one month. That is, as a visitor tries to read more articles, she/he is directed to a page asking for subscription (or login). This page is referred as a paywall. In our proposed approach, this interaction is used as an indicator of a user's interest in subscription. We calculate the total number of paywalls each user hits in all of her/his visits.

Average Number of Paywalls per Visit: This measure is calculated by normalizing the total number of paywalls by the number of visits.

Total Article Read: This measure is simply defined as the number of articles read by the user. There is difference between page visit and this measure. While in page visit we consider all of the pages (e.g., navigational or search pages), in this measure we only count article pages since they may better show the interest of users in contents and could

be more close to the real user engagement, considering situations where, e.g., we count the number of page visits when a user visits a lot of navigational pages while looking for a single article.

Average Number of Articles per Visit: This measure is the number of articles read by the user normalized by the number of visits.

Average Spent Time per Article: The time a user spent on each article is calculated based on the method described in §**??**. The average time spent per article is calculated by dividing the total time that the user spent on articles by the number articles she/he visited. This measure roughly shows how much a user is interested in articles.

Average Spent Time per Visit: This measure is defined as the time that the user spent on visits divided by the number of visits. Each visit time is calculated based on the sum of time that the user spent on all articles during the respective visit.

Total Spent Time: The total spent time is measured as the sum of time that a visitor spent on each article during all her/his visits.

Although these measures are the indirect proxy of real engagement our experimental results show their effectiveness for user subscription prediction.

5.2.3 Time-aware Subscription Prediction Model (TASP)

Given the set of engagement measures, in this section, we first outline the problem statement; then, in subsequent sections we describe our proposed Time-aware Subscription

Prediction (TASP) model in details. We utilize the generalized linear model as the building block of the model. By assuming an underlying distribution for subscription time (i.e., Weibull), we cast the problem into the maximum likelihood optimization. Finally, we derive the solution to learn the parameters of the model.

5.2.3.1 Problem Statement

Given the processed data for all the users, we refer to the time period of this data set as "exploration period". We first remove the users who subscribed before the exploration period. The remaining users either subscribed during the exploration period (i.e., *subscribers*) or never subscribed either before or during the period (i.e., *non-subscribers*). Note that we do not consider the users who subscribed before the exploration period since we do not have their information before their subscription and our targeted problem is to build a model to predict how and when the unsubscribed users turn to subscribed ones.

Definition 5.2.1 *(Subscription Occurrence Time): The subscription time $\widetilde{t_i}$ is defined as the time that passed since the first visit of user i until her/his subscription. Thus, given the absolute subscription time t'_i and the first visit time t_{f_i} for user i, $\widetilde{t_i}$ is computed as follows:*

$$\widetilde{t_i} = t'_i - t_{f_i} \tag{5.1}$$

The absolute subscription time refers to the timestamp that is recorded for each subscription. In our analysis, all timestamps are in day scale.

For non-subscribers, we define the possible subscription period as follows:

Definition 5.2.2 *(Possible Subscription Period): We define* $(\bar{t_i}, \infty)$ *as the possible subscription period for user i, where* $\bar{t_i}$ *is defined as:*

$$\bar{t_i} = t_{l_i} - t_{f_i} \tag{5.2}$$

where t_{l_i} is the last visit time in the exploration period for a non-subscriber. Alternatively, $\bar{t_i}$ might be considered as the time that subscription might occur afterward since the first visit for the user i. Please note that if the subscription occurs we know the exact time of subscription $(\widetilde{t_i})$, whereas in the case that the subscription does not occur, all we know is that the subscription time exceeds $\bar{t_i}$.

The training set for the subscription time prediction problem is defined as follows:

$$L = \{(X_i, t_i, I_i) | i = 1, 2, \ldots, n\} \tag{5.3}$$

where $X_i = [x_{ji}]_{m \times 1}$ is the engagement measure vector for the user i (x_{ji} is the *j'th* engagement measure calculated for the user i, see §5.2.2). We calculate the user engagement measures for subscribers based on the visits before the subscription time and for non-subscribers based on the first visit till the last visit in the exploration period. For simplicity, the vector of X_i is append by 1 to address the bias in the linear system. I_i is defined as the indicator function which specifies whether user i subscribed during the

exploration period or not:

$$I_i = \begin{cases} 1 & \text{if user i is a subscriber} \\ 0 & \text{otherwise} \end{cases} \tag{5.4}$$

and t_i is defined as $\widetilde{t_i}$ for subscribed users (i.e., $I_i = 1$) and $\bar{t_i}$ for non-subscribed user (i.e., $I_i = 0$). We refer to this arrangement in §5.2.3.4 as we want to formulate the optimization problem.

Let T be a non-negative continuous random variable representing the waiting time for subscription occurrence since the first visit. We assume $f_T(t)$ be the probability density function and $F_T(t) = P(T < t)$ (p.d.f.) be the cumulative distribution function (c.d.f.) of subscription occurrence by time t.

Now we define the problem of user subscription time prediction as follows. Given training data L (Eq. 5.3), we want to estimate the cumulative distribution function $F(t) = P(T < t)$ for any subscription time t.

5.2.3.2 Generalized Linear Model

In order to make a connection between subscription time (i.e., variable of interest) and engagement factors, we first develop a generalized linear model. The generalized linear model bridges the gap between the probability distribution of subscription time and engagement factors calculated for each user and parameterize our model from observed

data. Once the connection (i.e., a model) is established, we can predict the subscription time from the engagement behaviors.

Given vector X_i as the engagement measure vector (i.e., exploratory variables), subscription time observation for the user i is modeled as follows:

$$T_i = B^\top X_i + \epsilon \tag{5.5}$$

where ϵ is a stochastic residual coming from exponential family. The main idea is to model the expectation of subscription time as a function (i.e., link function) of linear combination of engagement measures. So,

$$E[T_i] = g^{-1}(B^\top X_i) \tag{5.6}$$

To ensure the strict positivity of $E[T_i]$, we assume g is the exponential function:

$$E[T_i] \propto \exp(-B^\top X_i) \tag{5.7}$$

This assumption also helps us simplify the objective function introduced in §5.2.3.4. Please note that if we choose Gaussian or Bernoulli distribution, the model will be reduced to linear regression or logistic regression respectively.

5.2.3.3 Underlaying Distribution for Subscription Time

As our goal is to model relationship between user engagement and subscription time, we need to find the proper distribution for predicting the subscription time. The Weibull

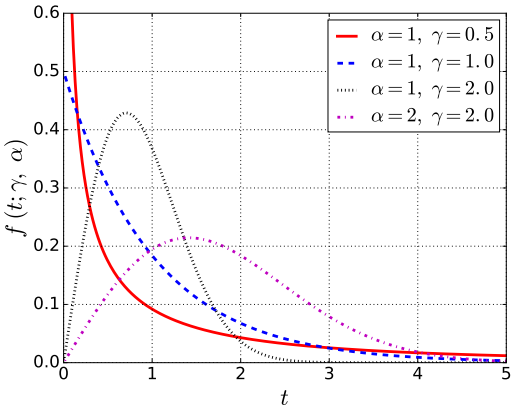


Figure 5.2: Weibull Distribution.

distribution has the flexibility to model right-skewed, left-skewed or even symmetric distributed data. Thus, we chose to use it in our model. It has been used in different domains to model the waiting time of an event [39]. The Weibull probability distribution for subscription time is as follows:

$$f_{T_i}(t;\gamma,\alpha) = \frac{\gamma}{\alpha_i}\left(\frac{t}{\alpha_i}\right)^{\gamma-1} \exp\{-\left(\frac{t}{\alpha_i}\right)^{\gamma}\} \tag{5.8}$$

where α_i and γ are scale and shape parameters respectively. The shape parameter γ can be learned to model the waiting time where the rate of event (i.e., hazard function) decreases ($\gamma < 1$), increases ($\gamma > 1$), or is constant ($\gamma = 1$) with time. Increasing the value of scale parameter (α_i) while holding shape parameter (γ) constant has the effect of stretching out the probability density function. Figure 5.2 shows the Weibull distribution for different parameters. The expectation of Weibull distribution is expressed as:

$$E[T_i] = \alpha_i \Gamma(1 + \frac{1}{\gamma}) \tag{5.9}$$

where Γ is the Gamma function. Given (Eq. 5.7), we can assume that:

$$\alpha_i = \exp\left(-B^\intercal X_i\right) \tag{5.10}$$

The cumulative distribution function is written as follows:

$$F_{T_i}(t) = P(T_i \leq t) = 1 - \exp\{-\left(\frac{t}{\alpha_i}\right)^\gamma\} \tag{5.11}$$

Note that the distributions in (Eq. 5.8) have the same shape parameter γ, but different expectation values via parameter α_i. In fact, the basic assumption is that each value of a random variable T_i is drawn from a distribution indicated in (Eq. 5.8) where the expectation of distribution depends on the data point in (Eq. 5.9 and 5.10).

5.2.3.4 Optimization Problem

Assuming that observations (i.e., data points) are statistically independent and drawn from the distribution (Eq. 5.8), the log-likelihood of the model is formulated as follows:

$$\log \ell = \sum_{i=1}^{n} \{I_i \log f_{T_i}(t_i; \gamma, \alpha_i) + (1 - I_i) \log P(T_i > t_i)\} \tag{5.12}$$

where t_i is the subscription occurrence time ($\widetilde{t_i}$) for the subscriber i (i.e., $I_i = 1$) and the start of possible subscription period ($\bar{t_i}$) for the non-subscriber i (i.e., $I_i = 0$). The basic idea is that subscribers contribute to the log-likelihood by the probability density function f_{T_i} while non-subscribers contribute to log-likelihood by the probability $P(T_i > t_i)$. If we plug in the probability density function in (Eq. 5.8) and the cumulative distribution

function in (Eq. 5.11) into the log-likelihood function (Eq. 5.12), we can simplify the log-likelihood of model in vector format as follows:

$$\log \ell = I^\intercal(\log(\gamma)\mathbb{1} + (\gamma - 1)\log(T_s)) + \gamma\, I^\intercal \mathbf{X}B - \mathbb{1}^\intercal \exp\{\gamma\,(\log(T_s) + \mathbf{X}B)\} \tag{5.13}$$

where $I = [I_i]_{n\times 1}$ is the indicator vector whose components are defined in (Eq. 5.4), $\mathbb{1} = [1]_{n\times 1}$ is the identity vector (i.e., all components are 1), $T_s = [t_i]_{n\times 1}$ is the vector of subscription time defined in (Eq. 5.3), $\mathbf{X} = [X_i]_{n\times m}$ is the matrix of engagement measures, where each row is X_i defined in (Eq. 5.3), and γ (scaler) and $B = [\beta_i]_{m\times 1}$ (vector) are parameters.

5.2.3.5 Learning Algorithm

We use the gradient ascending method to maximize the log-likelihood and learn the parameters of the proposed model. First, we derive the gradient of log-likelihood of model (Eq. 5.13) with respect to γ and B. The gradient of model with respect to B is specified as follows:

$$\nabla_B[\log \ell(B, \gamma)] = \gamma \mathbf{X}^\intercal I - \gamma \mathbf{X}^\intercal \exp\{\gamma\,(\log(T_s) + \mathbf{X}B)\} \tag{5.14}$$

and gradient of log-likelihood with respect to the γ is derived as follows:

$$\nabla_\gamma[\log \ell(B, \gamma)] = I^\intercal\{(1/\gamma)\mathbb{1} + \log T_s + \mathbf{X}B\} - (\log(T_s) + \mathbf{X}B)^\intercal \exp\{\gamma(\log(T_s) + \mathbf{X}B)\} \tag{5.15}$$

Algorithm 1: TASP Learning Algorithm

1 Initialize $B_1^{(0)}, B_2^{(0)}, \ldots B_m^{(0)} randomly$

2 Initialize $\gamma^{(0)} \leftarrow 1$

3 $t \leftarrow 0$

4 **while** *not converge* ***and*** $t < max\ iterations$ **do**

5 $\quad B^{(t+1)} \leftarrow B^{(t)} + \eta \nabla_B [\log \ell(B^{(t)}, \gamma^{(t)})]$

6 $\quad \gamma^{(t+1)} \leftarrow \gamma^{(t)} + \eta \nabla_\gamma [\log \ell(B^{(t+1)}, \gamma^{(t)})]$

7 $\quad t := t + 1$

8 **return** B, γ

The overall procedure is outlined in Algorithm 1. We use the coordinate ascending method [80] to learn the B and γ iteratively. In step 5, we update the parameter B based on the gradient derived in (Eq. 2.14), then in step 6, keeping B fixed, the parameter γ is updated according to the gradient in (Eq. 2.15).

5.2.4 Inference models

After the parameters of the model (i.e., γ and B) are learned, inference with the model is straightforward. Particularly, we are interested in answering two types of questions: (1) what is the probability that a user be subscriber by the given time t since the first visit? (time-aware subscriber prediction) (2) when will a user be a subscriber since the

first visit? (subscription time prediction).

5.2.4.1 Time-aware Subscription Occurrence Prediction

To find the users who will be subscriber by time t since the first visit, we need to estimate the $P(T \leq t)$. Given the user $\widehat{u}$ has a engagement vector $X_{\widehat{u}}$, we calculate the scale parameter $\alpha_{\widehat{u}}$ using (Eq. 5.10):

$$\alpha_{\widehat{u}} = \exp\left(-B^{\intercal} X_{\widehat{u}}\right) \tag{5.16}$$

The desired probability is calculated as follows:

$$F_T(t) = P(T \leq t) = 1 - \exp\{-\left(\frac{t}{\alpha_{\widehat{u}}}\right)^{\gamma}\} \tag{5.17}$$

We consider $F_T(t) = P(T \leq t) \geq 0.5$ as the subscription occurrence.

5.2.4.2 Subscription Time Prediction

For the subscription time prediction, as the final distribution can be skewed, we propose to use the median as prediction time. This measure is less susceptible to outliers and extreme values and empirically performs better in our experiments. Given user $\widehat{u}$ with $X_{\widehat{u}}$ as the engagement vector, the subscription time t for the user $\widehat{u}$ is calculated as follows:

$$t_{\widehat{u}} = \alpha_{\widehat{u}} \log(2)^{\frac{1}{\gamma}} \tag{5.18}$$

where $\alpha_{\widehat{u}}$ is estimated using (Eq. 5.16).

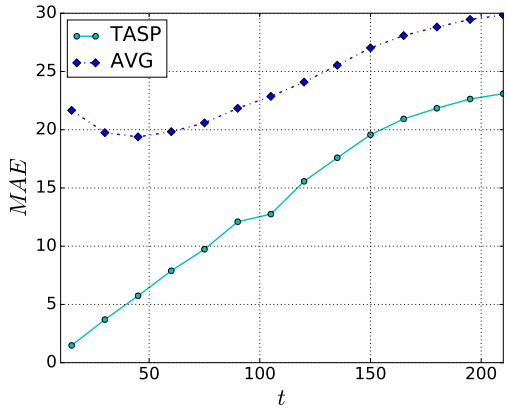


Figure 5.3: Subscription time prediction performance.

5.3 Empirical Evaluation

In this section, we evaluate our proposed Time-aware Subscription Prediction (TASP) model and compare it with the state-of-the-art techniques as the baselines. We compare our model with Logistic Regression (LR), Random Forest (RF), Decision Tree (J48) and Naive Bayes (NB). We use the Mean Absolute Error (MAE) and F_1-Measure as performance measures for “subscription time" and “subscription occurrence prediction" accordingly. All the experiments in this section are based on the 10-fold cross validation. All the time values in the experiments are in day scale. We use The Globe and Mail dataset in our experiments. We set the learning rate η and maximum number of iterations (i.e., *max iterations*) in Algorithm 1 to 0.01 and 1000 respectively. All the experiments are conducted on The Globe and Mail dataset described in §1.4.

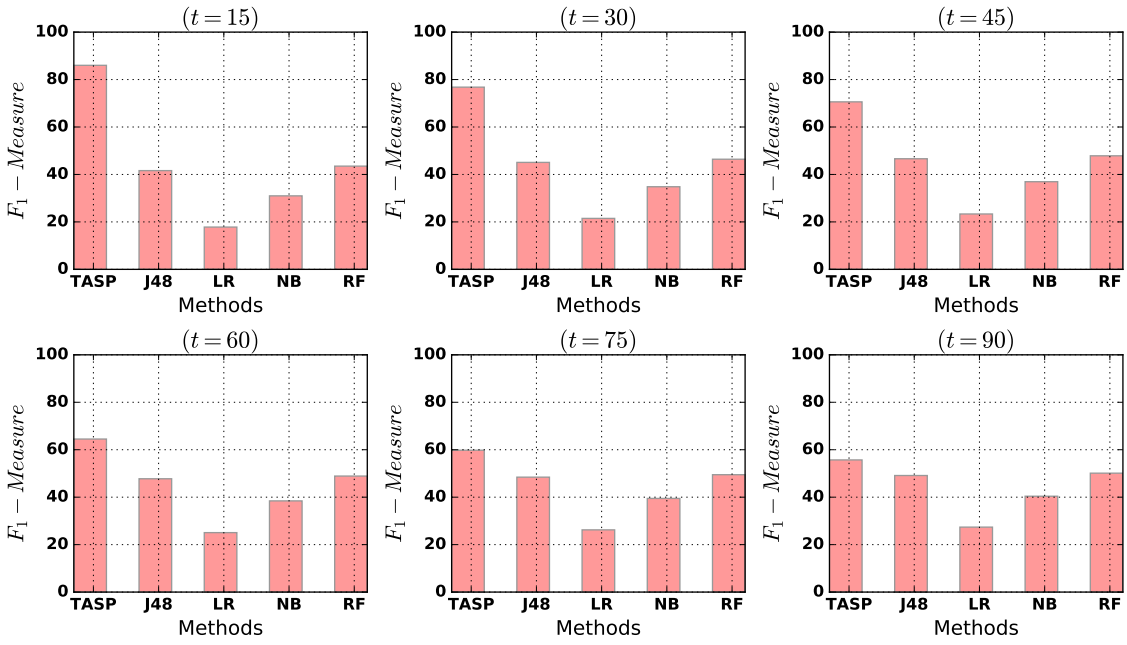


Figure 5.4: Subscription occurrence prediction performance (all time values are in days).

5.3.1 Subscription Time Prediction

Figure 5.3 shows the results of subscription time prediction for the proposed model (TASP) and Average Time (AVG) as the baseline. Each point in the figure shows the MAE between the predicted subscription time and actual subscription time for users who subscribe before time *t* (all time values are in days). For the AVG model we calculate the average subscription time of visitors who subscribed before time *t* in the training set. Then, the MAE is calculated based on the difference between the actual subscription time of users in the test set and the respective average time value. As observed, MAE

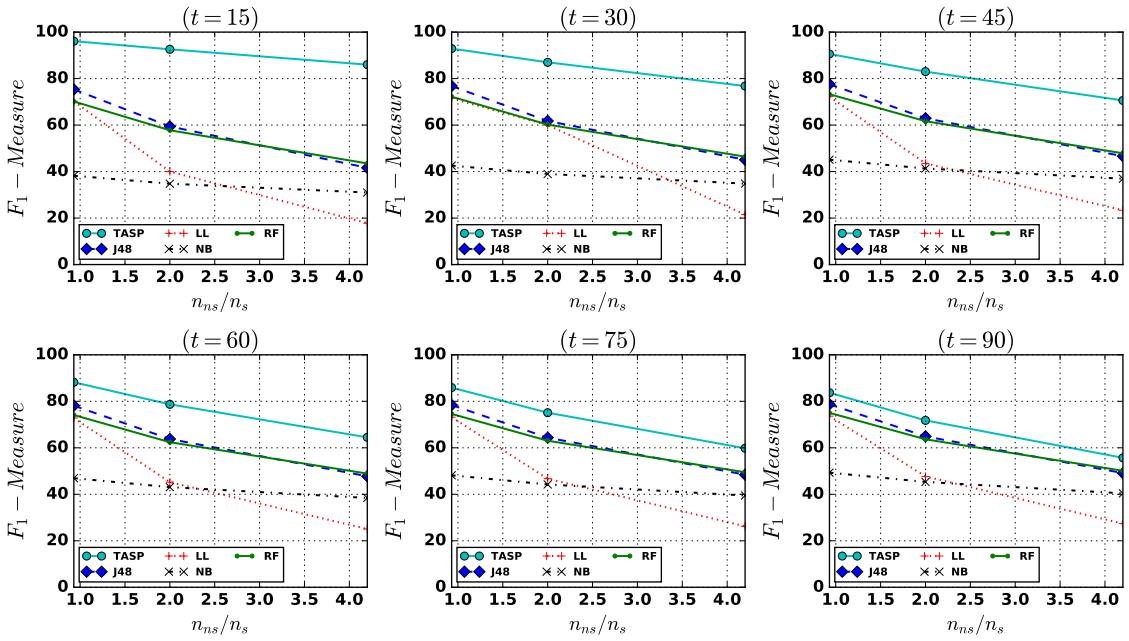


Figure 5.5: The subscription occurrence prediction performance sensitivity with respect to the number of non-subscribers to subscribers (n_{ns}/n_s).

for the proposed method is much less than the AVG method for different t. In particular, for small values of t, the proposed model performs better than bigger time values, which means that the proposed method works better in short-time subscription time prediction than it does in longer term prediction although it performs better than the AVG method in both short term and long term.

5.3.2 Subscription Occurrence Prediction

Figure 6.7 shows the performance of TASP compared to the other baselines for different values of t. Each figure shows the performance of the different models in predicting the subscription occurrence before time t. Note that the proposed model (TASP) considers the time in the training stage and answers the queries about subscription with respect to the time (i.e., probability of subscription before given time t). Figure 6.7 shows that the proposed model outperforms the baselines for different t values. Moreover, it can be seen that the TASP model performs better in short-time subscription prediction. Among the baselines, tree-based models (i.e., J48 and Random Forest) perform the best and the Logistic Regression has the worst performance.

5.3.3 Imbalanced Sensitivity Analysis

In this section, we study the performance sensitivity of the proposed model under different portions of non-subscribers to subscribers as the training data. As such, we vary the portion of non-subscribers to subscribers by down sampling the non-subscribers. Figure 6.5 shows the performance of the proposed model (TASP) as well as the baselines for different portions of non-subscribers to subscribers (n_{ns}/n_s) where n_{ns} and n_s are the number of non-subscribers and subscribers respectively in the training set. The performance of the proposed model in predicting the subscriber is better when the dataset is

balanced and consistently better than the baselines for different portions. As our model performance is better in the case that the dataset is balanced, we are aiming to embed a mechanism in our model to deal with imbalanced data as the future work. Figure 5.6 shows the performance sensitivity of the proposed model in predicting the subscription time. As can be seen the MAE has a small sensitivity to portion of non-subscribers to subscribers in the training set.

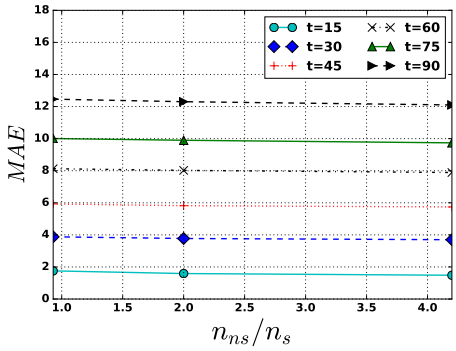


Figure 5.6: The subscription time prediction performance sensitivity with respect to the number of non-subscribers to subscribers (n_{ns}/n_s).

5.4 Summary

User acquisition for digital news portals are one of the most pressing issue as the users are exposed to many available news sources. In this chapter, we addressed the problem by predicting users who are prone to subscription in a given period of time. One important challenge is to define the measures that have enough power to predict the subscription

(since the subscription is a complex decision depending on many factors). We simply showed that engagement measures had the good capability in predicting the subscription. The intuition is that the engagement as a positive experience has a direct impact on subscription. We proposed a time-aware prediction model that not only could predict the subscription in a given period of time, but also the subscription time. The empirical study on a real dataset showed that the proposed model performed well compared to the baseline models. In the future, we plan to improve and embed a mechanism in the model to deal with imbalanced data (for the situation that the number of subscribers to non-subscribers are very low). We will also investigate the capability of the proposed model in other domains.

6 Adaptive Paywall Mechanism for Digital News Media

6.1 Introduction

Most online newspapers across the world generate revenue by displaying advertisements or/and using a pay model that restricts the reader access to articles via a paid subscription. In the former one, news agencies (e.g., USA Today) operate based on an ad-supported free content model, in which the articles are accessible for free and the revenue is derived from displaying advertisements. However, advertisement revenues may not be sufficient to sustain existing forms of news production as they do not create long term relationships with customers [8, 55]. Therefore, pay models, also known as *paywall mechanisms*, were developed by digital media to increase revenues through subscription. In such models, news agencies (e.g., The New York Times, and The Globe and Mail) offer a certain number of free articles in a period of time (e.g., a month) and then redirect visitors to the subscription page (i.e., paywall) to continue reading articles. The ultimate goal of a paywall mechanism is to persuade users to subscribe and as a result boost the profit. However, user persuasion for subscription (i.e., user acquisition) is not an easy task in

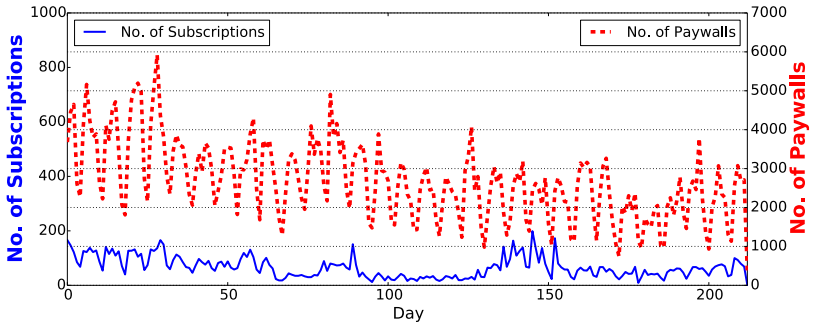


Figure 6.1: The number of subscriptions vs. the number of paywalls in The Globe and Mail dataset for the period 2014-01 to 2014-07. There is a weak correlation (i.e., $\rho = 0.59$) between the two numbers.

news domain since users usually have many choices in selecting news sources. Moreover, in most cases there is no direct relationship between the number of paywalls presented to readers and the number of subscriptions. That is, by increasing the number of times that paywalls presented to readers, we may not necessarily raise the number of subscriptions (see Figure 6.1). Therefore, the traditional paywall mechanism based on the total number of articles read in a period may not serve its purpose of increasing revenue. It may actually turn away many potential subscribers.

Due to the increase of technological obstacles for ad-supported free content revenue models (e.g., ad blockers) and the number of online news providers seeking for sustainable relationship with customers (78 % of U.S. newspapers with circulations over 50,000 are using a digital subscription model [79]), the need to generate revenue by developing effective paywall mechanisms is demanding. This is due to the fact that blocking a user from reading articles at a wrong time may disengage the user too early or allow a

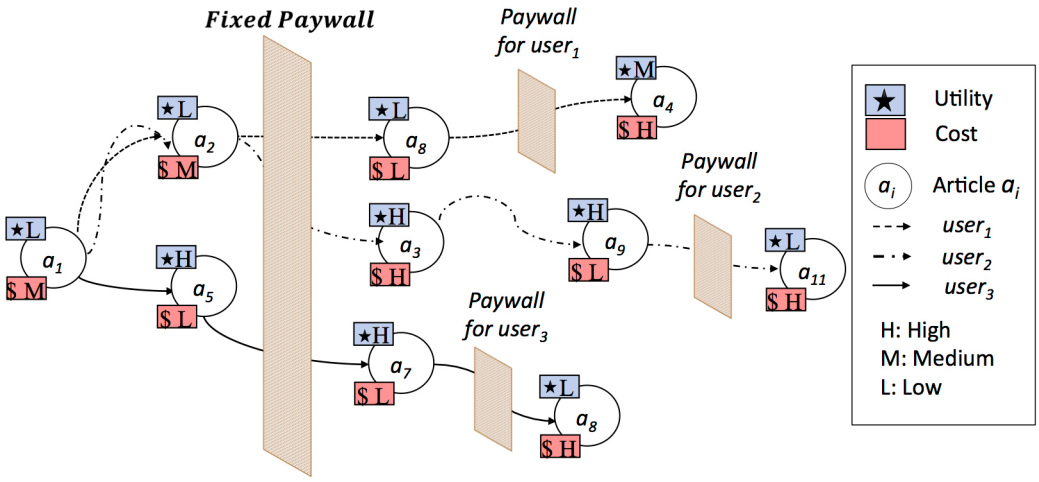


Figure 6.2: Incorporating utility and cost in paywall mechanisms.

non-potential subscriber to read too much content for free. Therefore, developing a smart paywall policy is of paramount importance to the prosperity and profitability of an online newspaper. The availability of users' interaction data and advances in machine learning techniques raise an interesting question: *Can we estimate how many and what articles a particular user should be allowed to read before a paywall?* It is unrealistic to expect the same answer for all users. Moreover, the answer should consider business objectives even if they could be in conflict with one another. For example, allowing readers to read more articles leads to more display of ads, increasing the ad-based revenue. However, from the subscription point of view, this is not desirable as offering too much free content makes subscription unnecessary from the reader's point of view.

Finding the optimal paywall time is a sequential decision-making process (or a sequential decision problem), in which a decision whether to show a paywall needs to be

made at each time point during a reading session and once the paywall is presented, the session terminates. To define an objective function for this problem, we introduce the notion of *utility* and *cost*. The utility of an article measures the effectiveness/usefulness of the article in achieving a business objective (e.g., user engagement which can increase the subscription possibility). The cost of an article measures the amount of resources consumed to prepare it (e.g., the amount of money paid to the author). Traditional paywall models (called *fixed paywall* or *metered paywall*) block users from reading articles after visiting a certain number of articles (e.g., two articles) without considering the above two or any other factors. Figure 6.2 shows a toy example demonstrating how the utility and cost of an article can be used to make smarter decisions. Assume that $user_1$ and $user_2$ both visit two low-utility articles a_1 and a_2 (e.g., ones that can be found in many other news sources). If the fixed paywall policy with the limit of two articles is used, both of them would be directed to the paywall. Assume we do not use the fixed paywall but consider what they would like to read next. Suppose $user_1$ clicks on a_8. Since a_8 has a low cost, we can show it to the user and take advantage of other benefits (e.g., displaying ads when showing a_8). However, $user_1$ might not be a good target for subscription as all the articles she has read have a low utility (e.g., can be found somewhere else), so once she clicks on the next article (i.e., a_4), which has a high cost, the paywall is presented. On the other hand, $user_2$ can see the article a_3 as its high cost can be justified by its high utility, and then receives the paywall after visiting a_9, where the next article the user se-

lects has low utility and high cost (i.e., a_{11}). Moreover, a user (e.g., $user_3$ in Figure 6.2) who visits high utility articles (i.e., a_5, a_7), which may be articles visited by subscribers before their subscription, is more likely to become a subscriber. Therefore, we show the paywall (after a_7) since the next article (e.g., a_8) has a high cost and low utility. Note that this user is more likely to become a subscriber and presenting the paywall at the right time might persuade her to subscribe. This example shows that different reading behaviors may require different paywall strategies, and making decisions based on the utility and cost model can serve as an effective approach to make a smart decision per user visit.

However, finding the optimal paywall time for a user is a challenging problem. First, while the concepts of utility and cost provide insights into the paywall decision problem, incorporating them into the optimal decision making process in a disciplined way is not a trivial task. Second, when making a decision at a time point (i.e., when the user clicks on an article), we only know the articles that the user has visited and the next article she is trying to read. The articles beyond the next one are unknown. Thus, when making an optimal decision at an early point, we need to consider the uncertainty as to what would happen later. Third, a proposed model is better to be flexible so that any objective can be plugged in as desired. Last but not least, the proposed approach should be efficient to work in an online setting.

To address these challenges, we formulate the problem in a unified stochastic decision

making framework by considering the utility and cost of articles. We define the main components of the model, propose an effective approach to solving the problem, and provide theoretical supports for the proposed approach accordingly. The main idea is that at each stage the paywall will be presented to a user if the prospective articles (which are likely to be visited by the user) are not promising in terms of the utility and cost.

Our main contributions are as follows:

- We define the new problem of adaptive paywall mechanism in digital news media. While this problem is a major issue in subscription-based online news agencies, to best of our knowledge, it has not been studied before.

- We cast the problem into a sequential stochastic optimization problem in a disciplined way, and propose an effective data-driven solution accordingly. In particular, we provide the theoretical analysis and design an effective policy for the problem. The proposed framework is general in that it can be applied with any given business objective.

- We apply the proposed framework to a real dataset obtained from a major Canadian newspaper and show that it outperforms some baseline approaches in terms of different business objectives.

The rest of the chapter is organized as follows. The problem and framework components are defined in Section 6.2. We describe the proposed model in Section 6.3.

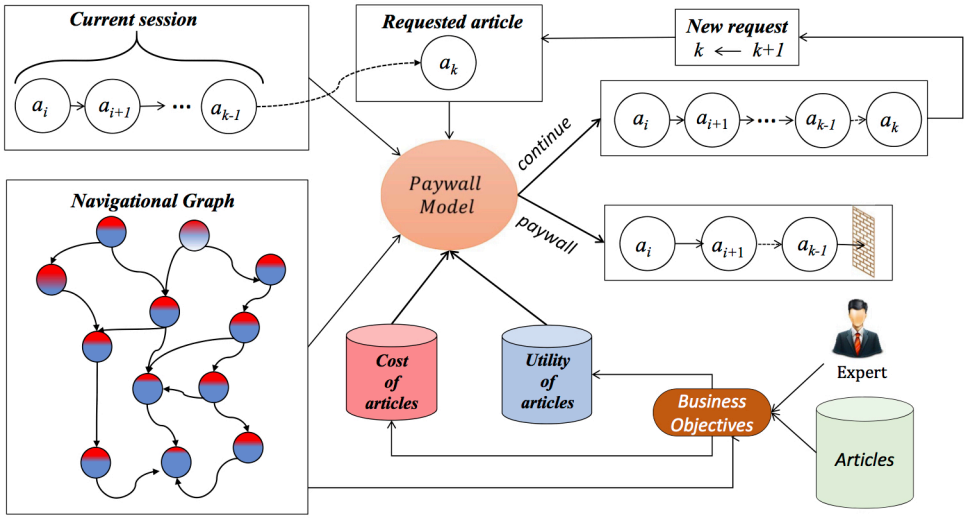


Figure 6.3: The proposed adaptive paywall framework.

Section 6.4 presents an application of the proposed method and its empirical evaluation, and finally Section 6.5 concludes the chapter.

6.2 Problem Definition

We describe the main components of the proposed model and define the problem accordingly. Figure 6.3 shows the proposed framework for adaptive paywall. The main components of the framework are: *utility*, *cost* and the *navigational graph* as well as the *paywall model*. The paywall model receives an article request, makes the decision, and changes the current state of the user session accordingly.

Definition 6.2.1 (Utility of article) *The utility of article a_i, denoted as $\phi(a_i)$, measures the effectiveness/usefulness of the article in achieving a business objective (e.g., user*

engagement which can increase the subscription possibility).

The utility of an article can be determined by domain experts or learned from historical navigational patterns in a data-driven fashion. For example, if a high percentage of the non-subscribed users reading an article subscribed to the newspaper later, the article has a high utility.

Definition 6.2.2 (Cost of article) *The cost of article a_i, denoted as $\psi(a_i)$, specifies the amount of resources (e.g., time, monetary cost) allocated to produce it.*

The cost can be specified in different ways. For example, the amount of money which the newspaper has to pay to the author, number of pages, etc.

Definition 6.2.3 (Navigational graph) *Navigational graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{W})$ is a directed graph, where $\mathcal{V}$ is a set of vertices representing articles, $\mathcal{E}$ is a set of directed edges where an edge e_{ij} from article a_i to a_j indicates a_j has been viewed right after a_i by at least one user, and weight $w_{ij} \in \mathcal{W}$ on e_{ij} represents the number/percentage of the users that read a_j right after reading a_i.*

The navigational graph encodes historical user navigation behaviours and is used in our model to estimate what article(s) a user is likely to visit next. Although the graph can be built based on all reading sessions that have occurred, we build the navigational graph based on the sessions made by subscribed users since they do not receive paywalls, and thus less bias is introduced.

A *session* is a group of activities (e.g., requesting and reading an article) that a user spends during one visit to the online newspaper. Traditional paywall mechanisms may consider the information in one or more sessions of an unsubscribed user in making a paywall decision (e.g., presenting the paywall if the total number of articles read by a user exceed a limit for a month). Since an unsubscribed user does not have an ID, user identification for each session (e.g., based on IP addresses or cookies) is necessary in order to consider multiple sessions of a user. However, since different users may use the same IP address and cookies can be blocked, tracking users across multiple sessions may be problematic. Thus, we focus on session-based paywall, although the proposed model can be applied to multiple sessions of a user.

Definition 6.2.4 (Session-based Paywall) *In this model, the paywall decision (i.e., whether to present the paywall at a time point in a session) is made based on the information that the user provides in the session without considering historical records of this particular user beyond the session. Once the paywall is presented, the session terminates.*

An example of session-based paywall is the traditional fixed/metered paywall that allows a user to read a fixed maximum number of articles (e.g., 2 articles) in a session. In this work, we propose an adaptive session-based paywall model in which the number of articles that a user can read depends on what the user has read/requested in the session and estimations of what the user might read in the future. Our adaptive paywall problem is defined as followed.

PROBLEM STATEMENT (**Adaptive paywall**): Given a navigational graph of subscribed users and a session that an unsubscribed user started, the goal is to determine at what time point during the session the paywall is presented so that the following objective function is maximized:

$$\frac{\sum_{i=0}^{k} \phi(a_i)}{\sum_{i=0}^{k} \psi(a_i)}, \tag{6.1}$$

where $\phi(a_i)$ and $\psi(a_i)$ are the utility and cost of the ith article that the user reads and $k+1$ is the total number of articles the user reads in the session before the paywall is presented. Although we define the objective function using the utility-to-cost ratio, it can be defined in different ways based on business objectives.

6.3 Proposed Method

The adaptive paywall problem is a sequential decision problem, that is, at each time step when the user requests an article in a session, a decision needs to be made regarding whether to allow the user to read the requested article or to present the paywall. Formulating and solving such a problem in a disciplined way is not a trivial task. The major challenge is the huge search space due to the high dimensionality of the problem. At each time step when the user requests a new article, in order to find an optimal solution, we need to look at not only what the user has read and requested, but also what the user would request or read in the future if the paywall is not presented at the current time, in order to compare with the values of the objective function at future time steps.

However, what the user will request or read is uncertain at the current time. Considering all possibilities at each time step to find exact solutions is prohibitable due to the large number of articles and combinations of them over multiple time steps. Thus, we resort to approximate solutions. Below we formulate the adaptive paywall problem using the *approximate dynamic programming* paradigm [61] and design a data-driven lookahead policy that makes decisions based on predicted behaviour of the user in the future to solve the problem.

6.3.1 Proposed Paywall Model

One of the most important tasks in approximate dynamic programming (and in particular, in sequential decision making) is to design a model of the problem, that is, to design the components of the problem. However, despite the importance of this step, there is no standard approach to modeling the problem [61]. A good model can facilitate the design of policy for solving the problem and may also allow the change of the assumptions (e.g., how to combine utility and cost) based on business objectives and providing alternative solutions accordingly. The major components of a sequential stochastic decision problem are: state variable, decision variable/function, transition function, and contribution function. Given the main components of the model the objective function can be defined accordingly.

State variable S_t: The state variable at time step t of a session is defined as S_t =

$(\overline{u}_t, \overline{c}_t, a_t)$, where a_t is the *article requested* at time t by the user, and $\overline{u}_t$ and $\overline{c}_t$ are the sum of utilities and costs of visited and requested articles (including a_t) by time t in the session, respectively.

The state S_t encapsulates the accumulated information in a *session* by time t, which is used to make the decision at time t. By defining the state in this way, the next state will not depend on the states before this state, which makes the decision process satisfy the Markov property. Meanwhile, the state representation should contain *minimally* necessary information, otherwise it may make the problem computationally intractable. For example, listing all the articles the user has visited in the session so far in a state results in a much richer representation, but it causes the state space to grow exponentially and makes the problem computationally intractable.

Decision function $X^\pi(S_t)$: The decision function determines the decision/action given state S_t using a policy π, where π is a function that maps S_t into a decision/action. In our problem, there are two possible decisions: presenting the paywall or not presenting it. The decision function is defined as follows:

$$x_t \triangleq X^\pi(S_t) \triangleq \begin{cases} 1 & \text{if } \pi(S_t) \text{ is "presenting paywall"} \\ 0 & \text{otherwise} \end{cases} \tag{6.2}$$

where x_t is called a *decision variable* indicating the decision taken at time step t. In our framework, policy π is to be designed using a data-driven method.

Transition function S^M: This function depicts the way that the proposed model evolves from one state to another one as a result of decision and exogenous information (i.e., a requested article at the next time step). The transition function S^M determining the transition from state S_t to S_{t+1} given decision x_t is defined as follows:

$$S_{t+1} = S^M(S_t, x_t, \hat{a}_{t+1}) \tag{6.3}$$

where S^M maps the components of S_t to S_{t+1} as follows:

if $x_t = 0$

$$a_{t+1} = \hat{a}_{t+1} \tag{6.4}$$

$$\overline{u}_{t+1} = \overline{u}_t + \phi(\hat{a}_{t+1}) \tag{6.5}$$

$$\overline{c}_{t+1} = \overline{c}_t + \psi(\hat{a}_{t+1}) \tag{6.6}$$

else

$$a_{t+1} = \text{“}paywall\text{”}$$

where $\hat{a}_{t+1}$ is the article to be requested at time $t+1$. Note that at time t, $\hat{a}_{t+1}$ is uncertain, and thus it is information that arrives exogenously, representing a source of randomness. As a result, its utility $\phi(\hat{a}_{t+1})$ and cost $\psi(\hat{a}_{t+1})$ are random. Also note that if $x_t = 1$ (i.e., the decision at time t is to present the paywall), we assign *paywall* to a_{t+1} to indicate the end state of the decision process. We denote the end/paywall state as S_p.

Contribution function: The immediate contribution/reward function of decision x_t in

state S_t measures how much decision x_t at state S_t contributes towards the final objective of the decision process, and is defined as follows:

$$C(S_t, x_t) \triangleq \begin{cases} (\overline{u}_t - \phi(a_t))/(\overline{c}_t - \psi(a_t) + 0.05) & \text{if } x_t = 1 \\ 0 & \text{if } x_t = 0 \text{ or } S_t = S_p \end{cases} \tag{6.7}$$

Note that when $x_t = 1$ (i.e., when the decision is to present the paywall), a_t is not presented to the user, thus a_t's utility and cost are not included in the accumulated utility and cost when the ratio is computed in the contribution function. 0.05 in the denominator is to avoid zero division. Also, we set the contribution to zero when $x_t = 0$, so that the contribution is only collected at the paywall time because the contribution at the paywall time considers the utilities and costs of all the articles that user reads in the whole session.

Paywall decision problem: We define the paywall decision problem as finding a policy π that maximizes the following objective function:

$$\mathbf{E}\{\sum_{t=0}^{\infty} \gamma^t \, C(S_t, X^{\pi}(S_t))\} \tag{6.8}$$

where $\gamma \leq 1$ is a discount factor (emphasizing that contributions in the future is not important as the current time contribution).

6.3.2 Policy Design

Solving the optimization problem defined in (6.8) directly is computationally intractable [61]. In this section, we convert Equation (6.8) into state value functions and analyze different

possibilities for the policy design, and in the next section discuss the proposed method.

Let $V^{\pi}(S_t)$ (called the value of state S_t with respect to policy π) be the expected total contribution of a session starting from state S_t and following policy π. That is,

$$V^{\pi}(S_t) = \mathbf{E}\{\sum_{t'=t}^{\infty} \gamma^{t'-t}\, C(S_{t'}, X^{\pi}(S_{t'}))\}$$

$$= C(S_t, X^{\pi}(S_t)) + \gamma E[V^{\pi}(S_{t+1})|S_t]. \tag{6.9}$$

Clearly, Equation (6.9) is the same as the objective function (6.8) when using S_t to denote the initial state. Thus, maximizing Equation (6.8) is equivalent to maximizing (6.9).

Theorem: The optimal value of Equation (6.9) is given by:

$$V^{\pi^*}(S_t) = \max_{x_t \in \{0,1\}} \{x_t C(S_t, x_t) + (1 - x_t)\, \gamma \mathbf{E}[V^{\pi^*}(S_{t+1})|S_t]\} \tag{6.10}$$

where π^* is the optimal policy and x_t is the optimal decision at state S_t based on π^* (i.e., $x_t = X^{\pi^*}(S_t)$).

Proof:

According to Equation (6.9) and Bellman's Principle of Optimality [74] that states an optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision, we have:

$$V^{\pi^*}(S_t) = \max_{x_t \in \{0,1\}} \{C(S_t, x_t) + \gamma \mathbf{E}[V^{\pi^*}(S_{t+1})|S_t]\} \tag{6.11}$$

where $x_t = X^{\pi^*}(S_t)$. We need to show Equation (6.11) is equivalent to (6.10). There are two possible decisions in state S_t:

a) If $x_t = 1$, the paywall is presented and thus $S_{t+1} = S_p$ (the end state). Since no contribution can be obtained at the end state and the process ends, $\mathbf{E}[V^{\pi^*}(S_{t+1})|S_t] = 0$. Thus, according to (6.11), $V^{\pi^*}(S_t) = C(S_t, x_t) = x_t C(S_t, x_t)$.

b) If $x_t = 0$, $C(S_t, x_t) = 0$ according to (6.7). Thus, according to (6.11), $V^{\pi^*}(S_t) = \gamma \mathbf{E}[V(S_{t+1})|S_t] = (1 - x_t)\gamma \mathbf{E}[V(S_{t+1})|S_t]$.

The final optimal value based on (6.11) is the maximum value between case (a) and (b) and obviously can be written by Equation (6.10). Equation (6.10) provides the insight on how to make an optimal decision for the paywall decision problem. At each state we can make the decision by comparing the optimum value of the state with the expected optimum value of the next states. Note that the value of a next state is computed recursively. One common approach to solve the Equation (6.10) is *value* (or similarly *policy*) *iteration* [74], which initializes the optimum value of each state randomly or using a guess and updates the value iteratively using the immediate contribution and the expected value of the future states according to the value function until convergence. However, this type of methods only applies to problems whose states are discrete and enumerable. The state space in our model is not discrete, so the iteration over the whole space is not possible.

An approach that avoids the iteration over the whole state space is to use the *value function approximation* [61, 76], in which the value function $V(S)$ is estimated explicitly by representing states with features and learning a model, e.g., linear combinations of features and neural networks, using, e.g., the gradient descent method. While feature

extraction from our states can be done, its process is not trivial if we would like use features from the articles involved in a state. In addition, to obtain the target value (e.g., V(S)) for the gradient descent training, an immediate reward is needed as part of the target value. However, in our problem, the reward function is very sparse. Moreover, our environment changes over time where new articles arrive frequently which may cause the state transition function and thus the value function change over time. Re-learning or updating the learned value function online may not be feasible.

Alternatively, it is possible to design a policy function directly by solving an approximation of the problem over a horizon. This class of techniques is called *lookahead policy* in approximate dynamic programming, and is distinguished from the value function approximation in that it uses random samples to simulate the future and make decisions directly based on the simulated future online without explicitly learning a value function.

6.3.3 Lookahead Paywall Policy

In this section, we propose a solution that uses the *lookahead* technique to estimate the decision function directly. In our solution, we replace the expectation in Equation (6.10) with an estimation. In particular, we make the following approximations when formulating the model: 1) we uses a short horizon H to limit the number of future time steps to look into, and 2) instead of using the full set of possible outcomes, we use Monte Carlo sampling to select a subset of outcomes starting at time t. Moreover, we use the

two-stage approximation for decision making. That is, we assume that we are in a known state S_t at time t; the second stage starts at time $t+1$, where we have different sample paths (i.e., realizations) of the future states from time $t+1$ to $t+H$. Let ω be a sample path of possible article requests from time $t+1$ to $t+H$ (which are stochastic), and $\widetilde{S}_{t'}(\omega)$ and $\widetilde{x}_{t'}(\omega)$ be the state and decision variables at time t' for the sample path ω accordingly (when we are in time t). Decisions are based on all stochastic variables over the horizon t to $t+H$ as follows:

$$X^{\pi^*}(S_t) = \arg \max_{\substack{x_t, \widetilde{x}_{t'}(\omega) \in \{0,1\}, \\ t+1 \le t' \le t+H, \forall \omega \in \Omega}} \{x_t C(S_t, x_t) + (1 - x_t) \sum_{\omega \in \Omega} p(\omega) \sum_{t'=t+1}^{t+H} \gamma^{t'-t} C(\widetilde{S}_{t'}(\omega), \widetilde{x}_{t'}(\omega))\} \tag{6.12}$$

where, Ω is the set of sample paths. In fact, at time t, we solve the problem optimally over horizon t to $t+H$ (using sampling) and find x_t, $\widetilde{x}_{t+1}(\omega)$ to $\widetilde{x}_{t+H}(\omega)$. However, we are not interested in values of $\widetilde{x}_{t+1}(\omega)$ to $\widetilde{x}_{t+H}(\omega)$. We are only interested in x_t, which is a decision at time t. After decision x_t is made, we advance through time and the process is repeated. Note that in Equation (6.12), x_t is common among all realizations. This procedure results in a simple and efficient method which can be applied to each article request in a session online.

Algorithm 1 shows the designed approach for the paywall problem based on Equation (6.12). This algorithm receives the current state S_t and navigational graph $\mathcal{G}$ and returns either 1 (i.e., show the paywall after this state) or 0 (continue without showing

Algorithm 2: Lookahead Paywall Policy Algorithm

Input: $S_t, \mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{W})$

Output: x_t

$P \leftarrow 0$

for $i = 1$ *to* $|\Omega|$ **do**

  $a_t \leftarrow$ Requested article of S_t

  $\omega \leftarrow []$

  $t' \leftarrow t + 1, Stop \leftarrow False$

  while $t' \leq t + H$ ***and*** $Stop = False$ **do**

    if $entropy(a_t) \leq 0.5$ **then**

      $\hat{a}_{t'} \sim Pr(\mathcal{N}_e(a_t, \mathcal{G}))$ ▷ $\mathcal{N}_e$ is the set of adjacent vertices for a_t

      $\omega \leftarrow [\omega\,,\ \hat{a}_{t'}]$

      $a_t \leftarrow \hat{a}_{t'}$

    else

      $Stop \leftarrow True$

    $t' \leftarrow t' + 1$

  $P_i \leftarrow 0$

  $\widetilde{S}_t \leftarrow S_t\ t' \leftarrow t + 1$

  while $t' \leq t + |\omega|$ **do**

    $\widetilde{S}_{t'} \leftarrow S^M(\widetilde{S}_{t'-1}, 0\,, \omega_{t'})$ ▷ $\omega_{t'}$ is the article at time t' in ω

    if $P_i < \gamma^{t'-t}\ C(\widetilde{S}_{t'}, 1)$ **then**

      $P_i \leftarrow \gamma^{t'-t}\ C(\widetilde{S}_{t'}, 1)$

    $t' \leftarrow t' + 1$

  $P \leftarrow P + P_i$

if $C(S_t, 1) > P/|\Omega|$ **then**

  return 1

else

  return 0

the paywall). Line 6-13 in the algorithm create a sample path (i.e., ω) based on the potential sessions encoded in $\mathcal{G}$. The only modification is that we do not proceed with sampling if the entropy of the current vertex is greater than 0.5. The entropy is calculated based on the probability of adjacent/next articles (determined based on weights of edges) in $\mathcal{G}$ as follows:

$$entropy(a_i) = -\sum_{a_j \in \mathcal{N}_e(a_i,\mathcal{G})} \frac{w_{ij}}{\sum_k w_{ik}} \log_b \frac{w_{ij}}{\sum_k w_{ik}} \tag{6.13}$$

where w_{ij} is the weight of edge between node a_i and a_j in the navigational graph, and $\mathcal{N}_e(a_i, \mathcal{G})$ is the set of neighbor vertices of a_i in $\mathcal{G}$. Note that we utilize the relative entropy (i.e., the base of $\log$ is b, where b is the number of adjacent vertices), so it is always between 0 and 1. The reason for using entropy is that if the article entropy is high, jumping to the next article is likely to introduce some noise (i.e., irrelevant articles). In particular, line 8 samples a next potential article $\hat{a}_{t'}$ from the set of neighbor vertices in $\mathcal{G}$ (i.e., $\mathcal{N}_e(a_t, \mathcal{G})$) based on the distribution $Pr(\mathcal{G}, \mathcal{N}_e(a_t, \mathcal{G}))$. In this distribution, for each vertex the probability of each adjacent vertex a_j is calculated by dividing the weight of the outgoing edge to a_j by the sum of all weights of outgoing edges of the vertex in navigational graph $\mathcal{G}$. Line 16-20 determines the best paywall time for different sample paths of articles by going over the states (line 17) in the path and finding the best contribution (line 18 and 19). Note that we need to calculate the second part of Equation (6.12) if we assume that x_t is 0, and in case that x_t equals 1, the first term in the equation (6.12) results in the whole contribution. Given that our sampling is unbiased,

we have $p(\omega) = 1/|\Omega|$. Finally, we can determine the best value for x_t by comparing the contributions of two cases ($x_t = 0$ or 1) accordingly (line 22).

Algorithm 3: Adaptive Paywall Algorithm

Input: User Requests, Navigation Graph $\mathcal{G}$

Output: Paywall decision

```
1  forall Requested article â do
2      if t = 0 then
3          Initialize the state S_0
4          Show â
5      else
6          S_t = S^M(S_{t-1}, x_{t-1}, â)
7          x_t ← LookaheadPaywallPolicy(S_t, G)
8          if x_t = 1 then
9              Show paywall and terminate the session
10         else
11             Show â
12     t ← t + 1
```

Algorithm 2 illustrates the overall procedure for the adaptive paywall approach. It receives a user request and makes a decision accordingly. Line 3 initializes the session by setting $\overline{u}_0$, $\overline{c}_0$, and a_0 of the first state (i.e., S_0) to $\phi(a_0)$, $\psi(a_0)$ and requested article

$\hat{a}$ accordingly. The algorithm always shows the first article (i.e., $x_0 = 0$). If the requested article $\hat{a}$ is not the first one, we first change the state to the next state (line 6), and then based on the result of the policy (Algorithm 1) we either show the article, or go to the paywall and terminate the session.

The above algorithms makes decisions online by simulating future states based on the navigation graph. New articles and changes in user navigation patterns can be incorporated easily by updating the graph. As long as the graph is updated, the algorithms can capture the new changes in the environment.

6.4 Empirical Study

We applied and evaluated the proposed method on a real dataset from The Globe and Mail, a major newspaper in Canada (See §1.4).

6.4.1 Utility and Cost Models

While the utility and cost of an article can be defined differently in our model, we use two intuitive utility and cost models in our experiments, namely, *Engagement to Cost (E/C)* and *Acquisition to Cost (A/C)*. For both models the cost (i.e., ψ) is defined as the article length (in terms of the number of 1KB pages) as the lengthy articles often need more efforts and resources to produce. The utilities for these models are defined as follows:

- The utility of article a in (E/C) model is defined as:

$$\phi(a) = \frac{Total\ dwell\ time\ by\ all\ users\ on\ a}{Total\ number\ of\ visits\ of\ a} \tag{6.14}$$

  where the unit of time is second.

- The utility of article a in (A/C) model is defined as:

$$\phi(a) = \frac{Total\ number\ of\ visits\ of\ a\ before\ subscription}{Number\ of\ visits\ of\ a} \tag{6.15}$$

  where the numerator is the number of subscribers who visited the article before the subscription (i.e., in the subscription session).

Note that the second measure has very small values as the number of subscribers is much smaller than non-subscribed users.

We use the utilities defined in Equation (6.14) and (6.15) to evaluate the proposed model as well as the baselines. However, in a real system, it is possible to use the outcomes of the model developed in chapter 4 (i.e., article dwell time prediction model) as the estimation of articles utilities. Alternatively, we can exploit the results of the model proposed in chapter 5 (i.e., time-aware subscription prediction model) to estimate the articles utilities based on Equation (6.15) (by treating predicted subscribers as the users who will subscribe). We consider this integration as the future work.

6.4.2 Baselines and Performance Measures

We compare the proposed Lookahead Paywall Policy (denoted as **LAP**) model with the following baselines:

- **Fixed Policy (FP):** This is a commonly-used paywall mechanism which allows a user to visit a maximum number, T, of articles during a session.

- **Average Threshold (AT):** This is a type of myopic policy [61] which defines the analytical decision function only based on the *current state* using a threshold. The decision function in this method is defined as follows:

$$x_t \triangleq X^{\pi}(S_t|\tau) \triangleq \begin{cases} 1 & \text{If } (\overline{u}_t/\overline{c}_t \leq \tau) \\ 0 & \text{Otherwise} \end{cases} \tag{6.16}$$

  where τ is set as the average ratio of the utility to cost, calculated based on the sessions in the training set.

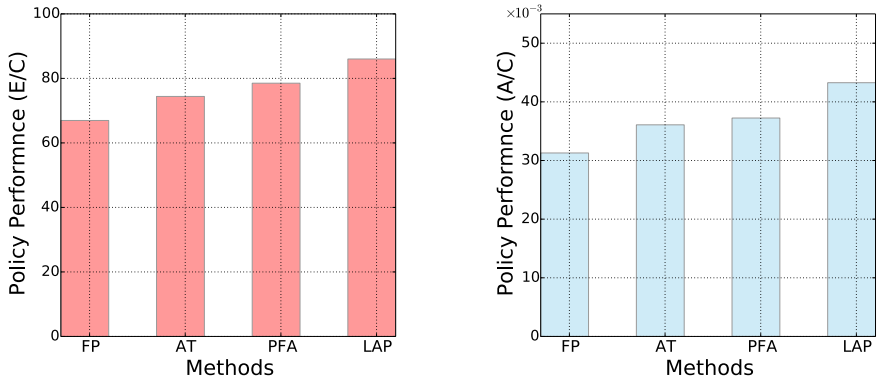
- **Policy Function Approximation (PFA):** This policy is based on Equation (6.16), but the parameter τ is optimized using sessions in the training set. Given (6.16), the policy search in the base optimization problem (6.8) is changed to a parameter search. Therefore, we can rewrite the optimization function in (6.8) as:

$$\max_{\tau} \; J(\tau) = \max_{\tau} \mathbf{E}\{\sum_{t=0}^{\infty} \; \gamma^t \; C(S_t, X^{\pi}(S_t|\tau))\} \tag{6.17}$$

Finding the best value of τ is a stochastic search problem. However, we cannot compute $J(\tau)$ in a compact form and $X^{\pi}(S_t|\tau)$ is not differentiable. Therefore, we use finite difference gradient [20, 75] which is a common approach for solving Equation (6.17).

In our experiments, we vary the value for T in the FP method. For our proposed LAP method. We set γ to 1 (which means that the future articles visits have the same value as the current one) and the horizon size H in Algorithm 1 to 4 unless indicated otherwise and sample size to 10 for all experiments.

We use the following *performance measures* when comparing the methods. (1) *Policy performance*, which is defined as the average ratio of aggregated utility of all articles in a session to the aggregated cost of all articles in the session over all the tested sessions. The utility of an article is defined using either the E/C or A/C model (according to Equation (6.14) or (6.15)). Thus, we have two policy performance measures: E/C based or A/C based. (2) *Policy performance at different percentages of articles delivered to users*, which measures the average utility-to-cost ratio of the sessions when a percentage of articles is presented to users. Obviously, the higher the ratio, the better the performance. (3) *The percentage of active sessions at each time point*, which is the percentage of sessions that have not received the paywall at each time point. A method A with more active sessions at time t than method B is better at t if it has at least the same or better policy performance than B at t. This is because as long as the trade-off between cost and



(a) E/C model

(b) A/C model

Figure 6.4: Average policy performance of different methods.

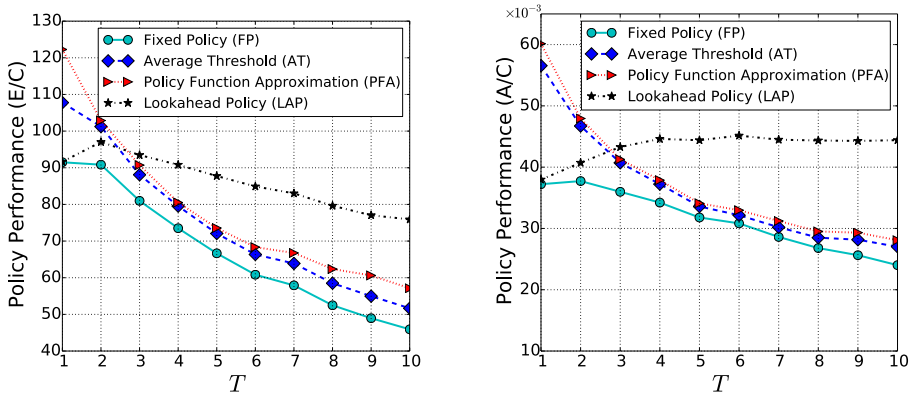
utility is fine, keeping the user active without presenting the paywall can further engage the user and also deliver advertisements.

6.4.3 Policy Performance Analysis

Figure 6.4 shows the overall average policy performance of the proposed model compared to the baselines for the engagement (i.e., E/C) and acquisition (i.e., A/C) utility models. For the FP method, the result is the average over T values ranging from 1 to 10. The Lookahead Policy model (LAP) shows 28.4 % and 38.3 % performance improvements over the traditional Fixed Policy (FP) for the E/C and A/C model respectively. It also outperforms the Average Threshold (AP) and Policy Function Approximation (PFA) on both E/C and A/C models.

In Figure 6.5, we compare the policies at different T values, where all policies are

allowed to show maximum T articles. Figure 6.5a illustrates the performance of the polices for the E/C model, which shows that the commonly-used FP policy has the lowest performance for all T values. It also shows that the performances of PFA and AT are better than the other methods at the beginning when T is small (i.e., less than 3). This is because they greedily terminate sessions if the articles requested so far do not look promising based on the threshold τ without looking at the future. On the other hand, if the LAP method is forced to stop early, it does not have the opportunity to make a better decision later on even if it finds one by looking ahead. As T increases, LAP performs better as it can present the paywall at a later time if it thinks it is better than current ones. Another observation is while the performances of all policies for the E/C model decline as we increase T (due to fact that users may visit more engaging articles at the start of sessions), the LAP policy outperforms the other policies by keeping alive the sessions with the promising future. Finally, between PFA and AT, PFA is marginally better than the AT method. Similar observations are found in Figure 6.5b, which illustrates the performances of the policies for the A/C model at different T values. The only difference is that as T increases, LAP's performance does not decline, which means that the LAP policy successfully navigates users to the articles that are useful for user acquisition (i.e., those which have been visited by the converted users and have good utility-to-cost ratios).



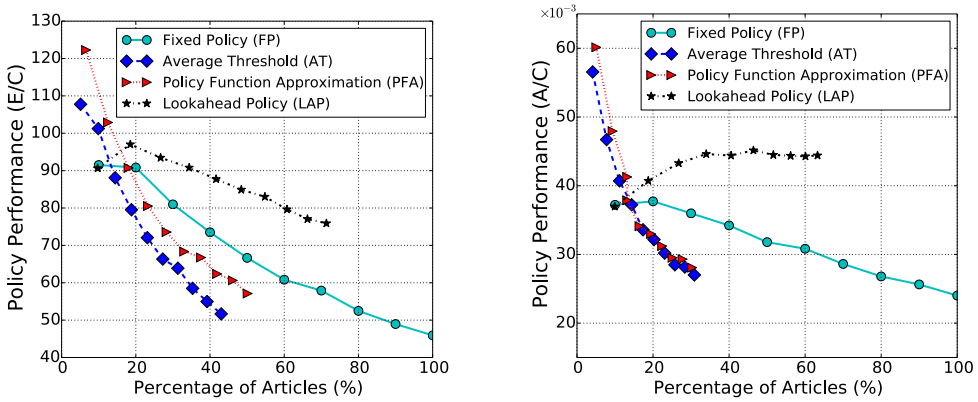
(a) E/C model

(b) A/C model

Figure 6.5: Policy performance for different models.

6.4.4 Performance vs. Delivery Percentage

In this section, we study the performance of each policy at different percentages of delivered articles (similar to precision at each recall level in information retrieval). To do this, for each policy and each test session s and at each time point t (from 1 to 10), we compute the percentage of delivered articles by time t as the number of articles delivered by the policy by time t divided by the total number of articles in session s, and also compute the ratio of the aggregated utility to the aggregated cost of all the delivered articles in the session by time t. In this way, for each policy and each session we obtain 10 ⟨delivered article percentage, policy performance⟩ pairs, one for each time point. We finally take an average of each pair over all the test sessions for the policy. Thus, each policy has 10 averaged pairs, one for each time point. For the FP method, we vary T from 1 to 10 to



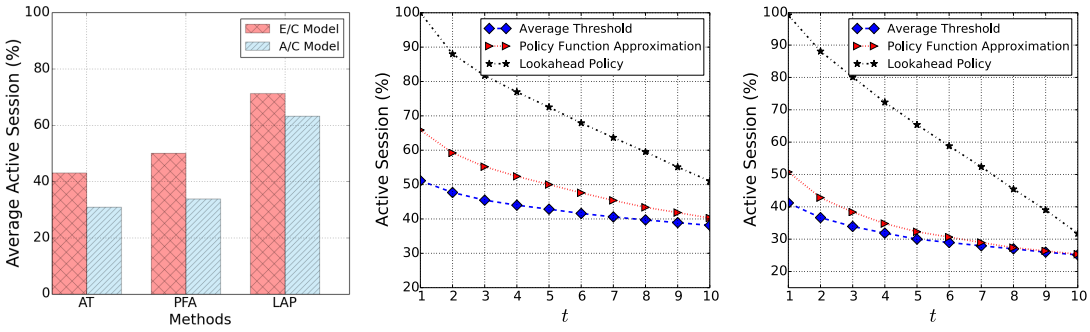
(a) E/C model

(b) A/C model

Figure 6.6: Policy performance vs. article delivery percentage.

collect the data for each time point.

Figure 6.6a shows the policy performance against the percentage of delivered articles for the E/C model. As can be observed, at the small article percentage levels (less than 15%), PFA and AT are better than LAP and FP. However, as we increase the percentage of articles, both LAP and FP outperform PFA and AT, with LAP being significantly better than all other polices consistently. In particular, LAP performs 31.0 % better than FP when 70 % of articles are shown to users. Similar results are found in Figure 6.6b, which shows the policy performance against the percentage of delivered articles for the A/C model. For example, the LAP performance is 48.0 % better than that of FP when 63 % of articles are shown to the users.



(a) Average active session (b) E/C model (c) A/C model

Figure 6.7: Active session for different utility to cost models.

6.4.5 The Effect of Policies on Users' Sessions

In this section, we investigate the effect of different policies on the number of active sessions (i.e., the percentage of sessions/users which have not received paywall before time point t). Intuitively, an online news agency prefers to have a high number of active readers all the time as long as the articles the reader reads have good utilities and their total cost is reasonable (i.e., the ratio of utility to cost is acceptable). Keeping a user active without presenting the paywall can further engage the user and allow more display of advertisements. Figure 6.7a illustrates the average percentage of active sessions (averaged over different time points in a session) for both E/C and A/C models. As can be seen, LAP has more active sessions on average compared to other policies. For example, the average number of active sessions of LAP is 42.5 % and 66.0 % more than PFA and AT methods respectively for the E/C model. Note that we did not show the FP method

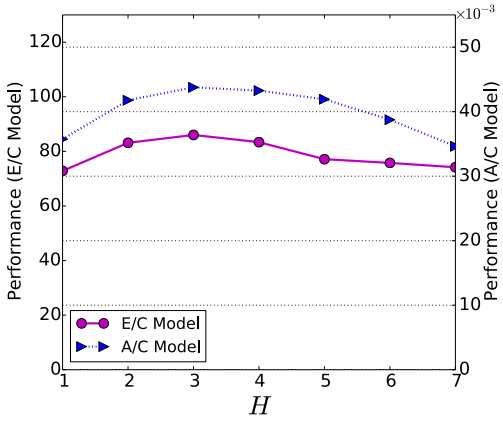


Figure 6.8: performance for different H.

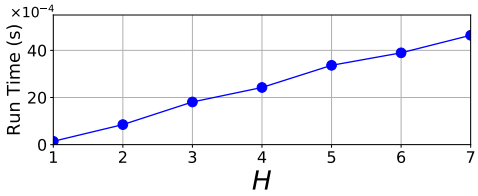


Figure 6.9: Average Runtime per request.

in this study because all sessions are active till time T in FP (which is assumed to be 10). Figure 6.7b and 6.7c illustrate the percentage of active sessions at different t values. PFA and AT terminate many sessions at the beginning by showing the paywall because close to half of the articles do not meet their threshold which is the average or close to average utility-to-cost ratio learned from the training data, while in LAP there are more active sessions at all the time points and the percentage of active sessions decreases more smoothly. Considering that LAP has better utility-to-cost ratios than other methods at almost all the time points (as shown in Figure 6.5), LAP is better as it has more active users without sacrificing the utility-to-cost ratio.

6.4.6 Sensitivity and Runtime Analysis

We analyze the performance sensitivity of LAP policy with different horizon sizes (i.e., H). We change the horizon size H and calculate the policy performance accordingly. Figure 6.8 shows the effect of horizon size H on the LAP performance for the E/C and A/C models. As can be observed, by incrementing H, the performance increases slightly for both E/C and A/C models, becomes stable in range of 2 to 4, and then starts to decline. This suggests that any value in range 2 to 4 would be a good choice for this parameter. Figure 6.9 shows the average run time per request for the LAP method. The experiments run on 2.2 GHz Intel Core i7 machine with macOS operating system. As can be seen, the response time per request is fast (a few milliseconds) and increases almost linearly with H.

6.5 Summary

We proposed an adaptive paywall mechanism for digital news media. The traditional paywall model allows a user to see a fixed number of articles and then directs them to the subscription page. We argued that this approach does not lead to more subscriptions and sacrifices other business objectives (e.g., increasing the user dwell time, the number of visits, etc.). We proposed a solution by formulating the paywall problem as a sequential decision problem that optimizes the ratio of the aggregated utility of the articles presented

to the user to their aggregated cost. We defined our problem and its components in an approximate dynamic programming paradigm, analyzed possible ways to solve it, and proposed a solution that uses a data-driven lookahead policy. We applied the proposed method to a real dataset from a national newspaper in Canada and showed the benefits and superiority of the method over the existing method and other baselines. To the best of our knowledge, the adaptive paywall problem has not been studied previously.

7 Conclusions and Future Work

Digital newspapers advertising revenues have declined remarkably in recent years. Cheap classified ads are popular and can easily target the right people. Moreover, ads blockers can effective filter out advertisements. Therefore, major newspapers still rely on user subscriptions as a major source of revenue. Thus, user engagement and acquisition are the most important issues for news companies. Although these are quite challenging tasks for all digital newspapers, there are great opportunities from data analytics perspective to address these problems. However, despite the impressive achievements of data mining/machine learning approaches in many domains, there is a significant gap between journalism and data mining/machine learning communities. Furthermore, while these demanding issues (e.g., needs for data-driven approaches in user acquisition) in journalism are clear, translating them into data mining and machine learning problems is not a trivial task. In this dissertation, we took the first steps towards bridging the gap between needs in digital journalism and capacities of data mining/machine learning approaches to address these demands.

7.1 Summary of Approaches and Contributions

In this thesis, we framed the main challenges of user acquisition and engagement in online newspapers into feasible problems in data mining/machine learning. First, we argued that user engagement plays a crucial role in building the sustainable relationships with users. As such, it is of paramount importance to understand the factors which make articles more engaging. Given this fact, we designed a data exploratory framework to provide insights into dwell time of articles, which is a measure of user engagement of an article. The summary of contributions is as follows:

- We considered the content-based engagement analysis, and in particular, investigated the article dwell time under the scope of factors such as topics, events, and emotions.

- The interesting dwell time patterns (i.e., negative and positive aging reading) were defined and the presence of main factors on these patterns was uncovered.

Moreover, inspired by the latest advances in neural network based representation learning approaches, we proposed a neural network architecture to predict dwell time of articles based on their contents. To best of our knowledge, this is the first study considering the content-based approaches to address the problem. The contributions are:

- We proposed an augmentation-based neural network architecture for the problem of dwell time engagement prediction.

- The proposed model can learn the high-order interactions of article contents as well as the low-order interactions of the article story main constituents (e.g., events) to produce the best prediction accuracy.

In line with our research goals, we designed a novel and an effective algorithmic tool to predict the users who are prone to subscription. The proposed model effectively utilized the engagement features and other historical data in digital news data collection platform to predict the potential subscribers. The summary of contributions is summarized as follows:

- We proposed a novel end-to-end framework which can effectively predict the user who are potential subscribers.
- The proposed model could effectively utilize the time in the prediction task and predict the subscription time.

Finally, we proposed an effective paywall model which balanced the benefit of showing article against presenting the paywall. To best of our knowledge, this is the first time that this problem was studied in data mining domain. The proposed model is a lookahead policy which can effectively exploit the navigation graph to find the optimal decision. The summary of contribution is as follows:

- We formulated the paywall problem as a sequential stochastic decision process, and proposed an effective solution which could effectively make a balance between

the benefit and cost of showing an article.

- We evaluated the proposed methods based on different notions of utility and showed that it effectively could be used to achieve different business objectives.

Last but mot least, the outcomes of this study have been published in the top tier conferences proceedings in the field of data mining such as the proceedings of the 24'th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [18], and proceedings of the 2017 SIAM International Conference on Data Mining [19], and are under review in 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2019), and the top journal in the field such as IEEE Transactions on knowledge and data engineering (TKDE).

7.2 Future Work

As this work is among the first steps towards the acquisition and engagement in digital news media, there are a lot of interesting directions for the future work.

7.2.1 Integrating the Proposed Models into One System

The notions of utility and cost in the proposed paywall model are general. However, it is possible to integrate the models developed in chapter 4 (article dwell time prediction model) or chapter 5 (time-aware subscription prediction model) into the proposed pay-

wall approach. The article dwell time prediction model can serve as a utility model since it can estimate articles dwell times. Moreover, the Time-aware Subscription Prediction (TASP) model can be used to estimate the number of articles which are visited by those who will subscribe. Therefore, it can be used as the utility model in the proposed paywall approach as well. One interesting direction is to integrate such predictive utility models into the proposed paywall model.

7.2.2 Online Evaluation of the Paywall Approach

The current evaluation of the proposed paywall model was done in an off-line setting to provide a proof of concept. Before the system can be deployed, an online evaluation in a real-world environment could be conducted for further assessment.

7.2.3 Investigating Other Engagement Measures

In this work, we focused on the dwell time as the engagement measure and built the data exploratory and predictive models for it. It is possible to investigate the other engagement measures and build the similar models for them. For example, most of digital news media allow users to comment on the news article and discuss about it. The number of comments that an article receives is the strong indicator showing the level of engagement of the article. Therefore, understanding and predicting this signal are interesting research questions.

7.2.4 Investigating Different Utility Models

In our experiment, we used the article dwell time and the number of visits to the article (e.g., by those who convert to subscribers) in the browsing history as the utility of article. However, one can define the utility of articles in numerous ways based in different business objectives. Investigating different models of utility could be an interesting research direction. Moreover, most of time for new articles, such information does not exist. Therefore, working on how to predict the utility of an article based on its content and other information would be another direction for investigation.

7.2.5 Dynamic of Utility and Cost Over Time

The utility and cost of article may change over time. A such, investigating how to model the dynamics of utility and cost over time and incorporate them into the proposed paywall model would be another interesting problem.

Bibliography

[1] A. Agrawal, A. An, and M. Papagelis. Learning emotion-enriched word representations. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 950–961, 2018.

[2] A. Agrawal, R. Sahdev, H. Davoudi, F. Khonsari, A. An, and S. McGrath. Detecting the magnitude of events from news articles. In *Web Intelligence (WI), 2016 IEEE/WIC/ACM International Conference on*, pages 177–184. IEEE, 2016.

[3] W.-H. Au, K. C. Chan, and X. Yao. A novel evolutionary data mining algorithm with applications to churn prediction. *IEEE transactions on evolutionary computation*, 7(6):532–545, 2003.

[4] S. Baluja, R. Seth, D. Sivakumar, Y. Jing, J. Yagnik, S. Kumar, D. Ravichandran, and M. Aly. Video suggestion and discovery for youtube: taking random walks

through the view graph. In *Proceedings of the 17th international conference on World Wide Web*, pages 895–904. ACM, 2008.

[5] H. Becker, A. Broder, E. Gabrilovich, V. Josifovski, and B. Pang. What happens after an ad click?: quantifying the impact of landing pages in web advertising. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 57–66. ACM, 2009.

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

[7] M. Broadie, P. Glasserman, and Z. Ha. Pricing american options by simulation using a stochastic mesh with optimized weights. In *Probabilistic Constrained Optimization*, pages 26–44. Springer, 2000.

[8] J. Cagé. *Saving the media : capitalism, crowdfunding, and democracy*. The Belknap Press of Harvard University Press, Cambridge, Massachusetts, 2016.

[9] H. Cai, K. Ren, W. Zhang, K. Malialis, J. Wang, Y. Yu, and D. Guo. Real-time bidding by reinforcement learning in display advertising. In *Proceedings of WSDM'17*, pages 661–670. ACM, 2017.

[10] S. Chakraborty, A. Venkataraman, S. Jagabathula, and L. Subramanian. Predicting socio-economic indicators using news events. In *Proceedings of the 22nd ACM*

SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 1455–1464. ACM, 2016.

[11] B. P. Chamberlain, A. Cardoso, C. H. Liu, R. Pagliari, and M. P. Deisenroth. Customer lifetime value prediction using embeddings. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1753–1762. ACM, 2017.

[12] H.-T. Cheng, L. Koc, J. Harmsen, T. Shaked, T. Chandra, H. Aradhye, G. Anderson, G. Corrado, W. Chai, M. Ispir, et al. Wide & deep learning for recommender systems. In *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*, pages 7–10. ACM, 2016.

[13] F. Chollet et al. Keras. `https://keras.io`, 2015.

[14] H. Chyi. Willingness to pay for online news: An empirical study on the viability of the subscription model. *Journal of Media Economics*, 18(2):131–142, 2005.

[15] M. Claypool, P. Le, M. Wased, and D. Brown. Implicit interest indicators. In *Proceedings of the 6th international conference on Intelligent user interfaces*, pages 33–40. ACM, 2001.

[16] A. C. Cohen. Maximum likelihood estimation in the weibull distribution based on complete and on censored samples. *Technometrics*, 7(4):579–588, 1965.

[17] H. Davoudi and A. An. Ontology-based topic labeling and quality prediction. In *International Symposium on Methodologies for Intelligent Systems*, pages 171–179. Springer, 2015.

[18] H. Davoudi, A. An, M. Zihayat, and G. Edall. Adaptive paywall mechanism for digital news media. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, pages 205–214, New York, NY, USA, 2018. ACM.

[19] H. Davoudi, M. Zihayat, and A. An. Time-aware subscription prediction model for user acquisition in digital news media. In *Proceedings of the 2017 SIAM International Conference on Data Mining*, pages 135–143. SIAM, 2017.

[20] P. Deisenroth, G. Neumann, J. Peters, et al. A survey on policy search for robotics. *Foundations and Trends in Robotics*, 2(1–2):1–142, 2013.

[21] D. Dheeru and E. Karra Taniskidou. UCI machine learning repository, 2017.

[22] G. Dupret and M. Lalmas. Absence time and user engagement: evaluating ranking functions. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 173–182. ACM, 2013.

[23] P. Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.

[24] N. B. Ellison, C. Steinfield, and C. Lampe. The benefits of facebook âĂIJfriends:âĂİ social capital and college studentsâĂŹ use of online social network sites. *Journal of Computer-Mediated Communication*, 12(4):1143–1168, 2007.

[25] R. Fletcher and R. K. Nielsen. Paying for online news: A comparative analysis of six countries. *Digital Journalism*, 5(9):1173–1191, 2017.

[26] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White. Evaluating implicit measures to improve web search. *ACM Transactions on Information Systems (TOIS)*, 23(2):147–168, 2005.

[27] M. Goyanes. An empirical study of factors that influence the willingness to pay for online news. *Journalism Practice*, 8(6):742–757, 2014.

[28] H. Guo, R. Tang, Y. Ye, Z. Li, and X. He. Deepfm: a factorization-machine based neural network for ctr prediction. *arXiv preprint arXiv:1703.04247*, 2017.

[29] S. Ioannidis, Y. Jiang, S. Amizadeh, and N. Laptev. Parallel news-article traffic forecasting with admm. In *SIGKDD Workshop on Mining and Learning from Time Series*. ACM, 2016.

[30] S. Jackson. *Cult of Analytics: Driving online marketing strategies using web analytics*. Routledge, 2009.

[31] C. Jennett, A. L. Cox, P. Cairns, S. Dhoparee, A. Epps, T. Tijs, and A. Walton. Measuring and defining the experience of immersion in games. *International journal of human-computer studies*, 66(9):641–661, 2008.

[32] J. H. Kim, A. Mantrach, A. Jaimes, and A. Oh. How to compete online for news audience: Modeling words that attract clicks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1645–1654. ACM, 2016.

[33] J. H. Kim, A. Mantrach, A. Jaimes, and A. Oh. How to compete online for news audience: Modeling words that attract clicks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery in data mining*. ACM, 2016.

[34] Y. Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, 2014.

[35] Y. Kim, A. Hassan, R. W. White, and I. Zitouni. Modeling dwell time to predict click-level satisfaction. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 193–202. ACM, 2014.

[36] R. Kohavi, A. Deng, B. Frasca, R. Longbotham, T. Walker, and Y. Xu. Trustworthy online controlled experiments: Five puzzling outcomes explained. In *Proceedings*

of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 786–794. ACM, 2012.

[37] H.-P. Kriegel, P. Kröger, and A. Zimek. Outlier detection techniques. In *Tutorial at the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2010.

[38] J. Lafferty, A. McCallum, and F. C. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

[39] C.-D. Lai, D. Murthy, and M. Xie. Weibull distributions and their applications. In *Springer Handbook of Engineering Statistics*, pages 63–78. Springer, 2006.

[40] M. Lalmas, H. O'Brien, and E. Yom-Tov. Measuring user engagement. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 6(4):1–132, 2014.

[41] Q. Le and T. Mikolov. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196, 2014.

[42] J. Lehmann, C. Castillo, M. Lalmas, and R. Baeza-Yates. Story-focused reading in online news and its potential for user engagement. *Journal of the Association for Information Science and Technology*, 68(4):869–883, 2017.

[43] J. Lehmann, M. Lalmas, G. Dupret, and R. Baeza-Yates. Online multitasking and

user engagement. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 519–528. ACM, 2013.

[44] J. Lehmann, M. Lalmas, E. Yom-Tov, and G. Dupret. Models of user engagement. In *International Conference on User Modeling, Adaptation, and Personalization*, pages 164–175. Springer, 2012.

[45] C. Liu, R. W. White, and S. Dumais. Understanding web browsing behaviors through weibull analysis of dwell time. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 379–386. ACM, 2010.

[46] D. C. Liu and J. Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1-3):503–528, 1989.

[47] I. Lopatovska and I. Arapakis. Theories, methods and current research on emotions in library and information science, information retrieval and human–computer interaction. *Information Processing & Management*, 47(4):575–592, 2011.

[48] Z. Luo, M. Osborne, J. Tang, and T. Wang. Who will retweet me?: finding retweeters in twitter. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 869–872. ACM, 2013.

[49] A. F. Martins, M. A. Figueiredo, P. M. Aguiar, N. A. Smith, and E. P. Xing. An augmented lagrangian approach to constrained map inference. 2011.

[50] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[51] R. G. Miller Jr. *Survival analysis*, volume 66. John Wiley & Sons, 2011.

[52] S. M. Mohammad and P. D. Turney. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 26–34. Association for Computational Linguistics, 2010.

[53] S. M. Mohammad and P. D. Turney. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465, 2013.

[54] M. C. Mozer, R. Wolniewicz, D. B. Grimes, E. Johnson, and H. Kaushansky. Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry. *IEEE Transactions on neural networks*, 11(3):690–696, 2000.

[55] M. Myllylahti. Newspaper paywalls-the hype and the reality: A study of how paid

news content impacts on media corporation revenues. *Digital journalism*, 2(2):179–194, 2014.

[56] N. Newman, R. Fletcher, A. Kalogeropoulos, D. A. Levy, and R. K. Nielsen. Reuters institute digital news report 2017. 2017.

[57] N. Newman, D. A. Levy, and R. K. Nielsen. Reuters institute digital news report 2016. *Available at SSRN 2619576*, 2016.

[58] K. Ng and H. Liu. Customer retention via data mining. *Artificial Intelligence Review*, 14(6):569–590, 2000.

[59] E. W. Ngai, L. Xiu, and D. C. Chau. Application of data mining techniques in customer relationship management: A literature review and classification. *Expert systems with applications*, 36(2):2592–2602, 2009.

[60] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[61] W. B. Powell. *Approximate Dynamic Programming: Solving the curses of dimensionality*. John Wiley & Sons, 2011.

[62] C. Raffel and D. P. Ellis. Feed-forward networks with attention can solve some long-term memory problems. *arXiv preprint arXiv:1512.08756*, 2015.

[63] S. Rendle. Factorization machines. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 995–1000. IEEE, 2010.

[64] F. Ricci, L. Rokach, and B. Shapira. *Introduction to recommender systems handbook*. Springer, 2011.

[65] M. Richardson, E. Dominowska, and R. Ragno. Predicting clicks: estimating the click-through rate for new ads. In *Proceedings of the 16th international conference on World Wide Web*, pages 521–530. ACM, 2007.

[66] M. Röder, A. Both, and A. Hinneburg. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408. ACM, 2015.

[67] S. Rosset, E. Neumann, U. Eick, and N. Vatnik. Customer lifetime value models for decision support. *Data mining and knowledge discovery*, 7(3):321–339, 2003.

[68] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA, 1986.

[69] G. Shani, D. Heckerman, and R. I. Brafman. An mdp-based recommender system. *Journal of Machine Learning Research*, 6:1265–1295, 2005.

[70] E. Shearer and J. Gottfried. News use across social media platforms 2017. *Pew Research Center, Journalism and Media*, 2017.

[71] B. V. Srinivasan, A. Natarajan, R. Sinha, V. Gupta, S. Revankar, and B. Ravindran. Will your facebook post be engaging? In *Proceedings of the 1st workshop on User engagement optimization*, pages 25–28. ACM, 2013.

[72] M. Steyvers and T. Griffiths. Probabilistic topic models. *Handbook of latent semantic analysis*, 427(7):424–440.

[73] R. Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3(1):9–44, 1988.

[74] R. Sutton and A. Barto. Reinforcement learning: An introduction. 2011.

[75] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pages 1057–1063, 2000.

[76] G. Taylor and R. Parr. Kernelized value function approximation for reinforcement learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1017–1024. ACM, 2009.

[77] C. Walker, S. Strassel, J. Medero, and K. Maeda. Ace 2005 multilingual training corpus. *Linguistic Data Consortium, Philadelphia*, 57, 2006.

[78] C. J. Watkins and P. Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.

[79] A. Williams. Paying for digital news: The rapid adoption and current landscape of digital subscriptions at us newspapers. *American Press Inst.*, 2016.

[80] S. J. Wright. Coordinate descent algorithms. *Mathematical Programming*, 151(1):3–34, 2015.

[81] L. Xiang, Q. Yuan, S. Zhao, L. Chen, X. Zhang, Q. Yang, and J. Sun. Temporal recommendation on graphs via long-and short-term preference fusion. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 723–732. ACM, 2010.

[82] B. Yang and T. Mitchell. Joint extraction of events and entities within a document context. *arXiv preprint arXiv:1609.03632*, 2016.

[83] X. Yi, L. Hong, E. Zhong, N. N. Liu, and S. Rajan. Beyond clicks: dwell time for personalization. In *Proceedings of the 8th ACM Conference on Recommender systems*, pages 113–120. ACM, 2014.

[84] E. Yom-Tov, M. Lalmas, R. Baeza-Yates, G. Dupret, J. Lehmann, and P. Donmez. Measuring inter-site engagement. In *Big Data, 2013 IEEE International Conference on*, pages 228–236. IEEE, 2013.

[85] Q. Zhao, M. A. Erdogdu, H. Y. He, A. Rajaraman, and J. Leskovec. Seismic: A

self-exciting point process model for predicting tweet popularity. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1513–1522. ACM, 2015.