Universiteit Leiden

ICT in Business and the Public Sector

What to do when recommendation systems become gatekeepers? The impact on news consumption

Name: Wanyi Wu
Student-no: s2573458

Date: 21/10/2021

1st supervisor: Dr. A.H. Zohrehvand
2nd supervisor: Dr. S.N. Giest
3rd supervisor: Prof.dr.ir. J.M.W. Visser

MASTER'S THESIS

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

The rapid growth of the Internet in recent years has led to the widespread use of news aggregators. News aggregators often use recommendation systems to increase news consumption for economic benefits, but at the same time, these recommendation systems also act as digital gatekeepers in the news distribution process, with an impact on society reflected in the diversity of the articles they recommend.

This study aims to find out the effects of recommendation systems on the news consumption, and to investigate the relationship between the diversity of user news consumption and the amount of user news consumption. In order to achieve the target, this thesis will first study different literatures, and generate two hypotheses based on these theories. Hypothesis 1 assumes that the user's exposure to diverse channels has a positive influence on the amount of his or her online news consumption. Hypothesis 2 assumes that 1): users who have large online news consumption are likely to have large online news consumption in their next sessions; 2) users who are exposed to more diverse topics are likely to have large online news consumption in their next sessions; 3) users who have long intervals between two sessions are likely to have large online news consumption in their next sessions. Then I collate user data from a database and analyse the integrated data using statistical tools to empirically test the hypotheses. The statistical results support hypothesis 1 and hypothesis 2.1, but are against hypothesis 2.2 and 2.3. Finally, based on the statistical results, this study provides some practical suggestions for the design of news recommendation systems, including developing algorithms to automatically remove highly similar articles, combining human editing and algorithms together, and more interdisciplinary research across computer, social, and legal sciences to propose more policies and regulations.

Keywords

News aggregators; Recommendation systems; News consumption; Gatekeeping; Media diversity

1 Introduction

1.1 Problem statement

The fast expansion of the Internet in recent years has resulted in widespread usage of online news. Online news platforms are easily accessible and user-friendly, overcoming the restrictions of traditional media. For news aggregators, the amount of clicks on their sites, in other words, the news consumption, translates into money. As a result, news aggregators employ a range of approaches to boost the amount of people who click on their materials. The most commonly used strategy is to employ recommendation systems that attempt to predict users' preferred content in order to increase traffic to their online news sites.

In order to give convincing suggestions to consumers, recommendation systems utilize algorithms to analyze previous user data and predict user behaviors. It's proven that recommendation systems are associated with better performance in online news areas, such as increasing user satisfaction, driving traffic, and increasing news consumption (Greenstein‐Messica et al., 2017; De et al., 2010). However, news differs from other recommended products due to its "public good nature", thus when designing a news recommendation system, we should not only aim to increase user consumption for financial gain, but also consider the impact on society (Claussen et al., 2019). Under this situation, media diversity is an important factor to consider because it can influence not only how people think and social trends, but also how users consume news. Previous research has also criticised inappropriate recommendation systems which negatively impact media diversity of the digital news ecosystem, such as creating filter bubbles (Pariser, 2010). However, there is little evidence to suggest whether the design of recommendation systems actually affects media diversity, and not much research on the relationship between media diversity and users' news consumption.

Although many news aggregators are using recommender systems, it is difficult for designers to consider media diversity because it is hard to quantify and there are not many relevant regulations to use as a reference. The complexity comes not only from designing the algorithms, which are sets of computer-implementable instructions for solving a problem, but also from understanding the role of recommender systems in the news distribution process. In summary, how recommendation systems can be designed to take into account media diversity and positively influence user consumption seems to be a valuable research topic.

1.2 Research question and research objectives

The main goal of this study is to investigate the technical and social considerations when designing recommendation systems. This study examines whether recommendation systems lead to changes in user news consumption and empirically tests the link between the diversity of user news consumption and the amount of user news consumption. In addition, based on the results of the study, I make some practical suggestions for the design of recommendation systems of news aggregators.

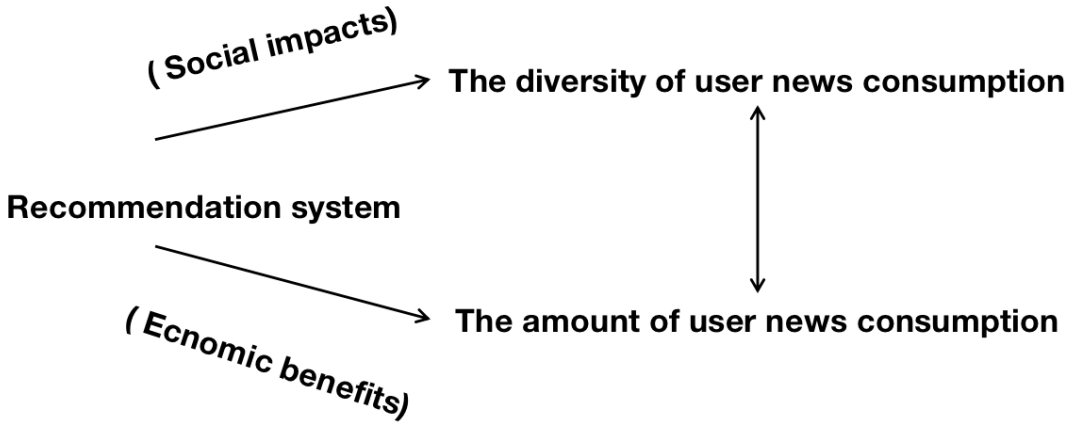

Figure 1: Visualization of the variables

The primary research question of this thesis is:

"How to design a recommendation system with better performance on user news consumption?"

The sub questions of this thesis are:

1. *What are the technical considerations when designing news recommendation systems?*
2. *What are the social considerations when designing news recommendation systems?*

In particular, the conclusions of the above research questions will be obtained by achieving the following research objectives.

- A literature review of the recommendation system, its technical principles, economic benefits and social impacts.

- To identify, through the analysis and study of databases, the factors that can be used in the assessment of user news consumption.
- To investigate the relationship of the diversity of user news consumption and the amount of user news consumption.
- Practical suggestions for designing news recommendation systems.

1.3 Thesis structure

The structure of this thesis is as follows. The first part outlines the goals and objectives of the study. The second part introduces the background of the study. The third part examines relevant research literature critically and formulates two hypotheses based on the findings. The fourth part presents the general research approach of this thesis. The fifth part presents the data and results of the study as well as a critical analysis. The sixth part contains a discussion of the findings as well as further summary conclusions.

2 Background

2.1 Recommendation system

2.1.1 What is recommendation system

Recommendation systems predict users' attitude towards an item, so as to deliver compelling suggestions to users (Schafer et al., 2001). Companies often apply recommendation systems to provide better user service, for instance, 70% of watch time on Youtube is recommended by its algorithms (Solsman, 2018). Because recommendation systems gather and analyze user data in order to create appropriate recommendations for users, they are always linked to improved online company performance, and benefit both service providers and users. As a result, recommendation systems are broadly used in many different areas, such as online video platforms, news aggregators, and food delivery. Especially for the companies offering streaming services, such as Netflix, Amazon Prime, and Disney+, the business model and success are based on their recommendation systems.

There are two broad types of recommendation systems algorithms: collaborative filtering methods and content based methods. Content-based recommendation systems discover the users' likes and dislikes based on history data, and recommend items which match with the user's preferences. Therefore, content-based recommendation systems always suggest similar items based on the users previous preference. By contrast, collaborative

filtering recommendation systems are completely dependent on past interactions between users and items. In addition, the basic premise of this system is that past interactions are sufficient for discovering similar users or items, as well as making predictions.

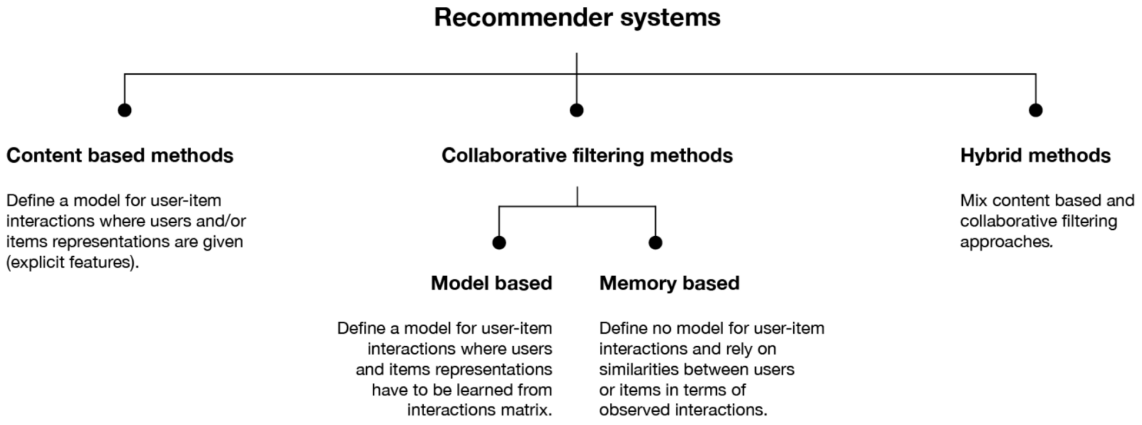

Figure 2: Summary of the different types of recommender systems algorithms (Rocca, 2019)

2.1.2 Collaborative filtering methods

Goldberg et al. (1992) suggest that collaborative filtering is associated with the relationship among multiple users and items. In their Tapestry experimental mail system, users are encouraged to make comments on documents, and these comments can subsequently be applied to filtering. The basic premise of collaborative filtering is that if two users rate or act on a certain amount of items similarly, they will have similar behaviors towards other items as well.

Collaborative filtering may be defined as a method of filtering items that a user might have preference for based on the responses of similar users. There are two types of techniques which are always used in collaborative filtering. These are: memory-based and method-based.

1) Memory-based collaborative filtering

In collaborative filtering systems which use memory-based algorithms, statistical methods are applied to the complete user rating dataset for calculating the similarity between users or items, resulting in scientific predictions (Su & Khoshgoftaar, 2009). If the user rating data is used for calculating the similarity between users, the method is called user-based collaborative filtering, and if it's used for calculating the similarity between items, the method is called item-based collaborative filtering. Memory-based collaborative filtering systems are widely used in business because they are easy to assemble, and extremely effective. For instance, Amazon.com, which is famous for its precise recommendations,

actually developed item-based collaborative filtering (Linden et al., 2003). By providing every user with a customized shopping experience, users spend less time and effort on searching, therefore, user friendliness is enhanced and customer loyalty is built up in commercial systems. In addition, companies can generate more sales and utilize personalized recommendations as an outstanding marketing tool (Ansar et al., 2000).

2) Model-based collaborative filtering

However, approaches of memory-based algorithms carry with them various limitations. For example, one major drawback of this approach is data sparsity (Acilar & Arslan, 2009). As the number of items increases, the number of common rated items decreases, which leads to difficulty of calculating similarity. For this reason, models are created using data mining or machine learning algorithms for more accurate predictions. These models include Bayesian network (Breese et al., 2013), matrix factorization algorithms such as singular value decomposition (Bokde et al., 2015), clustering based algorithms such as k-nearest neighbors (Sarwar et al., 2001), etc.

2.1.3 Content-based methods

Aside from collaborative filtering, content-based methods are also essential in recommendation systems. As its name suggests, content-based methods produce recommendations by studying the textual content of data and discovering patterns in the data. Comparing user interests with the item features, the items that have the most features in common with the user's previous preferences and interests are the ones that are recommended to the user (Lops et al., 2011).

When compared with collaborative filtering, the amount of data is little in a content-based approach, therefore, this method is more scalable (Thorat et al., 2015). Furthermore, because this method does not rely on the relationship with other users, it can tailor recommendations for every user. Moreover, content-based approach is more transparent, because it can clearly explain how the features of items match with the user’s interests, while collaborative filtering systems are more like black boxes, as the only reason behind an recommended item is that it was appreciated by unknown users with similar likes (Lops et al., 2011). On the other hand, this method is heavily reliant on previously known user interests. As a result, it cannot explore users’ unknown interests, or make accurate recommendations to new users.

2.1.4 Advantages and disadvantages about recommendation systems

As stated before, Recommendation systems collect users data and automatically analyse this data to generate suitable recommendations for users. And there is much research on the impacts of recommendation systems.

Recommendation systems have been widely associated with better performance in online business. One of the benefits is to increase sales or conversions without increasing marketing efforts. In 2010, De et al. compared the effects of different techniques (search and recommendation systems) in online sales, and they argue that both directed and undirected search have a less positive impact on product sales than recommendation systems. In addition, recommendation systems lead to a reduction in user efforts by offering suitable options, therefore increase user satisfaction and loyalty. In e-commerce, recommendation systems can also help with designing personalized pricing or discounts correlating to the specific contextual situation of the consumer, which increases the sales of online retailers (Greenstein‐Messica et al., 2017).

On the other side, because recommendation systems influence what people consume and experience on the internet, it's critical to discover whether they're influenced by any potential sources of bias that might have effects on society. For example, in order to maximize user interaction, social media news feeds may accidentally push incorrect or strongly politicized content, which is popular among like-minded audiences. More broadly, specialized yet high-quality content may not be well-received. To solve this problem, the diversity of online news consumers may be utilized to access the quality of news. Films that appeal to a wide range of age groups or ethnic groups, for example, may be of higher quality than other films (Bhadani ,2021). Estimating the influence of user diversity more accurately might help us better understand algorithmic bias and aid to the development of more reliable recommendation systems.

2.2 News aggregator

Over the past years, there has been a dramatic increase in the field of news, and news occupies a prominent place in today's society, people need to read news every day to stay up to date on latest information. With the development of technology, there are so many softwares and websites that news can be sent to smart devices automatically or can be accessed with a few clicks, and people are bombarded with information. Then news aggregators are developed to solve the problem of information overload, which means that the decision maker feels perplexed when too much information is displayed (Gross, 1964).

Enabled by the new technology of hyperlinking, news aggregators attract web traffic and generate revenue by hosting collections of links to third-party material. Instead of original content, news aggregators always provide titles and brief descriptions of stories they link to (Dellarocas et al., 2013). Some original news creators consider aggregators to be substitutes for traditional news consumption, whereas aggregators argue that they help the original news publishers by facilitating news discovery. In 2021, Athey et al. compared the news consumption of a large number of Google News users with a control group of similar non-Google News users, and they found that the shutdown of news aggregators reduces both overall news consumption and page views on third-party publishers.

Therefore, they suggest that news aggregators can lower search costs and increase the ability of small news creators to reach consumers. Furthermore, aggregators can link to news items as soon as they are published on third-party websites, so news aggregators always move fast to take the advantage of breaking news. Hamborg et al. (2020) identifies five typical phases involved in news aggregation, these are: crawling articles from websites; extracting articles from raw data; grouping articles on the same topic; generating summaries of related articles; visualization. Recommendation system is always used in the last phase to decide the priority of articles.

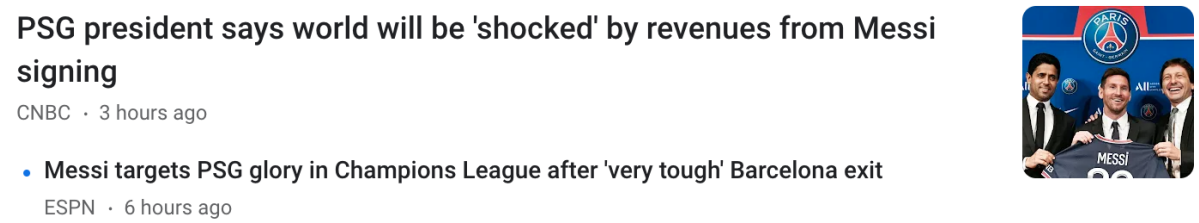

Figure 3: Example of hyberlinking in news aggregator (source: Google News, Aug 11th, 2021)

The explosion of online news necessitates the creation of news aggregator systems for identifying and filtering interesting information. However, recommending news faces unique obstacles when compared to recommending other things: users get bored when receiving too many similar stories, and adequate user information is frequently lacking (Garcin et al., 2013). In addition, news recommendation is challenging because the item of news is very special: there is limited data available about a very fresh news to generate recommendations (Garcin et al., 2012). Claussen et al. (2019) argue that the news is unlike any other product owing to its "public good nature". They believe that the algorithm is specifically calculated

based on past individual level data, however, personal preferences may conflict with “socially optimal reading behavior”, which results in chaos. They also ran a field experiment to compare human editors with recommendation algorithms, and found out that in terms of user engagement, automated recommendations can surpass human editors, under the circumstance of efficient training data. At last, they suggest that the best method of news aggregator appears to be to use a combination of algorithms and human editors.

Another interesting topic to consider is the privacy problem when collecting user data. In 2019, Claussen et al. propose that the decrease in data retention has no bearing on search engine performance. The findings also suggest that legislation proposed by various institutions, including the European Commission, on the amount of personal data retained by firms may not significantly erode firms' competitive advantage because adverse effects on consumer engagement and, as a result, recommendation system performance would be limited.

Considering these challenges in building news recommendation systems, Garcin et al. (2012) compare three approaches for customized news recommendation when users are anonymous and only current visit data can be utilized to produce recommendations, in other words, there is no information available regarding a user's previous behavior. They propose that collaborative filtering techniques provide much better performance than content-based techniques and hybrid techniques in this situation. Furthermore, they explain this by the fact that users easily get bored about the same topic: a user reads on average 7 distinct topics in 10 news.

Google News generates personalized recommendations combining three methods: collaborative filtering, as well as probabilistic latent semantic indexing and covisitation counts (Das et al., 2007). Das et al. believe that in areas like news, a user's interest in an item is not always characterized by what is present in the content. Furthermore, they have an ambitious target to create a system that could be used in other domains like photos and music, where analyzing the content is difficult, thus they built a content-agnostic system. In 2010, Liu et al. expanded on the Google News study by examining user click behavior to create accurate user profiles. They propose a Bayesian model for recommending news based on the user's interests and a group of users' news trends. To generate personalized recommendations, they combine this approach with that of Das et al.

In summary, news aggregators usually employ a recommendation system to determine the importance of items and tackle the problem of information overload. However, news is different from other products because of its "public good nature" and freshness, thus some

researchers suggest combining algorithms and human editors, while some researchers propose new models for recommending news.

3 Theoretical framework

3.1 Introduction to the theoretical framework

As mentioned before, the thesis aims to study the correlation between the diversity of user news consumption and the amount of user news consumption under the influence of recommendation systems, and to investigate technical and social considerations when designing recommendation systems. Due to the widespread use of technology in the news industry, recommendation systems play the role of digital gatekeepers in the news dissemination process, a role that emphasises the impact that recommendation systems can have on society, in addition to the economic benefits they bring. Recommendation systems were originally designed to generate revenue by increasing user news consumption, and their social impact is often expressed in the form of media diversity. Media diversity can be described by internal and external pluralism, which in this paper correspond to topic diversity and channel diversity. Finally, based on the theory of planned behavior and some other literature, the thesis makes some hypotheses about the correlation between diversity in user news consumption and the amount of user news consumption.

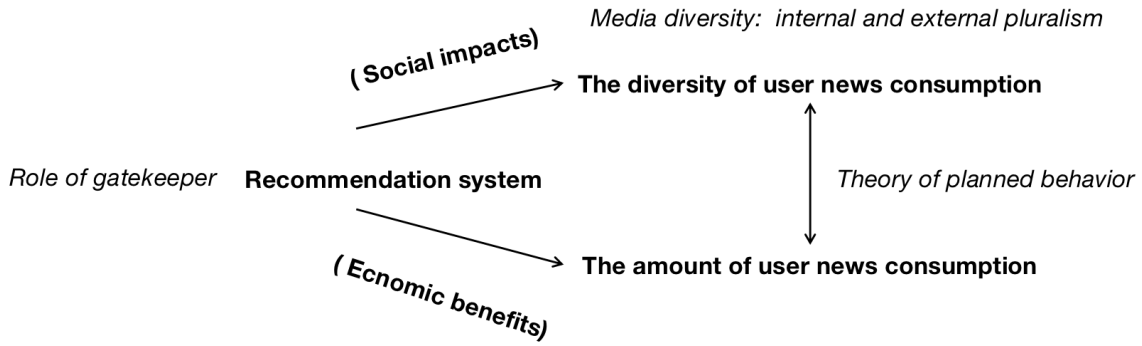

Figure 4: Theoretical framework

3.2 Gatekeeping theory

3.2.1 What is gatekeeping theory

Gatekeeping is the process of selecting, and filtering information, and gatekeepers decide about the content which would be seen by the audiences (Lewin, 1943). The gatekeeping theory is considered to be one of the most important and foundational theories in mass communication studies, because of the rapid development of mass-publication technologies such as newspapers, television, and the Internet.

Gatekeepers are high-level decision makers that manage the data flow across a whole social system, they discard some information and only allow specific information to get through to the audience. According to White (1950), the activity of news reporting is not reporting everything, but selecting from many news materials and further processing. From his point of view, gatekeeping is highly subjective, which depends on the gatekeeper's practical experience, personal attitudes, social influences, values or even bias. However, in 1965, Galtung and Ruge proposed that certain characteristics of original news events affect whether they will pass the gate and be communicated. In addition, Shoemaker and Vos (2009) proposed that gatekeeping is the “center of the media's role in modern public life”.

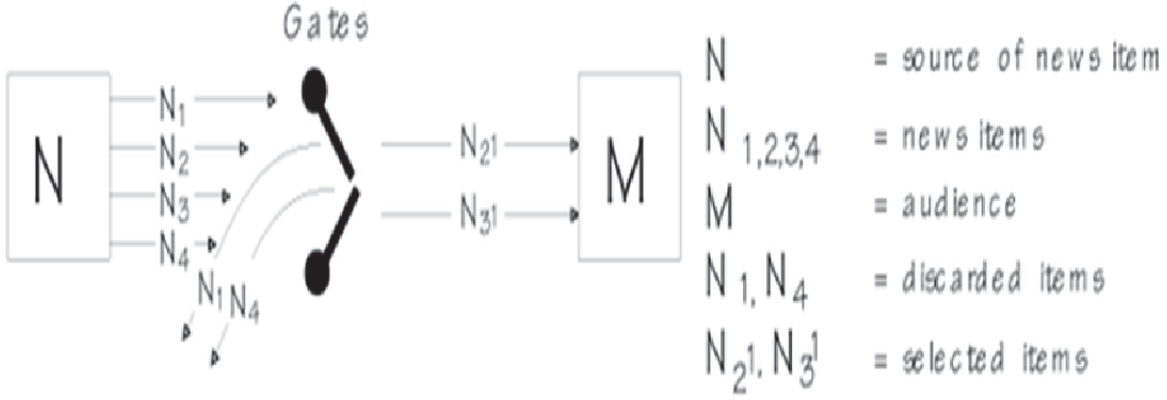

Figure 5: White’s model (source: McQuail & Windahl, 1981)

In the area of news media, some suggest that the journalists and editors play the role of gatekeeper. Each news channel receives numerous news items from around the world every day, and each channel has its own series of ethics, rules, and biases that the editor uses to decide whether the news items will be published or not. In some cases, the editor rejects news items due to external factors such as organization reputation and political impacts. Due to all these factors, different news organizations hold different views of what kind of information should be filtered. For instance, the exact same news story from different

channels, such as BBC, CBS or CNN, can be presented in a variety of ways, thus leaving varied impressions to the public. Based on how the news items are investigated and presented, the journalists and editors act as gatekeepers in the process of news publishing.

3.2.2 Digital gatekeeping

The Internet has changed how people consume news from several perspectives. Firstly, people have increasing control over the content they receive, because there are so many choices and they can easily get in touch with their interests. Secondly, the Internet provides a diverse range of sources for social events, so the audience is able to hear about different voices and develop new insights about these events. The extended access to the Internet has greatly increased the availability of media content over news and broadcasting, such as sharing on social media and content posted by users themselves, disrupting the traditional business model of news media. Also, the traditional way of gatekeeping has changed a lot.

For more than decades, journalists and editors, or news organizations are thought to be the only gatekeepers during the process of news communication. These traditional gatekeepers dominated how the truth is presented and what information reached society. For the past few years, other characters and platforms have gotten more involved with the matter of gatekeeping. For instance, the social media users actively publish their original contents and scramble to be the source of news. For another example, online news consumers like, hate or comment on the news stories to share their attitudes. Furthermore, news consumers are becoming more reliant on news aggregators which recommend particular news stories to them, because they need personalized recommendations among the large pools of news items.

There are also many other studies that discuss the impact of the internet on gatekeeping theory. For example, the opinion of network gatekeeping stresses the digital audience that creates and disseminates information. It refers to a broad range of institutions, including governments, search engine providers, and varied organizations and individuals. For another instance, the concept of gatewatchers emerged. Instead than generating and distributing information, gatewatchers make it public by pointing to the source. This method bypasses possible misinformation in the dissemination of information and gives other users effective access to the original source of the information. In order to become a gatewatcher, the recipient of information has to be more active in the process of consuming and producing data.

Past research has focused almost exclusively on human gatekeepers, such as editors and journalists, and some researchers have recently begun to focus on digital technology as gatekeepers (Tandoc, 2014; Welbers et al., 2018). Digital technologies in the field of news usually take the form of recommendation systems and search engines, which enhance the user experience through personalised services but also may lead to filter bubbles. For instance, Napoli (2014) argues that search engines and recommender systems provide content that is in fact closely related to the public interest. They, or their proprietary companies, are considered to be the managers of information flows on many digital platforms. The role of digital technologies in gatekeeping is growing and they have become an important part of social reality (Just and Latzer, 2016).

Due to the changes brought by the internet, traditional gatekeeping theory is no longer suitable for understanding online news consumption processes. Wallace(2017) developed a cohesive model for gatekeeping theory as he believes the gatekeeping methods are fragmented.

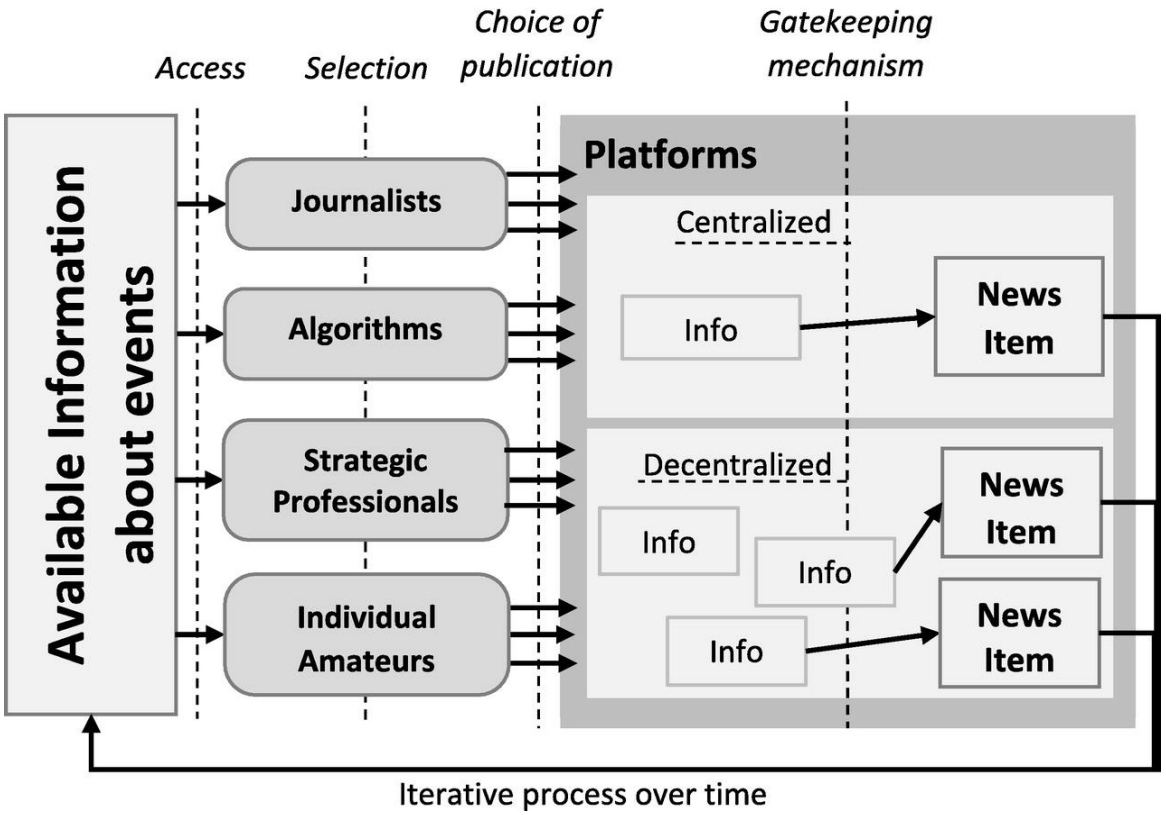

Figure 6: Digital gatekeeping as a news dissemination process (source: Wallace, 2017)

First of all, Wallace identified four kinds of gatekeepers: journalists, individual amateurs, strategic professionals and algorithms because they have different access, selection

preference, choices of publication. Furthermore, Wallace identified two kinds of platforms which gatekeepers operate on: centralized platforms and decentralized platforms. His model represents different gatekeeper archetypes and how they interact, however, the gatekeeping mechanism on platforms is not well explained.

3.2.3 gatekeeping theory in news recommendation systems

To focus on the application of algorithms to recommendation systems, this thesis develops a new model to explain the entire process of a news item from its generation to its reception by the news consumers, taking into account the previous discussion in 3.2.1 and 3.2.2.

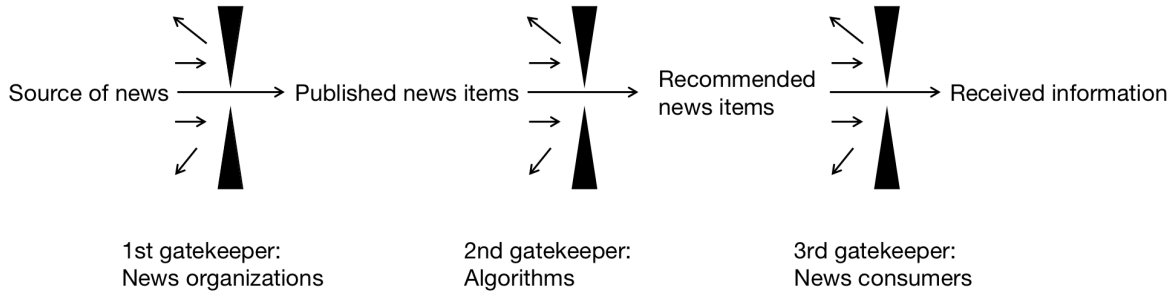

Figure 7: A new model of gatekeeping theory in news recommendation systems

As is shown in figure 7, this new model of gatekeeping theory is built in the context of news recommendation systems. In this model, there are three different kinds of gatekeepers in the whole process of online news consumption:

1. The news organizations which include journalists, editors, and other possible involved characters and factors. From an event to a published news story, the news organization plays the role of the traditional gatekeeper in this process.
2. The algorithms. Based on people's past reading preferences, the algorithm gives the most accurate recommendations possible from a wide range of news, which is the second gatekeeping.
3. The news consumers. The ideal algorithm should give results that exactly match the user's preferences. However, in reality, users' preferences are constantly changing, so they only choose to see the news they are interested in among the many recommendations. In this process, the news consumers themselves become the third gatekeeper.

Combining traditional gatekeeping theory with modern methods, the proposed gatekeeping model identified two gatekeeper archetypes and how they work in the context of news aggregators.

3.3 Theory of planned behavior

3.3.1 What is theory of planned behavior

The theory of planned behavior, which assumes that human behavior is the result of deliberate deliberation, helps to understand how people change their behavior patterns (Ajzen, 1991). The theory of planned behavior evolved from Fishbein and Ajzen's theory of reasoned action in 1975, and is recognised as a more complete model of explaining human behavior. Many empirical studies have also shown that the predictive power of the theory of planned behavior is indeed higher than that of the theory of reasoned action (Madden et al.,1992; Hansen et al., 2004).

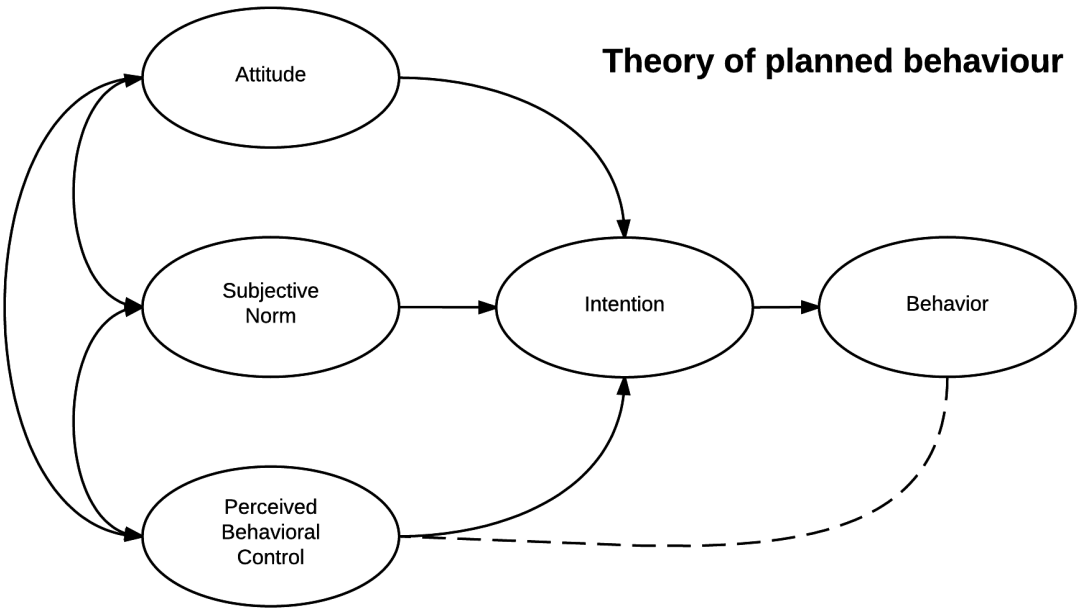

Figure 8: Theory of planned behavior (source: Ajzen, 1991)

As is shown in figure 8, there are four main factors in the theory of planned behavior:

1. Attitude refers to an individual's ongoing assessment of his or her liking or disliking of a particular object or idea, and can be used to predict his or her likely behavior. In other words, attitude is an individual's positive or negative evaluation of a particular

behavior. The more positive an individual's attitude towards a behavior, the higher his or her behavioral intentions will be.

2. Subjective norm refers to the social pressure that an individual feels when adopting a particular behavior, i.e. the pressure that the individual perceives from significant others or groups (e.g. parents, friends, colleagues, etc.) as to whether or not he should perform the particular behavior. The more positive the positive subjective norm, the more likely it is that the individual will be motivated to engage in the behavior.
3. Perceived behavioral control refers to the perception of the individual's ability to control the resources and opportunities needed to engage in a particular behavior. It includes non-motivational factors out of the individual's control, such as time, money, skills, opportunities, abilities, resources or policies. Therefore, even if an individual wants to engage in a particular behavior, he or she may not be able to do so because of a lack of resources. In addition, the limits of perceived behavioral control can be divided into self-efficacy, which refers to the perception of one's ability to perform the behavior, and external resources, which refers to the availability of resources and the degree of hindrance, both of which may influence the individual's decision to adopt a behavior.
4. Behavioral intention refers to an individual's propensity to engage in a particular behavior and the degree to which he or she wants to do so. In terms of measurement, it can be translated into questions such as whether or not an individual is willing to try hard or how much effort he or she is willing to put in, and this variable can be used to explain and predict an individual's actual behavioral performance.

In summary, the theory of planned behavior suggests that attitudes, subjective norms, and perceptions of behavioral control not only jointly determine an individual's behavioral intentions, but also interact with each other. In effect, behavioral intentions determine individual behavior, and behavioral intentions are determined by attitudes, subjective norms, and perceptions of behavioral control.

3.3.2 The theory of planned behavior in recommendation systems

To increase user satisfaction with the news recommendation system, I apply the theory of planned behavior to better understand user’s behavior while using a news aggregator. As Wang (2011) states, “behavioral intentions to use recommendation systems is defined as a person’s readiness to use the recommendation system to receive purchasing

advice". In the context of news aggregators, the user's intention is described as his or her readiness to use the news aggregator to receive reading advice.

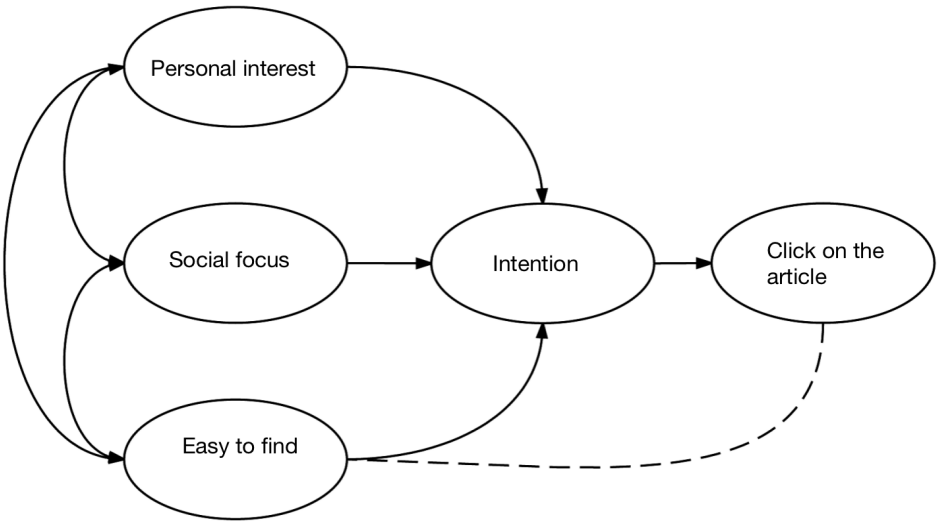

Figure 9: Applying the theory of planned behavior to news recommendation

Treating each user click on a recommended article as a positive action, within an ideal recommendation system users would click on every recommended article, in other words, they would find these reading suggestions extremely useful. The application of the theory of planned behaviors to news recommendations is presented in figure 9..

1. A user's attitude towards the behavior of clicking on the recommended articles is actually the user's attitude towards the recommended articles. In this case, the user's personal interest, in other words, the user's positive or negative evaluation of a particular article is positively correlated with the user's intention.
2. It's obvious that public news in the spotlight always gets more clicks. In this case, even if one is not interested in the recent social news, he or she may still click on and check what it is because people around are talking about it, and this is the presentation of the subject norm in the news area.
3. News aggregator solves the problem of information overload, and saves time and energy for users to find interesting content. On this occasion, resource-saving recommendation systems and user-friendly web design are central to promoting user clicks.

In summary, based on the analysis above, a well-designed news recommendation system should be able to predict user's interests, move fast to include social focus events, be resource-saving and user-friendly.

3.4 Media diversity

3.4.1 Media pluralism

Pluralism is a broad term which means diversity. The notion, however, covers a variety of characteristics and has been interpreted from numerous angles in accordance with different scenarios. In the area of media pluralism, there are two viewpoints: internal and external pluralism.

1) Internal pluralism refers to how media output reflects social and political diversity. In other words, the news stories should include various social groups, represent different political opinions, and analyze from different angles (Doyle, 2002). For example, UNESCO contributes to the public's access to a wider range of information, especially for women, because they are always underrepresented in media content, decision-making and media workforce (UNESCO, 2021; Macharia et al., 2015).
2) External pluralism refers to the news supplier diversity, such as the amount of news organizations, channels, media companies, individual websites, etc,. To make sure that the audience can easily access a variety of opinions and the truth of events, competition among various news organizations is deemed necessary.

3.4.2 Filter bubbles

Due to the huge volume of data online it is important to use algorithms, such as search engineers and recommendation systems to save people time and improve user experience. Nowadays, people obtain news via platforms such as Twitter or Google News. While visiting these websites, algorithms have picked some of the news that the users see. Algorithms made this choice based on data gathered by websites about the users' previous usage, as well as data the users willingly share with the websites (usually stated in data privacy agreement). Naturally, there is concern that this kind of prediction will encourage existing spending tendencies.

Pariser (2010) claims that these filtering algorithms are prejudiced and do not display materials that the user may not like. Pariser believes that algorithms generate filter bubbles that impact our society and have harmful repercussions unless we pay attention to the algorithms and take social responsibility when coding. And the issue grows much worse if everyone refuses to leave their own bubbles. For example, when everybody is sure that they are receiving the whole information on the current event but actually they are only seeing part of the truth, it becomes impossible for people to make fair judgements and not to favor a

certain kind of opinion. More seriously, filter bubbles lead to a refusal to examine competing ideas and negative facts, and make debates less meaningful.

On the other hand, some researchers suggest that focusing on filter bubbles can lead to misunderstandings of the mechanisms, as well as diverting us from somewhat more important issues. In 2020, Fletcher et al. found that people are not only increasing the diversity of news sources when obtaining news from social media, but also improving the balance among these different sources. However, there is a chance that this diversity is generating polarization in opinions and usage. This research is intriguing since, in some ways, it contradicts the filter bubble concept. In addition, Bruns (2019) also holds the view that the "filter bubble" concept should be examined critically, concluding that the emphasis on filter bubbles may be hindering us from facing underlying reasons of polarisation in politics and society.

The combination of consumer patterns, shifting economic models, and technological systems is causing a greater divide in the way individuals utilize news all over the world. As a result, while there is an extremely large amount of information available, it is also difficult for each individual to get the full picture. While there is an extremely large amount of information available, it is also difficult for each individual to get the full picture. Although there is some disagreement, the theory of filter bubbles does support the suggestion that ethical considerations should be taken into account when designing algorithms.

3.5 Hypothesis

3.5.1 Hypothesis 1

In this research, I use the user's exposure to diverse channels to describe media external pluralism. And the user's exposure to diverse channels may correlate with consumption on his or her online news consumption for two reasons.

First, users who are exposed to news from more diverse sources have more choices, thus they are likely to consume more news online. In accordance with the theory of planned behavior, a user's positive attitudes towards news sources can be used to predict that he is likely to visit these news sources in the future. As Flaxman et al.(2016) proposed, the most of online news consumption is composed of consumers merely visiting their favorite news providers. In this case, users who are exposed to more diverse news providers will have more

choices, and they are more likely to be exposed to their preferred channels, therefore their online news consumption may be more than in a less diverse environment.

Secondly, diverse news channels provide users with different views on the same event, thus users are able to understand events from multiple perspectives, rather than trust only one media outlet and see things from only one side. The exact same news event can be investigated and presented from multiple angles by different channels, thus leaving varied impressions to the public. At the same time, depending on the audience's learnings, they may have strong thoughts and beliefs about the authenticity of the news sources. And that's why external pluralism is deemed necessary to make sure that the audience can easily access a variety of opinions and the truth of events. People who are exposed to diverse channels are more eager to understand the different perspectives on social events, the user may read articles from multiple sources, thus having a high amount of online news consumption.

This leads to the hypothesis:

Hypothesis 1: The user's exposure to diverse channels has a positive influence on the amount of his or her online news consumption.

3.5.2 Hypothesis 2

1) the user's consumption amount of the current session

Considering the behavior of a user consuming online news in a day as a session, the user's consumption amount of the current session may correlate with his or her online news consumption amount of the next session for the following reason.

According to the planned behavior theory, the user's past behaviour can be used to predict future behaviour, thus the user's online news consumption amount of this session can be used to predict his or her consumption amount of the next session. Firstly, well-practised behaviours repeat in a continuous environment, as good responses automate these behaviours. Secondly, past behavior frequency, in turn, shows habit strength and has a direct impact on future performance (Ouellette, 1998). Under certain circumstances, past behaviour (together with attitudes and subjective norms) may influence intentions, and intentions, according to the theory of planned behavior, govern an individual's behaviour. For this reason, the user's consumption amount of this session can be positively correlated with the user's consumption amount of the next session.

This leads to the hypothesis:

Hypothesis 2.1: Users who have large online news consumption are likely to have large online news consumption in their next sessions.

2) the topic diversity of the current session

In this research, I use the user's exposure to diverse topics to describe media internal pluralism. The topic diversity of the current session may correlate with the user's news consumption amount of the next session for the following three reasons.

First, positive user attitudes towards a wide range of topics may lead to higher online news consumption. In agreement with the theory of planned behavior, attitudes can be used to predict the user's potential behavior, so it is deductible that if a user holds a positive attitude towards a topic, then this user is likely to click on posts covering this topic in the following session. A user who is exposed to more diverse topics will have more choices, and they are more likely to be exposed to topics they are interested in, therefore their online news consumption may be more than in a less diverse environment.

Secondly, users with less exposure to topics always continue to be active in a small area rather than being exposed to new things, and limited choices lead to limited news consumption. In accordance with the theory of filter bubbles, although there is unbelievable information accessible in the online news consumption area, individuals are increasingly exposed to a limited range of information because filtering algorithms do not display materials that users may not like. Instead of receiving novelty content, people exposed to less diverse topics tend to confirm their own particular preconceptions. In this case, they are not very eager to explore the topics they didn't know about before, but always stay on the same topics, therefore their choices are limited, and it's not surprising that they don't consume much news.

Thirdly, users exposed to a limited range of topics are prone to feeling conflicting values in the news, which can lead to avoidance of some news. Slaets et al. (2020) provided this explanation from the view of selective exposure. As confirmed by a survey of media users, they suppose that the diversity of information has the potential to affect the selection process. When presented with a huge amount of varied information, news consumers selectively interpret messages based on cognitive interpretation frameworks molded by their personal, family, and social life experiences. For a user with a limited range of topic exposure, there will be more topics that cause him to perceive contradictory values, leading to increased news avoidance and less amount of news consumption compared to users with more topic exposure.

This leads to the hypothesis:

Hypothesis 2.2: Users who are exposed to more diverse topics are likely to have large online news consumption in their next sessions.

3) *the time interval between the current session and the next*

The time interval between the current session and the next may correlate with his or her online news consumption amount of the next session for the following reason.

According to the theory of planned behavior, the social pressure that an individual feels when adopting a particular behavior can impact the individual's choice of adopting this behavior or not. As proposed by Barthel et al.(2020), nowadays many individuals regard news consumption as a "socially desirable" behavior, in other words, people may perceive social pressure for not reading news as they are afraid of falling behind. And because news is time-sensitive and sometimes consistent, people who have not read the news for a while are likely to read more news to make up for it, also to relieve the social pressure of not knowing what others are talking about.

This leads to the hypothesis:

Hypothesis 2.3: Users who have long intervals between two sessions are likely to have large online news consumption in their next sessions.

4 Research design

4.1 Research philosophy

As Saunders, Lewis and Thornhill (2012) said, "the research philosophy you adopt can be thought of as your assumptions about the way in which you view the world"(p.128). In other words, research philosophy is belief about how data is collected, analyzed and synthesized in a research. The two primary research philosophies are positivism and interpretivism, and positivism is used as the research philosophy in this research. Positivism adheres to the view that researchers are independent of research, and the researches are possible to be completely objective (Wilson, 2010). In addition, quantitative methods of analysis and large samples are always used in positivism.

4.2 Research approach

Research approaches are the approaches used in research projects based upon different ways of reasoning. There are three main research approaches: deductive approach, inductive approach, and abductive approach. Utilizing a deductive approach, the researcher begins the study with a theory, which is often generated through reading of academic literature, and then constructs a research plan to test the theory. If the researcher chooses an inductive approach, he or she begins by gathering data to investigate a phenomena and then develops or builds theory usually in the form of a conceptual framework. The abductive approach is when the researcher collects data to investigate a phenomena, find trends, and explain patterns in order to develop a new or alter an existing theory, which the researcher then tests with further data. In this research project, I choose the deductive approach, as the purpose of data collecting is to assess hypotheses that are connected to an existing theory.

4.3 Research methodology

This is quantitative research, because it investigates the connections between several variables that are quantitatively measured and analyzed using a variety of statistical approaches. Furthermore, I selected secondary data since it needs fewer resources, can give comparative and contextual information, as well as lead to unexpected new discoveries (Saunders, Lewis and Thornhill, 2012, p.318). More details about the data is explained in 5.1.

4.4 Research strategy

Research strategy outlines how a researcher will approach answering his or her research questions. In this research, I use the case of a news aggregator in reality, and study about the variables impacting user behavior. Case study, under this situation, can help with obtaining a thorough grasp of the research's background and the processes that are taking place.

5 Data analysis and findings

5.1 Data introduction

The raw data set used in this thesis is a dataset provided by a news aggregator service company. It's made up of 4,173,708 reading activity data from 95,094 individuals over the course of a month on the news aggregator site. There are 169,025 articles in this data

collection that are in the recommendation range, covering 135 different subjects. Each piece of reading behavior data represents one user's response to a specific article at a certain moment. Due to ethical and philosophical considerations, the supplier has desensitized the whole data collection because the majority of the data is directly connected to the user's personal information.

The obtained data was prepared prior to analysis. Missing data and outliers were verified in the dataset. Then I analyzed the data using python, the corresponding codes can be found in the appendix.

5.2 Measurement

5.2.1 User news consumption and time interval

A session is defined as all of a user's visits to the application in one day (from 0:00 to 24:00). The amount of user's online news consumption is measured by the number of articles the user visits during one month/ session depending on the characteristics of the hypotheses. Because the variable of "user news consumption during one month" is orders of magnitude bigger than other variables, I take the logarithm of it to eliminate this large difference, and to make the normal linear model used for testing hypotheses make more sense.

In this study, the time interval between two sessions is calculated by the day differences of these two sessions.

5.2.2 Channel diversity (in one month) and topic diversity (in one session)

I use the Herfindahl-Hirschman Index (HHI) to measure channel diversity and topic diversity. The Herfindahl-Hirschman Index is mostly used for assessing market competitiveness, however, I chose it because it's a practical statistical measure of concentration. When used to measure the channel diversity of user *i*, in other words, the distribution of different channels in the online news consumption of user *i*, HHI is defined as:

$$HHI_i = \sum_{j=1}^{n} S_j^2$$

where S_j is the number of visits as a percentage of total visits of the *j*th channel, and n is the number of total channels. For hypothesis 1, the channel diversity of user *i* is calculated on the basis of one month.

Similarly, when used to measure the topic diversity of user *i*, in other words, the distribution of different topics in the online news consumption of user *i*, S_j is the number of

visits as a percentage of total visits of the *j*th topic, and *n* is the number of total topics. For hypothesis 2, the topic diversity of user *i* is calculated on the basis of one session.

5.3 General analysis

Before testing the specific hypotheses, I first calculated and gathered the data of all the related variables. The general state of this data is presented in the table below. The channel/ topic diversity is 1 when the user sticks to only one channel, and the smaller the channel/ topic diversity data is, the more diverse the user's interests are.

	Maximum	Minimum	Medium
User news consumption (during one month)	4244.000	1.000	108.435
Log [user news consumption (during one month)]	8.353	0.000	2.483
User news consumption (during one day)	423.000	1.000	8.609
Channel diversity per user (during one month)	1.000	0.024	0.108
Topic diversity per user (during one month)	1.000	0.045	0.178
Time difference	29.000	1.000	1.856

Table 1: Comprehensive table of related variables

In addition, the distribution of these variables is presented below. As shown in the histograms, topic diversity per user, channel diversity per user, and user visits, these three variables all have skewed right distribution. In other words, most users read 0~500 articles in one month, with a few exceptions (big fans of the application.) that are distributed along a large range of higher values. Similarly, most users own relatively low topic diversity and channel diversity, which means that most of users' interests are quite diverse, with fewer users having less diverse interests.

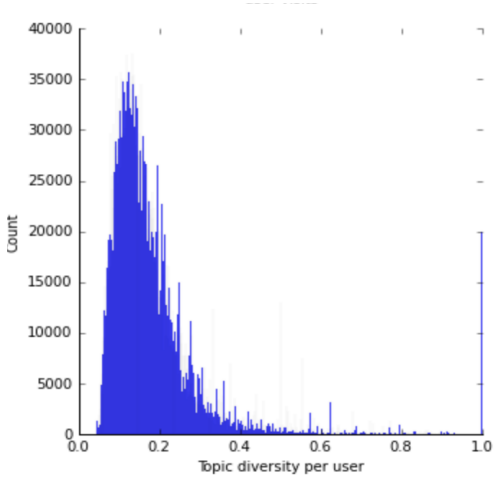

Figure 10: The distribution of topic diversity

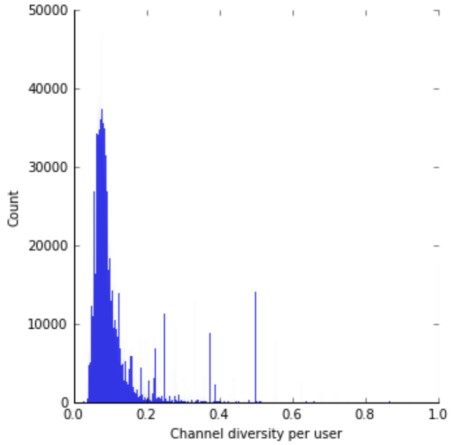

Figure 11: The distribution of channel diversity

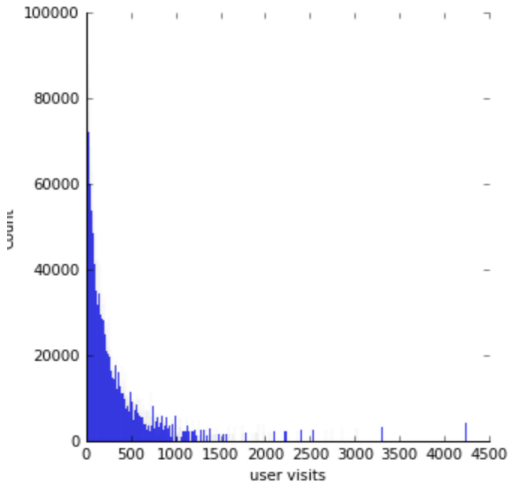

Figure 12: The distribution of user news consumption (during one month)

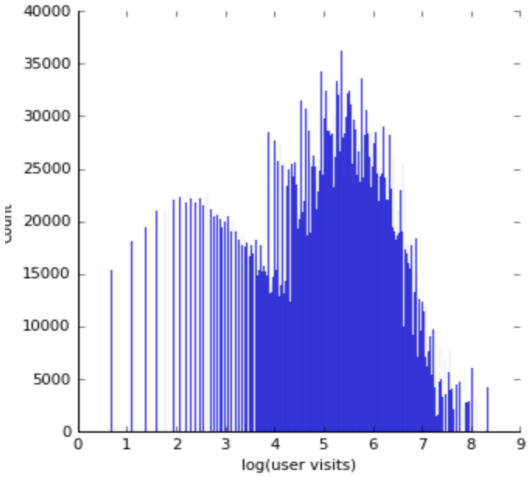

Figure 13: The distribution of log (user news consumption (during one month))

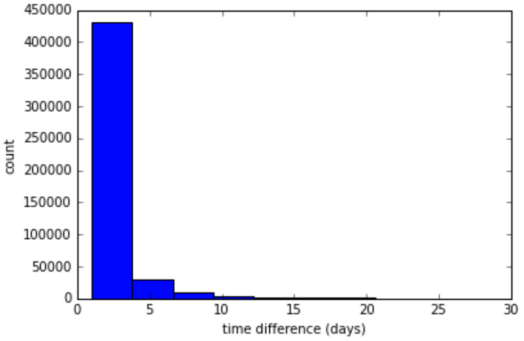

Figure 14: The distribution of time difference

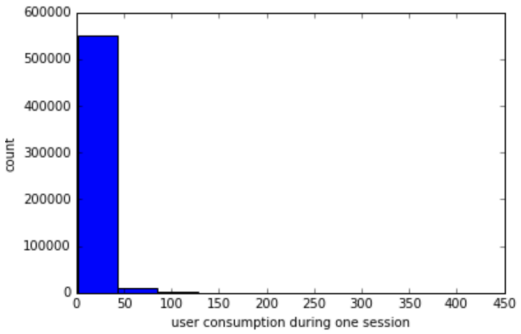

Figure 15: The distribution of user news consumption (during one session)

Scatterplots in the following illustrate the relationship between topic diversity per user and log (user visits), and the relationship between channel diversity per user and log (user visits). In agreement with the scatterplots, the direction of these relationships are both negative, which means that if a user visits the application more often, he/ she tends to have more diverse interests in topics/ channels.

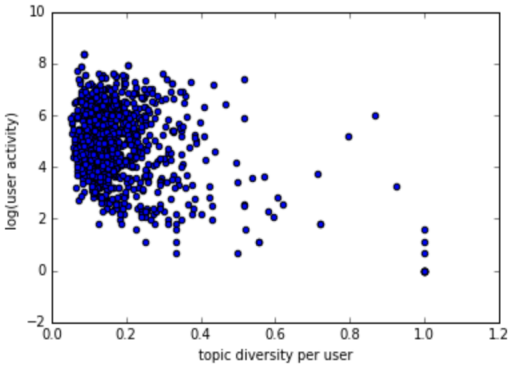

Figure 16: Log(user activity) vs. channel diversity per user

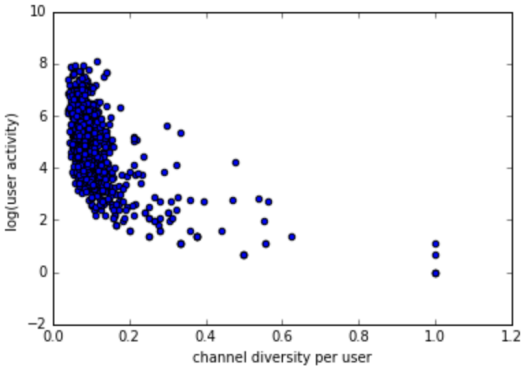

Figure 17: Log(user activity) vs. topic diversity per user

5.4 Statistic model

5.4.1 Testing hypothesis 1

Hypothesis 1 predicts the user's exposure to diverse channels increases the amount of his or her online news consumption. For every user, I calculated the channel diversity during one month, and took the logarithm of the user's online news consumption in one month.

I tested hypothesis 1 using OLS regression, and the results are shown below. R-squared is the measurement of how much of the independent variable is explained by changes in our dependent variables. In other words, 0.626 means our model explains 62.6% of the change in our user consumption variable. P>|t| is one of the most important statistics in the summary. It uses the t statistic to produce the p value, a measurement of how likely your coefficient is measured through our model by chance. The p value of 0.000 for channel diversity per user is saying there is a 0% chance the channel diversity per user variable has no effect on the dependent variable, user consumption. In addition, the negative coefficient means that these two variables have an inverse relationship. To be specific, the user is exposed to more diverse news, more news the user consumes; however, this experiment does not confirm a causal relationship between these two variables, and it is also possible that the more news the user consumes leads to a more diverse news consumption.

Dependant Variable: user_visits		
	coef	P>|t|
VARIABLES		
channel_diversity_per_user	-4.163	0
Observations	95094	
R squared	0.626	

Table 2: Results of testing hypothesis 1

5.4.2 Testing hypothesis 2

1) Preprocessing data

Hypothesis 2 predicts that users who 1) have large online news consumption, 2) are exposed to more diverse topics, and 3) have long intervals between two sessions are likely to have large online news consumption in their next sessions. For every user, I calculated the online news consumption and the channel diversity during every session, and the time

difference between two continuous sessions. Examples of data used for testing hypothesis 2 can be found in the appendix.

2) Empirical framework

According to hypothesis 2, users' online news consumption in their next sessions, i.e., the amounts of visited articles during next day, can be predicted together by three variables related with the previous session:

$$Consumption_{i,s+1} = Topic\ diversity_{i,s} + Consumption_{i,s} + \Delta t_{i,(s+1,s)}$$

The key dependent variable of interest is $Consumption_{u,s}$, which reflects the amount of news consumption by user *u* in session *s+1*. The first independent variable is $Topic\ diversity_{u,s}$, it is the topic diversity of news consumed by user *u* in session *s*. The second independent variable is $Consumption_{u,s}$, it is the topic diversity of news consumed by user *u* in session *s*. The second independent variable is $\Delta t_{(s+1,s)}$, it is the time difference between *s+1* and *s*.

3) Detecting Multicollinearity

In a multiple regression model, multicollinearity arises when two or more independent variables have a strong correlation. When several characteristics are highly linked, it might be difficult to separate their individual impacts on the dependent variable, so it is necessary to detect multicollinearity before testing the hypothesis.

Variance Inflation Factor (VIF) is one of approaches used to detect multicollinearity. Using the VIF technique, I choose each feature and regress it against all other features. The factor is determined as follows for each regression:

$$VIF = \frac{1}{1 - R^2}$$

Where, $R^2$is the coefficient of determination in linear regression. A VIF greater than 5 is considered as implying a high level of multicollinearity. And the following figure shows VIF for all the independent variables.

	feature	VIF
0	topic_diversity	1.673605
1	consumption	1.126195
2	time_diff	1.681827

Figure 18: results of VIF

As can be seen, all the VIF values are not greater than 5, showing that these variables are not closely linked, and there is no multicollinearity detected.

4) OLS regression

I tested hypothesis 2 using OLS regression, and the results are shown below.

R-squared of 0.334 means our model explains 33.4% of the change in our $Consumption_{u,\,s+1}$ variable. The p value of 0.000 for $\Delta t_{(s+1,\,s)}$ is saying there is a 0% chance $\Delta t_{(s+1,\,s)}$ variable has no effect on the dependent variable, $Consumption_{u,\,s+1}$. Also, there is a 0% chance $Topic\ diversity_{u,\,s}$ and $Consumption_{u,\,s}$ have no effect on $Consumption_{u,\,s+1}$.

Dependant Variable: next_con		
	coef	P>|t|
VARIABLES		
time_diff	-0.0213	0
topic_diversity	0.002	0
consumption	0.5714	0
Observations	478617	
R squared	0.334	

Table 3: Results of testing hypothesis 2

5) Including fixed effects

In statistics, complex phenomena in the real world are usually described by a combination of statistical models and errors. When using a random sample, the fixed effect model can only be used to draw inferences on the sample dataset. However, the random effects model permits predictions on population data relying on the theory of normal distribution. Thus, the individual characteristics are assumed to be associated with the independent variable in the fixed effects model, and not so in the random effects model.

In this study, every user has unique features which may or may not have an impact on the dependent variable, for example, the user's available time or age groups could influence the user's news consumption. Thus, fixed effects are chosen because it is believed that unobserved individual differences may influence outcome variables, and these unobserved individual differences should be accounted for.

To make the prediction closer to the reality, I include $factor(userID)$, which is individual fixed-effects to account for unobserved individual differences:

$$Consumption_{i,\,s+1} = Topic\ diversity_{i,\,s} + Consumption_{i,\,s} + \Delta t_{i,\,(s+1,\,s)} + factor(userID)$$

I tested hypothesis 2 again using panel OLS regression, and the results are shown below. R-squared is only 0.0821 this time, which means our model explains only 8.2% of the change in the dependent variable: the amount of the user's next news consumption. Because of the limitations of the database records, a significant number of users consume only once during one month thus have no data for their next consumption, and this big number of used data is possibly the reason for the low R-squared value in the fixed effects model. The p value of 0.000 for $\Delta t_{(s+1,\,s)}$ is saying there is a 0% chance $\Delta t_{(s+1,\,s)}$ variable has no effect on the dependent variable, $Consumption_{u,\,s+1}$. Also, there is a 0% chance $Topic\ diversity_{u,\,s}$ and $Consumption_{u,\,s}$ have no effect on $Consumption_{u,\,s+1}$.

In addition, the negative coefficient means that $\Delta t_{(s+1,\,s)}$ and $Consumption_{u,\,s+1}$ have an inverse relationship, and the positive coefficients refer to the positive correlation between $Topic\ diversity_{i,\,s}$ and $Consumption_{u,\,s+1}$, also the positive correlation between $Consumption_{i,\,s}$ and $Consumption_{u,\,s+1}$. In this case, the results of the test support hypothesis 2.1, but not support hypothesis 2.2 and hypothesis 2.3. However, comparing the coefficients of the independent variables, it can be concluded that the effect of the current consumption on the next consumption is much greater than that of the other two independent variables.

Dependant Variable: next_con		
	coef	P>|t|
VARIABLES		
time_diff	-0.0976	0
topic_diversity	0.0027	0
consumption	0.087	0
Included effects: entity		
Observations	478617	
R squared	0.0821	

Table 4: Results of testing hypothesis 2 (including fixed effects)

6 Conclusion

6.1 Interpretations

The results of the study show that many factors in recommendation systems have an impact on the user behavior and the weighting of the impact of these factors is different. The literature study and the results of statistical analysis have brought to light the answers to the research question.

What are the technical considerations when designing news recommendation systems?

The most used algorithms in recommendation systems are collaborative filtering methods and content based methods. Content-based recommendation systems discover the users' likes and dislikes based on history data, and recommend items which match with the

user’s preferences. By contrast, collaborative filtering recommendation systems are completely dependent on past interactions between users and items. Furthermore, the statistical results in this paper also demonstrate that users' past behaviour (news consuming and time between consumption behavior) is correlated with future behaviour, and therefore the appropriate algorithm can accurately predict user behaviour. While both approaches are effective in understanding users' interests and predicting news that may be of interest to them, it is also important to express respect for media diversity in the algorithms due to social impacts of recommendation systems, for example by automatically removing articles with high duplication rates from news aggregation sites, and by using a combination of human editing and algorithms.

What are the social considerations when designing news recommendation systems?

As the European Commission underlined in its Communication on tackling COVID-19 disinformation, free and plural media is important to address disinformation and enlighten citizens (European Commission, 2020). Access to news from different sources and on different topics, in other words, media diversity contributes to social cohesion, tolerance and the peaceful coexistence of different cultures, ideologies and perspectives. This is particularly true in this day and age, as society now relies heavily on online resources and digital tools to help us find and access information and news.

On the one hand, in the realm of online news, the proliferation of recommendation algorithms has sparked a lively debate about its potential negative impact on the social public sphere, as well as concerns about polarisation, filter bubbles, and misinformation and disinformation. However, these filter bubbles are not an inevitable consequence of digital technology, but rather the result of poor recommendation system design. Recommendation systems can make or break filter bubbles. They help to achieve or hinder public values and freedom of expression in a digital society. Much depends on the design of these systems. Some algorithms can be designed simply to generate clicks and short-term engagement, while others help users discover different news and views to consider social interests. Finding a way to realise the potential of algorithmic recommendation systems while promoting public values such as media diversity is a combined challenge for computer science, artificial intelligence, political science, media law and theory, and communication science.

On the other side, the significance of media diversity in promoting user consumption is not explored much in other literature. The results of this paper suggest that external pluralism in user consumption (source diversity) not only facilitates exposure to different

viewpoints, but also has a positive impact on the amount of user news consumption. In this case, recommendation systems should be designed with attention to the competition among news sources, also to prevent recommendation pages from being filled up by a few common news providers. In addition, internal pluralism (topic diversity), though having a small negative impact on the amount of user news consumption, still reflects social and political diversity. Therefore recommendation systems should be designed to provide a variety of topics as appropriate while taking into account user preferences.

6.2 Limitations

This study also has some limitations. Firstly, I used one month of news data to calculate the amount of news consumed by users, topic diversity, channel diversity, etc. However, if more data were used, the results of the calculations could be different and more representative. Therefore, we need to test our empirical model against more data. Second, the data for this study was sourced from a representative news aggregation website. However, this news aggregation platform is not comparable to news aggregation platforms such as google news. Thirdly, the study did not consider how long users spent on each article and how they reacted to the article (sharing, liking, commenting, etc.), and the future research needs to consider the amount of time news consumers spend on articles and their reactions. Using these factors as dependent variables., the users attitudes towards the articles could be better described. Lastly, the thesis only studies the impact of recommendation systems from the literature, and doesn't consist of a laboratory experiment in controllable environments, which helps to compare the impact of different recommendation systems on the diversity of user consumption and the amount of user consumption. The laboratory experiment could add more reliability to the study and provide more convincing suggestions about how to build recommendation systems for practitioners.

6.3 Strong sides

The thesis investigates the importance of media diversity and proposes design principles in recommendation systems by combining concepts from behavior science and data science. In summary, this paper seeks to do two things: report on a study that implements an innovative and effective approach to measuring media diversity in online news recommendations, and reflect more broadly on the challenges of determining normative criteria for what constitutes a 'good' recommendation system as they increasingly act as key gatekeepers in the online news environment. The central role of machines in the production,

distribution and consumption of news makes normative judgements about their behaviour a fundamental theme of research. In the news market, as in many other areas of human activity, machines and the cultural, political and economic interests behind them are reshaping the landscape, particularly in terms of quality, diversity, and quantity, at a faster rate than scholars and practitioners can assess their social impact. The need for empirical and normative investigation is particularly acute as this is a formative period for the integration of machines into the news, with significant implications for the production and dissemination of news. At the same time, this article shows that there is no universally accepted standard for humans as news gatekeepers, making it even more complex to assess the performance of machines in that role.

6.4 Recommendations for practitioners

These results have policy implications for news publishers and designers of recommendation systems. With the rise of recommendation systems in online news, it was irresistible that they would become the new gatekeepers in journalism. However, the role of gatekeeper in journalism has long been filled by humans, so the emerging role of recommender systems as gatekeepers has both a communication function and a public significance. Many researchers argue that democratic public life benefits from a wealth of voices, and recommendation systems should therefore highlight diverse information from a large number of sources that offer a wide range of perspectives. Such problems can be difficult without a deeper understanding of the design and deployment of algorithms. The designing of recommendation systems should be critically examined, and then evaluated against a normative background. More specifically, first, news publishers and aggregators need to work to reduce filter bubbles. News aggregators might propose policies to maintain news diversity or develop an algorithm to automatically remove articles with high duplication rates from their news aggregation sites. Another approach could be to consider a combination of human editing and algorithms to prevent filter bubbles.

These results also have policy implications for public policy-makers. Achieving digital media diversity is not a problem that can be solved by technical design alone. It is also the result of how professionals and end users interact with technology, the way digitisation and automation change the news value chain, the way decision-making power is organised and, more broadly, what the conditions are for digital media. These conditions are influenced and shaped by laws and regulations, starting with copyright and data protection laws, to the role of law in stimulating innovation, accountability and diverse media landscapes. Therefore,

to provide governments with a toolset for regulating the media environment without stifling innovation and democratic rights, it is critical to include legal and media regulation perspectives. To be more specific, instead of transferring existing regulations from broadcasting to news aggregators, more interdisciplinary research across computer, social, and legal sciences are needed to ensure diverse, transparent, explainable, and fair news recommendations.

References

Su, X., & Khoshgoftaar, T. M. (2009). A survey of collaborative filtering techniques. *Advances in artificial intelligence*, *2009*.

Linden, G., Smith, B., & York, J. (2003). Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet computing*, *7*(1), 76-80.

Ansari, A., Essegaier, S., & Kohli, R. (2000). Internet recommendation systems.

Goldberg, D., Nichols, D., Oki, B. M., & Terry, D. (1992). Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, *35*(12), 61-70.

Acilar, A. M., & Arslan, A. (2009). A collaborative filtering method based on artificial immune network. *Expert Systems with Applications*, *36*(4), 8324-8332.

Breese, J. S., Heckerman, D., & Kadie, C. (2013). Empirical analysis of predictive algorithms for collaborative filtering. *arXiv preprint arXiv:1301.7363*.

Bokde, D., Girase, S., & Mukhopadhyay, D. (2015). Matrix factorization model in collaborative filtering algorithms: A survey. *Procedia Computer Science*, *49*, 136-146.

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

Schafer, J. B., Konstan, J. A., & Riedl, J. (2001). E-commerce recommendation applications. *Data mining and knowledge discovery*, *5*(1), 115-153.

Lops, P., De Gemmis, M., & Semeraro, G. (2011). Content-based recommender systems: State of the art and trends. *Recommender systems handbook*, 73-105.

Thorat, P. B., Goudar, R. M., & Barve, S. (2015). Survey on collaborative filtering, content-based filtering and hybrid recommendation system. *International Journal of Computer Applications*, *110*(4), 31-36.

Das, A. S., Datar, M., Garg, A., & Rajaram, S. (2007, May). Google news personalization: scalable online collaborative filtering. In *Proceedings of the 16th international conference on World Wide Web* (pp. 271-280).

Liu, J., Dolan, P., & Pedersen, E. R. (2010, February). Personalized news recommendation based on click behavior. In *Proceedings of the 15th international conference on Intelligent user interfaces* (pp. 31-40).

Garcin, F., Dimitrakakis, C., & Faltings, B. (2013, October). Personalized news recommendation with context trees. In *Proceedings of the 7th ACM Conference on Recommender Systems* (pp. 105-112).

Dellarocas, C., Katona, Z., & Rand, W. (2013). Media, aggregators, and the link economy: Strategic hyperlink formation in content networks. *Management science*, *59*(10), 2360-2379.

Gross, B. M. (1964). *The managing of organizations: The administrative struggle* (Vol. 2). Free Press of Glencoe.

Kim, E., Nam, D. I., & Stimpert, J. L. (2004). The applicability of Porter's generic strategies in the digital age: assumptions, conjectures, and suggestions. *Journal of management*, *30*(5), 569-589.

Hamborg, F., Meuschke, N., & Gipp, B. (2020). Bias-aware news analysis using matrix-based news aggregation. *International Journal on Digital Libraries*, *21*(2), 129-147.

Garcin, F., Zhou, K., Faltings, B., & Schickel, V. (2012, December). Personalized news recommendation based on collaborative filtering. In *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology* (Vol. 1, pp. 437-441). IEEE.

Claussen, J., Peukert, C., & Sen, A. (2019). The editor vs. the algorithm: Targeting, data and externalities in online news. Data and Externalities in Online News (June 5, 2019).

Ebbes, P., & Netzer, O. (2021). Using Social Network Activity Data to Identify and Target Job Seekers. Management Science.

Athey, S., Mobius, M., & Pal, J. (2021). The impact of aggregators on internet news consumption (No. w28746). National Bureau of Economic Research.

Oestreicher-Singer, G., and Sundararajan, A. (2012a). “Recommendation networks and the long tail of electronic commerce.” MIS Quarterly, 36 (1).

De, P., Hu, Y., & Rahman, M. S. (2010). Technology usage and online sales: An empirical study. *Management Science*, *56*(11), 1930-1945.

Bhadani, S. (2021, September). Biases in Recommendation System. In *Fifteenth ACM Conference on Recommender Systems* (pp. 855-859).

Greenstein‐Messica, A., Rokach, L., & Shabtai, A. (2017). Personal‐discount sensitivity prediction for mobile coupon conversion optimization. *Journal of the Association for Information Science and Technology*, *68*(8), 1940-1952.

Doyle, G. (2002). Media ownership: The economics and politics of convergence and concentration in the UK and European media. Sage.

Fletcher, R., Cornia, A., & Nielsen, R. K. (2020). How polarized are online and offline news audiences? A comparative analysis of twelve countries. The International Journal of Press/Politics, 25(2), 169-195.

Bruns, A. (2019). Filter bubble. *Internet Policy Review*, *8*(4).

Lewin, K. (1943). Forces behind food habits and methods of change. Bulletin of the national Research Council, 108(1043), 35-65.

White, D. M. (1950). The “gate keeper”: A case study in the selection of news. Journalism quarterly, 27(4), 383-390.

Galtung, J., & Ruge, M. H. (1965). The structure of foreign news: The presentation of the Congo, Cuba and Cyprus crises in four Norwegian newspapers. Journal of peace research, 2(1), 64-90.

Shoemaker, P. J., & Vos, T. (2009). Gatekeeping theory. Routledge.

Vermeer, S., Trilling, D., Kruikemeier, S., & de Vreese, C. (2020). Online news user journeys: the role of social media, news websites, and topics. Digital Journalism, 8(9), 1114-1141.

Solsman, J. E. (2018, January 10). YouTube’s AI is the puppet master over most of what you watch. CNET. https://www.cnet.com/news/youtube-ces-2018-neal-mohan/

Wallace, J. (2018). Modelling contemporary gatekeeping: The rise of individuals, algorithms and platforms in digital news dissemination. Digital Journalism, 6(3), 274-293.

Ajzen, I. (1991). The theory of planned behavior. Organizational behavior and human decision processes, 50(2), 179-211.

Fishbein, M., & Ajzen, I. (1977). Belief, attitude, intention, and behavior: An introduction to theory and research. Philosophy and Rhetoric, 10(2).

Madden, T. J., Ellen, P. S., & Ajzen, I. (1992). A comparison of the theory of planned behavior and the theory of reasoned action. Personality and social psychology Bulletin, 18(1), 3-9.

Hansen, T., Jensen, J. M., & Solgaard, H. S. (2004). Predicting online grocery buying intention: a comparison of the theory of reasoned action and the theory of planned behavior. International Journal of Information Management, 24(6), 539-550.

Just, N., & Latzer, M. (2017). Governance by algorithms: reality construction by algorithmic selection on the Internet. *Media, culture & society*, *39*(2), 238-258.

Tandoc Jr, E. C. (2014). Journalism is twerking? How web analytics is changing the process of gatekeeping. *New media & society*, *16*(4), 559-575.

Just, N., & Latzer, M. (2017). Governance by algorithms: reality construction by algorithmic selection on the Internet. *Media, culture & society*, *39*(2), 238-258.

Vu, H. T. (2014). The online audience as gatekeeper: The influence of reader metrics on news editorial selection. Journalism, 15(8), 1094-1110.

Budak, C., Goel, S., & Rao, J. M. (2016). Fair and balanced? Quantifying media bias through crowdsourced content analysis. Public Opinion Quarterly, 80(S1), 250-271.

Liu, J., Dolan, P., & Pedersen, E. R. (2010, February). Personalized news recommendation based on click behavior. In Proceedings of the 15th international conference on Intelligent user interfaces (pp. 31-40).

McQuail, D. (1992). *Media performance: Mass communication and the public interest* (Vol. 144). London: Sage.

Ouellette, J. A., & Wood, W. (1998). Habit and intention in everyday life: The multiple processes by which past behavior predicts future behavior. *Psychological bulletin*, *124*(1), 54.

Slaets, A., Verhoest, P., d'Haenens, L., Minnen, J., & Glorieux, I. (2020). Fragmentation, homogenisation or segmentation? A diary study into the diversity of news consumption in a high-choice media environment. European Journal of Communication, 0267323120966841.

Barthel, M., Mitchell, A., Asare-Marfo, D., Kennedy, C., & Worden, K. (2020). Measuring news consumption in a digital era. Pew Research Center's Journalism Project, 8.

Macharia, S. (2015). Global media monitoring project. World Association for Christian Communication. Retrieved from http://whomakesthenews. org/gmmp/gmmp-reports/gmmp-2015-reports.

Saunders, M., Lewis, P., & Thornhill, A. (2012). Research methods for business students (6. utg.). Harlow: Pearson.

Wilson, J. (2014). *Essentials of business research: A guide to doing your research project*. Sage.

Rhoades, S. A. (1993). The herfindahl-hirschman index. Fed. Res. Bull., 79, 188.

Flaxman, S., Goel, S., & Rao, J. M. (2016). Filter bubbles, echo chambers, and online news consumption. Public opinion quarterly, 80(S1), 298-320.

EUROPEAN COMMISSION. (2020, October 6). Tackling COVID-19 disinformation - Getting the facts right. EUR-Lex. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52020JC0008

Möller, J., Trilling, D., Helberger, N., & van Es, B. (2018). Do not blame it on the algorithm: an empirical assessment of multiple recommender systems and their impact on content diversity. *Information, Communication & Society*, *21*(7), 959-977.

Bernstein, A., de Vreese, C., Helberger, N., Schulz, W., Zweig, K., Baden, C., ... & Zueger, T. (2020). Diversity in news recommendations. *arXiv preprint arXiv:2005.09495*.

Fleder, D. M., Hosanagar, K., & Buja, A. (2010). Recommender systems and their effects on consumers: the fragmentation debate. *EC*, *229*, 230.

Appendices

Appendix A) Codes for testing the hypotheses

1. Import data

```
#!/usr/bin/env python3
# -*- coding: utf-8 -*-
```

```
"""
Created on Sun Jul  4 23:58:33 2021

@author: vivian
"""

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from datetime import datetime

from scipy.stats import spearmanr
from scipy.stats import kendalltau

import statsmodels.api as sm
import pandas
from patsy import dmatrices
from statsmodels.graphics.api import abline_plot
from sklearn.preprocessing import scale
import seaborn as sns
import matplotlib.pyplot as plt
plt.style.use('classic')
%matplotlib inline
import numpy as np
import pandas as pd

from google.colab import drive
drive.mount('/content/gdrive')

metadata = pd.read_csv('/content/gdrive/My
Drive/WU_Wanyi_thesis/data/Data_masked_one_month.csv',
                       low_memory=False)


metadata['time'] = pd.to_datetime(metadata.time)
metadata['date'] = metadata['time'].values.astype('datetime64[D]')
metadata['count of topics'] = metadata.count(axis='columns')-5

pd.set_option('display.max_columns', None)

metadata.info()
print(metadata)
```

```
## for all the articles which are not marked with any topics, mark them
as topic 135
metadata['topic1'].replace(np.nan,135,inplace = True)

print(metadata.tail(10))
```

2. Define variables in h1

```
## user_visits = how many articles they read in one month

user_act = metadata.groupby('member_id').size().to_dict()
metadata['user_visits'] = metadata['member_id'].map(user_act)
print(metadata)

## channel_diversity_per_user = the diversity of channels in one month
per user
c_cols=['channel_id','member_id']
c_data = metadata[c_cols]

print(c_data)

def hhi(series):
   _, cnt = np.unique(series, return_counts=True)
   return np.square(cnt/cnt.sum()).sum()

user_c_div = c_data.groupby('member_id').agg({'channel_id': hhi})
print(user_c_div)
dict_user_c_div = dict(zip(user_c_div.index,user_c_div['channel_id']))
print(dict_user_c_div)

metadata['channel_diversity_per_user'] =
metadata['member_id'].map(dict_user_c_div)
print(metadata)
```

3. Test h1

```
## build the data for testing h1
h1cols = ['member_id', 'user_visits','channel_diversity_per_user']
h1data = metadata[h1cols]
h1data = h1data.drop_duplicates(keep='first')
h1data['user_visits'] = np.log(h1data['user_visits'])

print(h1data)
```

```
y1, X1 = dmatrices('user_visits ~ channel_diversity_per_user',
data=h1data, return_type='dataframe')

mod = sm.OLS(y1,X1)
res = mod.fit()
print(res.summary())
print(res.params)
```

4. Define variables in h2

```
ucols1 = ['member_id','topic1','date','post_id']
udata1 = metadata[ucols1]
udata1.rename(columns={'topic1':'topic'},inplace=True)

ucols2 = ['member_id','topic2','date','post_id']
udata2 = metadata[ucols2]
udata2.rename(columns={'topic2':'topic'},inplace=True)

ucols3 = ['member_id','topic3','date','post_id']
udata3 = metadata[ucols3]
udata3.rename(columns={'topic3':'topic'},inplace=True)

ucols4 = ['member_id','topic4','date','post_id']
udata4 = metadata[ucols4]
udata4.rename(columns={'topic4':'topic'},inplace=True)

frames = [udata1,udata2,udata3,udata4]
u_data = pd.concat(frames)
u_data = u_data.dropna(axis=0, how='any')

print(u_data)

## user visits per day, topic diversity per day
def hhi(series):
   _, cnt = np.unique(series, return_counts=True)
   return np.square(cnt/cnt.sum()).sum()

con_div = u_data.groupby(['member_id','date']).agg({'topic':
hhi,'post_id':np.count_nonzero})

con_div.rename(columns={'topic':'topic_diversity','post_id':'consumptio
n'},inplace=True)
con_div1 = con_div.reset_index()
```

```
print(con_div1)

## next consumption
con_div2 = con_div1.groupby(['member_id']).apply(lambda
x:x['consumption'].shift(-1))
con_div2 = con_div2.reset_index()
## print(con_div2)

dict_new_com = dict(zip(con_div2.index,con_div2['consumption']))
con_div1['next_com'] = con_div1.index.map(dict_new_com)
print(con_div1)

con_div3 = con_div1.groupby(['member_id']).apply(lambda
x:x['date'].shift(-1) - x['date'])
con_div3 = con_div3.reset_index()

dict_new_com = dict(zip(con_div3.index,con_div3['date']))
con_div1['time_diff'] = con_div1.index.map(dict_new_com)
print(con_div1.head(20))
```

5. Test h2

```
sel_cols = ['topic_diversity', 'consumption', 'next_com', 'time_diff']
statdata2 = con_div1[sel_cols]
statdata2 = statdata2.dropna()
statdata2['time_diff'] = (statdata2['time_diff'] /
np.timedelta64(1,'D')).astype(int)
print(statdata2)


##statdata2 = statdata2.apply(lambda x: (x - np.min(x)) / (np.max(x) -
np.min(x)))
##print(statdata2)

y2, X2 = dmatrices('next_com ~ time_diff + topic_diversity +
consumption', data=statdata2, return_type='dataframe')

mod = sm.OLS(y2,X2)
res = mod.fit()
print(res.summary())
print(res.params)
```

6. Apply fixed effects

```
from statsmodels.datasets import grunfeld
from linearmodels import PanelOLS
import statsmodels.api as sm

statdata3 = statdata2.set_index(["member_id","date"])
print(statdata3)

exog =
sm.add_constant(statdata3[['topic_diversity','consumption','time_diff']
])
grunfeld_fe = PanelOLS(statdata3['next_con'], exog, entity_effects=
True, time_effects=True)
grunfeld_fe = grunfeld_fe.fit()
print(grunfeld_fe)
```

Appendix B) Codes for building a primary recommendation system

1. import data

```
#!/usr/bin/env python3
# -*- coding: utf-8 -*-
"""
Created on Sun Jul  4 23:58:33 2021

@author: vivian
"""

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from datetime import datetime

from google.colab import drive
drive.mount('/content/gdrive')

metadata = pd.read_csv('/content/gdrive/My
Drive/WU_Wanyi_thesis/data/Data_masked_one_month.csv',
                       low_memory=False)


metadata['time'] = pd.to_datetime(metadata.time)
metadata['date'] = metadata['time'].values.astype('datetime64[D]')
```

```
metadata['count of topics'] = metadata.count(axis='columns')-5

pd.set_option('display.max_columns', None)

metadata.info()
print(metadata)
```

2. find all the posts

```
post_data = metadata.drop_duplicates(subset = ['post_id'],keep='first')

m_cols = ['post_id', 'topic1', 'topic2','topic3',
'topic4','topic5','topic6','count of topics','time']

post_data = post_data[m_cols].reset_index(drop = True)

##post_data.info()

## group posts

nan_posts =  post_data[post_data['topic1'].isnull()]

print(nan_posts)
```

3. for every post, build a 134 dimension vector

```
posts = post_data['post_id']

topics = np.arange(1,136,1).tolist()


allList = 0

df = pd.DataFrame(allList,columns=topics,index=posts)


##print(df.tail())


##one post has no more than six topics

topicsmax = np.arange(1,7,1).tolist()
```

```
##for every post, fill in the every topic with 1 or 0
for t in topicsmax:
 for i in posts:
   j = post_data.iloc[i-1,t]
   if j > 0:
     j = j.astype(int)
     df.iloc[i-1,j-1] = 1

print(df.tail())
```

4. calculate the similarity between two posts with Euclidean distance

```
def sim_distancec(post1, post2):

   score_post1 = df.iloc[post1-1]
   score_post2 = df.iloc[post2-1]

   distance = np.sqrt(((np.array(score_post1) - np.array(score_post2))
** 2).sum())
   return distance

##sim_distancec(1, 2)
```

5. sort all the other posts by the similarity between the input post

```
 """
def cal_all_post_distance(post) -> list:
   all_post_sim = [(sim_distancec(p, post),p) for p in posts if p !=
post]
```

```
    all_post_sim.sort()

    return all_post_sim[0:100]

"""
```

6. However, there are so many posts owning totally the same topics as the input post, thus we only output the newest posts for recommendation.

```
def cal_all_post_distance(post) -> list:


    all_post_sim = [(sim_distancec(p, post),p) for p in posts if p !=
post]

    all_post_sim.sort()


    ## find all the posts which own the same topics as the input post

    col =['distance','postid']

    same_topics = pd.DataFrame(all_post_sim,columns=col)

    same_topics = same_topics[same_topics['distance'].isin(['0'])]


    ## find the publish time of these posts

    same_topics.loc[:, 'time'] = None

    same_topics.set_index(["postid"], inplace=True)


    for i in same_topics.index:

      same_topics['time'].loc[i] = post_data['time'].loc[i-1]


    ##select the 10 newest posts

    same_topics = same_topics.sort_values("time",ascending = False )
```

```
    return same_topics[:10]



def main(post_id):
    print('recommendation list', cal_all_post_distance(post_id))



if __name__ == '__main__':
    main(9)
    main(18972)
    main(389)
```

Appendix C) Examples of data

	member_id	user_visits	channel_diversity_per_user
0	1	5.099866	0.062463
1	2	4.454347	0.054624
2	3	5.236442	0.081768
3	4	3.637586	0.240997
4	5	4.510860	0.088757

Figure 19: example of data used for testing hypothesis 1

member_id	date	topic_diversity	consumption	next_con	time_diff
1	2021-05-01	0.168242	23	29.0	1
	2021-05-02	0.103448	29	15.0	1
	2021-05-03	0.200000	15	40.0	1
	2021-05-04	0.143750	40	32.0	1
	2021-05-05	0.105469	32	28.0	1
...		...	...	...	...
94686	2021-05-29	1.000000	1	1.0	1
94708	2021-05-29	0.500000	4	3.0	1
94716	2021-05-29	0.500000	2	2.0	1
94742	2021-05-29	0.500000	2	1.0	1
94974	2021-05-30	0.500000	2	1.0	1

Figure 20: example of data used for testing hypothesis 2