








# What Prompts Users to Click on News Headlines? A Clickstream Data Analysis of the Effects of News Recency and Popularity

Tingting Jiang<sup>1,2</sup> , Qian Guo<sup>1</sup> , Yaping Xu<sup>1</sup> , Yang Zhao<sup>1</sup> ,  
and Shiting Fu<sup>1</sup> 

<sup>1</sup> School of Information Management, Wuhan University, Wuhan, Hubei, China  
tij@whu.edu.cn

<sup>2</sup> Center for Studies of Information Resources, Wuhan University,  
Wuhan, Hubei, China

**Abstract.** A new headline nowadays has to compete for readers' attention and sometimes it needs to entice readers to click and read the news article. The peripheral indicators of news headlines would provide visual suggestions for user to decide on which news to read and which to ignore. This study focused on the recency and popularity indicators of online news. For the purpose of revealing the relationships between news recency/popularity and users' clicking behavior, a 2-month server log file containing 39,990,200 clickstream records from an institutional news site was analyzed in combination with the news recency and popularity information crawled from its homepage. It was found that more recent or more popular news headlines received more clicks. The results have important implications for news providers in creating effective news headlines and in publishing and disseminating news more responsibly. The introduction of unobtrusive clickstream data to user behavior analysis is a major methodological contribution.

**Keywords:** News headlines · Recency · Popularity · Clickstream data

## 1 Introduction

In the era of print newspapers, people used to glance through a newspaper page by page and read the news articles of interest [1]. The headline was a succinct and accurate summary of the corresponding news story [2]. As more and more news is consumed online today, news reading has become increasingly rapid and shallow [3]. The major role of headlines has shifted to attracting attention [4].

In the face of a tremendous amount of information, users tend to avoid extra efforts and take advantage of peripheral indicators to make a selection. These indicators, e.g. source of information, are useful cues that help users determine information salience, especially when they are involved in casual scanning [5]. On social media, users make evaluation of information quality in virtue of bandwagon cues, i.e. the number of an individual's followers, and they are likely to imitate others' judgment and behavior [6]. With the development of Web technologies, online news sites have been enabled to

indicate past readers' interaction with the news, such as the frequency of reading and the overall rating of news story, to inform future readers. Such indicators are usually displayed around news headlines so that readers can find them easily [7].

Recency and popularity indicators are widely seen on news sites. News is time-sensitive information, and recency is a dominant criterion for judging newsworthiness in news media. A piece of news will carry more weight if it is recent [5, 8]. The popularity of a new story can be measured with the number of clicks, likes, or comments it has received. The lapse of time decreases recency while more reading increases popularity. The two indicators are essentially the variable properties of news headlines [8].

Although news sites make efforts to provide recency and popularity indicators, there still lacks empirical evidence of how users actually make use of them. As a result, this study aimed to investigate whether the indicated recency and popularity prompt users to click on news headlines. For this purpose, a 2-month server log file containing 39,990,200 clickstream records from an institutional news site was analyzed in combination with the news recency and popularity information crawled from its homepage.

## 2 Related Work

### 2.1 Recency and Popularity of Online Information

Both recency and popularity indicators have been used to attract attention and/or interaction to online news as well as other types of online information.

Recency or timeliness refers to how recent the information is. It is an important aspect of information quality and has an impact on users' assessment of the credibility of information [9]. According to a study of social news aggregators, when the credibility of the source was low, more recent news received more attention; whereas in the case of high-credibility source, more recent news had higher perceived newsworthiness [5]. Social media enable the real-time sharing of all kinds of information and cater to users' desire to be kept up-to-date. Further evidence has suggested that faster updates would lead to higher source credibility as mediated by cognitive elaboration [10]. Also for electronic word-of-mouth messages, i.e. online comments or reviews provided by customers about a product or company, their currency has been found to positively influence customers' perception of their usefulness which in turn predicts purchase intention [11].

As the overwhelming amounts of information may create cognitive overload, it is natural for individuals to follow others' choices and attitudes when evaluating the information rather than relying on their own judgement [12]. User-generated content on social media, e.g. micro-posts, will receive likes, reposts, and comments and replies which together indicate popularity. Higher popularity is often associated with greater perceived usefulness and stronger preferences and thus increases the intention of interaction and possibility of actual interaction [13, 14]. Similar results have been obtained for online shopping: the social popularity of a product, i.e. having a large number of people liking or purchasing it, would affect positively customers' trust, perception of product quality and value as well as purchase intention and behavior [15, 16].

Ksiazek and Peer [17] measured the popularity of online news videos in terms of the numbers of views, favorited, and ratings and found that more popular videos would attract more comments while less popular ones more replies to comments.

## 2.2 Clickstream Data Analysis

Clickstream data is a typical type of trace data generated on Web servers when users visit websites or APPs for their own purposes. It captures all the clicks or page requests made by users in sequence from entering to leaving a site. Each clickstream record basically informs us which user performs which type of action on which page at what time. Clickstream data analysis now can be found in user behavior research in many contexts, such as social media, social commerce, online courses, information portals, and so on [18]. A general framework for analyzing clickstream data has taken shape which consists of three levels, i.e. footprint, movement, and pathway. When a user visits a site, each click causes a movement, the changing of location from one page to another, and leaves a footprint, a mark showing the user's presence on a page. The click series during that visit, i.e. chaining all the movements in a chorological order, engenders a pathway, indicating the process in which the user interacts with the site. This framework has been successfully applied in the studies of users' information seeking behavior in social library systems [19, 20] and academic library OPAC systems [21].

## 3 Data Collection and Preparation

This study introduced clickstream data analysis to the investigation of real-world users' news reading behavior. A server log file was obtained from an institutional news site that affiliates to the official site of a renowned Chinese university. It contains a total of 39,990,200 clickstream records generated between March 1st, 2017 and April 30th, 2017. The six basic fields in the log are *User-IP* (client IP address, e.g. "202.114.65.\*\*\*"), *Date* (date on which request is made, e.g. "28/Mar/2017"), *Time* (time when request is made, e.g. "08:00:10"), *Method* (type of client to server request, e.g. "GET"), *URL* (URL of the resource requested, e.g. "/info/1002/40929.htm"), and *Status* (HTTP status code returned by server, e.g. "200").

The log was cleaned to eliminate corrupted and redundant records in the first place. Corrupted records were errors produced when the server performed logging incorrectly and were easily recognizable for not fitting the patterns of the normal data in the same field. Redundant records were those irrelevant to the objective of this study, including unsuccessful requests, data submission requests, and requests for pictures, styles, scripts, and other resources.

The next step was to define sessions. Each session is composed of all the records deriving from one visit to the site, and a user may have more than one session for visiting the site multiple times during the two months. So different users were identified with their IP addresses (*User-IP* field); for the same user, if the time interval between any two records exceeded 30 min, they would be divided into different sessions. Given that some visits recorded in the log file might be attributed to search engine spiders, a cut-off of 101 records was adopted to determine non-human sessions: if a session

contained 101 or more records, it was assumed to represent a non-human visit and thus excluded. As a result, approximately 10% ( $N = 3,987,030$ ) of the records remained which involved 839,685 sessions.

## 4 Data Analysis and Results

The institutional news site allows users to find news in multiple ways. With an explicit need or interest, users would perform keyword searching or browse through particular categories (e.g. *Academic News* and *Alumni News*). However, news reading has become a daily monitoring activity to many people [22]. It is common for users to start from the homepage where hundreds of selected news headlines were displayed to see what was happening to the university. Their undirected scanning on the homepage was more likely to be affected by the peripheral indicators. Therefore, this study analyzed the clickstream data mainly at the movement level with a focus on the movements from the homepage to news article pages. The type of a page was identified with a specific string in its URL.

During the two-month time period, a Web crawler tool was meanwhile employed to scrape the homepage of the news site once a day at 23:00, for the purpose of capturing its daily update of news headlines, including the changes of the recency and popularity indicators which refer to the date of publishing and the number of reads respectively. The crawled data could be linked to the clickstream data as each news headline points to a news article which has a distinct URL identifiable in the log file. Specifically, the crawled data provided information about a headline's recency and popularity, while the clickstream data how many clicks it attracted.

### 4.1 Headlines with Recency and Popularity Indicators on the Homepage

The homepage of the institutional news site demonstrates a traditional layout with a navigation bar on the top and news headlines enclosed into 18 blocks below it. These blocks correspond to different news categories. Only two of the blocks have recency and/or popularity indicators appended to each news headline. The *Important News* (news events happening to the university) block displays both indicators, whereas the *Media News* (news about the university published on mainstream media) block only the former.

Both blocks are updated every day. Newly published headlines will be inserted to the top and gradually edge out older ones on the bottom from the blocks. There was an overlap between the headlines on different days. Even if a headline stayed in these blocks for days, its recency would decrease and popularity might increase. Therefore, this study treated every headline on a single day as a distinct headline.

As extracted from the crawled data, all the distinct headlines that ever appeared in both blocks during the two months added up to 888, with 488 in *Important News* and 400 in *Media News* respectively. According to the movement level analysis based on the clickstream data, the headlines on the homepage attracted 98,016 clicks in total. About one third of the clicks ( $N = 33,919$ , 34.61%) were contributed by the headlines

in the above two blocks. This is reasonable given that they are placed in the most conspicuous positions on the homepage.

#### 4.2 Relationships Between News Recency/Popularity and Users' Clicking Behavior

The news published on the institutional news site became obsolete at a much slower rate than general online news for confined to the university. The oldest important news or media news that users might see on the homepage could be traced back to more than 2 months ago (probably due to the winter vacation during which news was published infrequently). This study defined two different levels of news recency for the convenience of analysis: high (published within 1 week, including 7 days) and low (published more than 7 days ago). High-recency headlines were in the majority ( $N = 698$ ) while low-recency ones much less frequently seen ( $N = 190$ ).

As mentioned above, the popularity indicator is only available to important news. The numbers of reads varied from news to news, ranging from hundreds to thousands. They averaged around 1,000 which was used as the boundary to distinguish two groups of headlines of different popularity: high (read more than 1,000 times) and low (read 0-1,000 times). There were almost twice as many low-popularity headlines ( $N = 312$ ) as high-popularity ones ( $N = 176$ ).

The Mann-Whitney U test was conducted to examine the relationships between news recency/popularity and users' clicking behavior. As can be found in Table 1, significant results were obtained for both recency ( $Z = -15.366$ ,  $p < .05$ ) and popularity ( $Z = -17.889$ ,  $p < .05$ ). To be more specific, users were more likely to click on high-recency headlines than on low-recency ones (mean rank: high 513.50 > low 191.00), and high-popularity headlines attracted more clicks than low-popularity ones (mean rank: high 396.50 > low 158.76) (Table 2).

**Table 1.** Results of the Mann-Whitney U test for recency and popularity

	Recency	Popularity
Mann-Whitney U	18,145	704.000
Z	-15.366	-17.889
Asymp. Sig. (2-tailed)	.000	.000

**Table 2.** Mann-Whitney U test statistics for recency and popularity

	Groups	N	Mean rank	Sum of ranks
Recency	High	698	513.50	358,426.00
	Low	190	191.00	36,290.00
	Total	888		
Popularity	High	176	396.50	69,784.00
	Low	312	158.76	49,532.00
	Total	488		



## 5 Discussion and Conclusions

Reading news online has become an integral part of modern life. The ever-increasing amount of available news and the fragmentation of users' time engender a great challenge to news providers. A news headline must compete for users' attention and be attractive enough to induce clicks. Attaching peripheral indicators to headlines is one way toward this end. This study shed light on the effects of news recency and popularity on users' clicking behavior. It was found through a clickstream data analysis that the presence of both indicators had an impact on the selection of news headlines. The higher the recency or popularity, the more the clicks a headline would attract. The significance of this study not only consists in its implications for news providers in creating effective news headlines and in publishing and disseminating news more responsibly, but also the introduction of clickstream data that is unobtrusive and more reliable in reflecting real behavior.

### 5.1 Implications for News Providers

The peripheral indicators are important visual cues that indicate news salience. They help users make quick judgments about the headlines before exploring their semantic content. This study echoed previous studies in terms of the findings that more recent or more popular news headlines received more clicks. Recency has been a major factor affecting the perceived value of information, especially news [5]. A low frequency of updating may decrease users' trust in information [10]. The pursuit of popular information has its psychological root. The "bandwagon effect" suggests that individuals tend to believe the information when many other people also believe it [6].

Nevertheless, the researchers have a concern about the potential risks of misusing users' preferences for recent and popular news. On the one hand, news providers nowadays strive to win attention by updating the news almost every minute, and some even make use of computer-written news articles to do so [23]. Valuable news may be mixed with reproduced stories, clickbaits, and advertisements, etc. It is undesirable for users to be addicted to or waste time on such "news" despite its high recency. On the other hand, following popular behavior may save people's efforts to make their own judgement, but it may also lock them in information cocoons [24]. If popular news contains bias, the negative effects of the bias can be strengthened as the news gets more popular. This is detrimental to the diversification of opinions, knowledge, and interests.

### 5.2 The Advantages of Unobtrusive Data

Most existing studies depended on self-report methods and experiments to investigate users' information behavior. They are typical obtrusive methods of data collection [25, 26]. The participants of surveys or interviews may not provide authentic, accurate, and complete information for various reasons. The problem of reactivity is even more obvious in experiments. The researchers may introduce consciously or unconsciously their own bias to the experiment design, and the participants may behave not as usual due to the pressure of being observed, artificial tasks, and the environment.

This study introduced two types of unobtrusive data, i.e. the clickstream data from the Web server and the data crawled from the homepage. The collecting of both types of data did not intervene the occurrence of users' clicking behavior. Therefore, the clickstream data analysis was able to provide the most reliable information about the clicks each headline received. Another advantage of the unobtrusive method is that it could achieve considerably larger data size, which made it easier to detect trends. It will be very difficult for obtrusive methods to record such a huge volume of clicks. With the rise of big data, the field of information behavior should consider making use of a wide variety of trace data generated as a result of users' interaction with the Internet. However, the generalizability of the findings of this study could be enhanced if clickstream data were collected from mainstream news sites. It should also be mentioned that clickstream data analysis only reveals users' behavioral patterns without considering what shapes their behavior.

### 5.3 Future Research

The current study focused on the recency and popularity indicators which were in essence external to the news headlines. When it comes to the internal elements of the headlines, their influences on users' attention and behavior should be even stronger. The researchers plan to extract more information from the crawled data, such as the text lengths of the headlines, and the use of numbers and punctuation marks in the headlines. It is interesting to explore whether differences in these characteristic dimensions will arouse differences in clicks. It is also desirable to increase the size of the clickstream data in order to enhance the validity of the analysis of user behavior. In addition, traditional obtrusive methods, e.g. surveys and interviews, will be introduced to complement the clickstream data analysis for the purpose of revealing users' motivations.

**Acknowledgement.** This research has been made possible through the financial support of the National Natural Science Foundation of China under Grants No. 71774125 and No. 71420107026.

## References

1. Holmqvist, K., Holsanova, J., Barthelson, M., et al.: Reading or scanning? a study of newspaper and net paper reading. In: Hyona, J.R., Deubel, H. (eds.) *The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research*, pp. 657–670. Elsevier, Oxford (2003)
2. Dor, D.: On newspaper headlines as relevance optimizers. *J. Pragmat.* **35**(5), 695–721 (2003)
3. Kruikemeier, S., Lecheler, S., Ming, M.B.: Learning from news on different media platforms: an eye-tracking experiment. *Polit. Commun.* **35**(1), 75–96 (2017)
4. Kuiken, J., Schuth, A., Spitters, M., et al.: Effective headlines of newspaper articles in a digital environment. *Digit. Journalism* **5**, 1300–1314 (2017)
5. Xu, Q.: Social recommendation, source credibility and recency: effects of news cues in a social bookmarking website. *Journalism Mass Commun. Q.* **90**(4), 757–775 (2013)
6. Lee, J.Y., Sundar, S.S.: To tweet or to retweet? that is the question for health professionals on Twitter. *Health Commun.* **28**(5), 509–524 (2013)

7. Knoblochwesterwick, S., Sharma, N., Hansen, D.L., et al.: Impact of popularity indications on readers' selective exposure to online news. *J. Broadcast. Electron. Media* **49**(3), 296–313 (2005)
8. Sundar, S.S., Knobloch-Westerwick, S., Hastall, M.R.: News cues: information scent and cognitive heuristics. *J. Assoc. Inf. Sci. Technol.* **58**(3), 366–378 (2007)
9. Metzger, M.J.: Making sense of credibility on the Web: models for evaluating online information and recommendations for future research. *J. Assoc. Inf. Sci. Technol.* **58**(13), 2078–2091 (2007)
10. Westerman, D., Spence, P.R., Heide, B.V.D.: Social media as information source: recency of updates and credibility of information. *J. Comput. Mediated Commun.* **19**(2), 171–183 (2014)
11. Cheung, R.: The influence of electronic word-of-mouth on information adoption in online customer communities. *Global Econ. Rev.* **43**(1), 42–57 (2014)
12. Wang, S.M., Lin, C.C.: The effect of social influence on the bloggers' usage intention. *Online Inf. Rev.* **35**(1), 50–65 (2011)
13. Chang, Y.T., Yu, H., Lu, H.P.: Persuasive messages, popularity cohesion, and message diffusion in social media marketing. *J. Bus. Res.* **68**(4), 777–782 (2015)
14. Chin, C.Y., Lu, H.P., Wu, C.M.: Facebook users' motivation for clicking the "like" button. *Soc. Behav. Pers. Int. J.* **43**(4), 579–592 (2015)
15. Yi, C., Jiang, Z., Zhou, M.: The effects of social popularity and deal scarcity at different stages of online shopping. In: *Thirty Fifth International Conference on Information Systems*, pp. 1–16. AIS eLibrary, Auckland (2014)
16. Mou, J., Shin, D.: Effects of social popularity and time scarcity on online consumer behavior regarding smart healthcare products: an eye-tracking approach. *Comput. Hum. Behav.* **78**, 74–89 (2017)
17. Ksiazek, T.B., Peer, L., Lessard, K.: User engagement with online news: Conceptualizing interactivity and exploring the relationship between online news videos and user comments. *New Media Soc.* **18**(3), 261–270 (2016)
18. Jiang, T., Xu, Y.P., Guo, Q.: Review of clickstream data analysis and visualization studies. *J. China Soc. Sci. Tech. Inf.* **37**(4), 436–450 (2018)
19. Jiang, T.: Characterizing and evaluating users' information seeking behavior in social tagging systems. Dissertation (2010)
20. Jiang, T.: A clickstream data analysis of users' information seeking modes in social tagging systems. In: *iConference 2014 Proceedings*, Berlin, pp. 314–329 (2014)
21. Jiang, T., Chi, Y., Gao, H.Q.: A clickstream data analysis of Chinese academic library OPAC users' information behavior. *Libr. Inf. Sci. Res.* **39**(3), 213–223 (2017)
22. Boczkowski, P., Mitchelstein, E., Matassi, M.: Incidental news: how young people consume news on social media. In: *Hawaii International Conference on System Sciences*, Hawaii (2017)
23. van der Kaa, H.A.J., Krahmer, E.J.: Journalist versus news consumer: the perceived credibility of machine written news. In: *The Computation and Journalism Symposium*, New York (2014)
24. Zuiderveen Borgesius, F.J., Trilling, D., Moeller, J., Bodó, B., De Vreese, C.H., Helberger, N.: Should we worry about filter bubbles? *Internet Policy Rev.* **5**(1), 1–16 (2016)
25. Jiang, T., Zhang, C., Li, Z., et al.: Information encountering on social Q&A sites: a diary study of the process. In: *International Conference on Information*. Sheffield, UK (2018)
26. Makri, S., Bhuiya, J., Carthy, J., et al.: Observing serendipity in digital information environments. In: *Proceedings of the 78th ASIS&T Annual Meeting*, St. Louis, MO, USA (2015)