How to Compete Online for News Audience: Modeling Words that Attract Clicks

Joon Hee Kim[‡] Amin Mantrach[+] Alejandro Jaimes[*] Alice Oh[‡]

joon.kim@kaist.ac.kr, {amantrac,ajaimes}@yahoo-inc.com, alice.oh@kaist.ac.kr

[‡]Korea Advanced Institute of Science and Technology Daejeon, Republic of Korea

[+]Yahoo Labs Barcelona, Spain

[*]Yahoo Labs New York, United States

ABSTRACT

Headlines are particularly important for online news outlets where there are many similar news stories competing for users' attention. Traditionally, journalists have followed rules-of-thumb and experience to master the art of crafting catchy headlines, but with the valuable resource of large-scale click-through data of online news articles, we can apply quantitative analysis and text mining techniques to acquire an in-depth understanding of headlines. In this paper, we conduct a large-scale analysis and modeling of 150K news articles published over a period of four months on the Yahoo home page. We define a simple method to measure click-value of individual words, and analyze how temporal trends and linguistic attributes affect click-through rate (CTR). We then propose a novel generative model, headline click-based topic model (HCTM), that extends latent Dirichlet allocation (LDA) to reveal the effect of topical context on the click-value of words in headlines. HCTM leverages clicks in aggregate on previously published headlines to identify words for headlines that will generate more clicks in the future. We show that by jointly taking topics and clicks into account we can detect changes in user interests within topics. We evaluate HCTM in two different experimental settings and compare its performance with ALDA (adapted LDA), LDA, and TextRank. The first task, *full headline*, is to retrieve full headline used for a news article given the body of news article. The second task, *good headline*, is to specifically identify words in the headline that have high click values for current news audience. For *full headline* task, our model performs on par with ALDA, a state-of-the art web-page summarization method that utilizes click-through information. For *good headline* task, which is of more practical importance to both individual journalists and online news outlets, our model significantly outperforms all other comparative methods.

KDD '16, August 13-17, 2016, San Francisco, CA, USA

DOI: http://dx.doi.org/10.1145/2939672.2939873

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications - Data mining; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

Keywords

Headline Prediction, Large-scale Analysis, Click-through Rate, Online News Analysis

1. INTRODUCTION

In recent years, a fast decline in print readership, coupled with spectacular growth in online news consumption have resulted in new types of journalism, new distribution methods and sources, as well as new business models. The entire field of journalism is experiencing an unprecedented amount of change and competition, particularly from online media. As news sources have multiplied, so have the number of articles that describe the same news event, and this poses a great challenge for journalists in attracting audiences to their stories. This problem is readily visible, as searching for any newsworthy topic on any given day is likely to yield thousands of results. At the same time, the rise of online social media and their mobile apps adds to this problem by stories being quickly shared and reaching millions of users worldwide. From the individual user's perspective, the number of articles he is exposed to on a daily basis has increased significantly; users can visit multiple news media sites, and each site can potentially host a vast number of articles.

Increases in news production and changes in user behavior have generated significant competition for users' attention, in a marketplace where different headlines "compete" for a user's click (both within a particular page and across media platforms). In many ways this is not new, and Tabloids in particular have historically been the masters of grabbing readers' attention with classic headlines such as "Ford to City: Drop Dead" and "Headless Body in Topless Bar" (a headline which inspired a movie of the same name)[1]. Arguably, the "art of headline writing" is a skill developed by journalists that requires creativity and use of some good ground rules. A good headline summarizes the news article, and at the same time entices the reader to want to read it. Guidelines include, for example, verbs and adverbs are pre-

[1]http://nymag.com/nymetro/news/anniversary/35th/n_8568/

ferred to nouns and adjectives, and verbs of active forms are more effective than verbs of passive form [17].

The combination of a surge in online news production and consumption, real-world datasets of user click behavior, and advanced machine learning techniques, presents a singular opportunity for large-scale data-driven analysis of this art. Good headlines have been historically important in attracting readers, but with online news, the difference between a good and a bad headline for a single article can have important revenue impact, affect the propagation of the story in social media, and result in either growth or decline of readership. Despite the potential and significance of a systematic approach to headlines, there has not been much scientific research on this topic, and journalists still rely on intuition and hand crafted rules of thumb. One possible exception is the Huffington Post, which uses A/B testing to choose some headlines [18],

In this paper, we conduct a large-scale quantitative and machine learning-based study of the relationship between user clicks and words in news headlines. We first present an in-depth analysis of user click-through data on 150K news articles published on the Yahoo front page over a four-month period. The analysis reveals important facets about words in headlines. First, some words significantly induce more clicks than others, which illustrates the need for a new metric for click-through rates of each word. Second, certain classes of words, such as named entities and past tense verbs, attract more clicks than others. Finally, word-level click-through rates rapidly vary over time, as events and topics emerge and dissipate. These results highlight the importance of considering context of news articles in formulating effective headlines, and thus we propose the *Headline Click-based Topic Model* (HCTM) to explicitly model the topical context of words with respect to clicks. HCTM extends the traditional Bayesian topic model, latent Dirichlet allocation (LDA) [6], by taking into account aggregate information of clicks on headlines from previously published news articles. The central idea behind HCTM is that it models "interest" on a given topic/word by aggregating click information on articles recently published (we use the previous week as the time frame), and leverage that to suggest headline word to journalists that could attract more clicks. HCTM models the distribution of click-value of words for each topic, and the distribution of clicks for each view as conjugate distributions (*Beta* and *Binomial* distributions, respectively), and incorporate them into the framework of topic modeling approach.

Our main contributions can be summarized as follows:

- We highlight the importance of analyzing individual words in their context with respect to click-inducing characteristics. We develop a new metric for word-level click-through rate.

- From a quantitative analysis of 150K articles with click information from real users, we show that the news articles' click-through rate is positively strongly correlated with average click-value of words in the headline ($r = 0.882$).

- We show that various attributes of words such as named-entities, and part-of-speech play import roles in the click-value of words. We confirm conventional journalistic wisdom such as verbs are more effective than nouns, and that adverbs are more effective than adjectives [17].

- We introduce a novel generative click-based topic model (HCTM) that utilizes previously published news contents and its corresponding click data, improving performance over state-of-the-art techniques.

- We show that by taking topics together with clicks into account we can detect quick changes in user interest while topics themselves remain stable.

The rest of this paper is organized as follows. In Section 2 we discuss related work. In Section 3, we define a new metric for word-level click-through rate. We then perform quantitative analysis of word-level click-through rate with large-scale news and user click data. In Section 4 we present our proposed model and its inference. In Section 5 we present and discuss our experiments on the headline prediction task. Finally, in Section 6 we conclude the paper.

2. RELATED WORK

We explain the contributions of our research within various domains. First we discuss the common practice of headline writing in journalism and what we can learn from that process. Then we look at previous work on automatic headline generation which is a well-studied field but has a significantly different focus than ours. We then discuss topic modeling, which has been widely used for analyzing unstructured text, and we discuss how our model fits into the topic modeling literature. Lastly, we discuss previous research on predicting which online news articles will be widely read, and how our research question is related to those papers.

2.1 Research on Headlines

Traditional studies in journalism tend to focus on the news value of the newspaper as a whole [24] partly because each newspaper has an audience of stable readers [21], and an important way to increase the readership is to improve its reputation by having a good set of well written articles. Consequently, in journalism research, the wording of headlines has been studied from a stylistic, pragmatic and linguistic perspective (cf. [12, 20]). A methodological shortcoming of these studies is that they are small in scale, covering around 1,000 articles, and they rely on study participants' answers on questionnaires rather than direct observation of readers' behavior.

Research on online media and their headlines requires a different approach. One reason is that unlike newspapers, online news delivery offers a much more convenient way for readers to skim through several headlines and choose what to read from thousands of news outlets, free from any physical constraint of newspapers. Another reason is that user behavior can be directly observed through click-through data. To best of our knowledge, not much research has been conducted for online news headlines. One recent report revealed that the Huffington Post performs A/B testing to choose headlines that get more clicks [18], but there is no scientific research for this heuristic. In this paper, we present a large-scale analysis and a predictive model to learn the patterns of clicks and words in headlines based on observations of actual user clicks.

2.2 Automatic Headline Generation

Many previous studies have looked at automatically generating headlines. For example, [4] presents a statistical model for learning the likelihood of a headline. [27] describes using a Hidden Markov Models for generating informative headlines for English texts, and [25] uses singular value decomposition for a similar task. And [8] describes a system that creates a headline for a newspaper story using linguistically-motivated heuristics. More recently, an external corpus, such as Wikipedia, has been used in headline generation [26]. A widely used method in this line of research is TextRank [15], based on a graph where the nodes are the words of the article and links are weighted with the number of sentences where the connected words co-occur. The PageRank algorithm is then applied on the constructed graph to rank the keywords. We use this state-of-the art method as a comparative baseline in the experimental section.

Though we share the same motivation of finding better headlines, the goal of our work differs from the line of researches discussed here as our model does not generate the full headline phrase, but rather proposes specific words that is relevant to the news content and, if included in the headline, increases the chance of the article being clicked.

2.3 Topic modeling

Topic modeling techniques have acquired a lot of popularity since the seminal work of Blei *et al.* [6]. These models have been extended or applied in various domains such as document summarization [11], recommendation [5] or community detection [14]. As a variant of latent semantic analysis (LSA), ALSA (adapted LSA) utilizes click-through information to improve the quality of web-page summarization [22]. In forming the term vector of each page, weight of each word is boosted by the number of times users click on the page after making a query that contains the word. We apply this method to our experimental setting as illustrated in Section 5.3.1, and build ALDA (adapted LDA) as a comparative method.

2.4 Online news forecasting

In the context of online news popularity prediction, the main goal consists of identifying interesting features that can forecast the future popularity of a news articles [3, 1, 23, 9, 13, 16]. State-of-the-art models extract different features using user generated content (such as the number of clicks, of views, the publication time, number of comments, named entities) to predict the popularity of the news, i.e. the CTR. While not directly related to news forecasting, our problem can be seen as the reverse problem. Indeed, in this paper we design a machine learning algorithm that is able to learn headline words that are likely to generate more clicks by using the CTR as input. In [3] the authors show that named entity is a strong signal for news forecasting. In the same sense, in Section 3.3 we show that the presence of celebrity names (i.e. a category of named entities) as well as linguistic features significantly affect how well headlines attract clicks, which gives a strong encouragement in taking into account of both news contents and click information in predicting attractive words for headlines.

police	bombing	suspects	planned	more	attacks
0.0697	0.0601	0.0740	0.0484	0.0531	0.0515

Table 1: Example of wCTR from headline, "Police: Bombing Suspects Planned More Attacks". Words and their corresponding wCTR values are shown. Some words (suspects, police) are more likely to generate clicks than others (planned, attacks). CTR of this headline is 0.0659

3. HEADLINE ANALYSIS AT WORD LEVEL

In this section, we present the details of the quantitative analysis of headlines, including a new metric for word-level click-through rate ($w\text{CTR}(w,t)$).We also examine how its average value over time ($\overline{w\text{CTR}(w)}$) and daily variability ($\Delta(w)$) could reveal the role that an individual word plays in a headline. We describe our dataset in Section 3.1, and define wCTR in Section 3.2. Then we examine the extent to which word-level click-value influences CTR of headlines in Section 3.3. Finally, we discuss how we use the aforementioned metrics to discover the effect of linguistic attribute of a headline on its CTR in Section 3.4.

3.1 Dataset Description

Our dataset consists of a large set of articles published on the Yahoo homepage (yahoo.com) and their corresponding CTR data. A user visiting the homepage might perform several actions including checking mail, browsing photos, or reading news. We only consider user sessions that contain at least one click on a news article. We take news articles published over a period of four months, from March to June 2013, and extract the number of times each article is presented to users (views), and the number of times it is actually clicked (clicks). We filter out articles viewed less than 100 times, and select a random sample of 150K articles.

3.2 Word-level Click-through Rate

In this section we investigate how individual words in a headline impact the CTR of that headline[2]. More precisely, we hypothesize that each word carries an intrinsic click-value depending on current trends and interest manifested by online users, which is mainly revealed in click information. A widely used method to measure click-value of a news article is CTR. It is defined as

$$CTR(d) = \frac{clicks(d)}{views(d)}$$

where $views(d)$ is the number of times a news article d is shown to users, and $clicks(d)$ is the number of times it is actually clicked.

We propose a new measure: word Click-Through Rate (wCTR) that calculates click-value of individual words from a set of headlines and its associated click information. On a given day, a word w can appear in multiple headlines, and in multiple user sessions.

[2]Note that there is a one-to-one correspondence between news articles and their headlines. Thus one could use the terms "article" and "headline" interchangeably when discussing clicks and views.

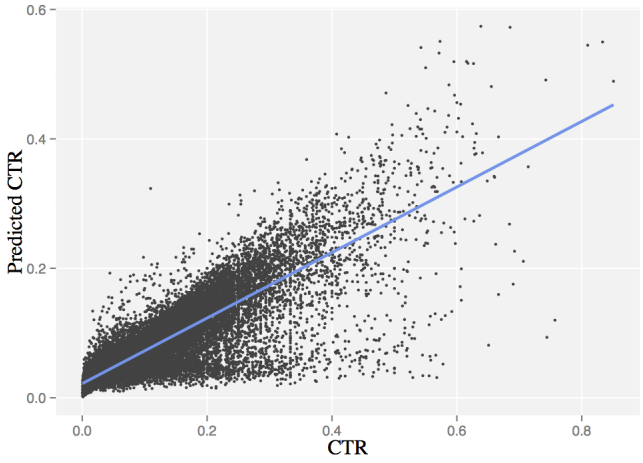

Figure 1: Correlation analysis between CTR and predicted CTR of 150K news articles. Predicted CTR is calculated by averaging wCTR of individual words in each headline. Correlation Coefficient is 0.882

Formally, we define wCTR of word w on day t as

$$wCTR(w,t) = \frac{clicks(w,t)}{views(w,t)}$$

where $views(w,t)$ is the number of times headlines that contain word w are shown to users on day t, and $clicks(w,t)$ is the number of times such headlines are clicked. With this definition, a word with high wCTR value tends to generate more clicks than others. Table 1 gives an actual example of CTR and wCTR for the headline *Police: Bombing Suspects Planned More Attacks*, and it illustrates that individual words in the headline have different wCTR values.

3.3 Correlation Analysis

We verify the extent to which click-value of individual words in the headline could explain the variability of CTR of the article. To do so, we calculate predicted CTR value of a headline by averaging wCTR of each word in the headline. Correlation analysis reveals that the predicted CTR is positively strongly correlated with CTR ($r = 0.882$, Figure 1).

This expected result validates the assumption that individual words carry certain click-value, and they have a strong influence on the popularity of the news article. This encourages us to develop an unsupervised statistical method that learns from recent news articles and click history, and models click-value of individual words based on the context they are used in.

3.4 Effect of Linguistic Attributes

We discover interesting groups of words by analyzing temporal patterns of wCTR value. At each day, wCTR of each word is computed from the news articles published on that day. Then, we compute the mean of wCTR for each word ($\overline{w\text{CTR}(w)}$) on the four month news data as well as its average daily variability ($\Delta(w)$) calculated as the following:

High $\overline{w\text{CTR}(w)}$		Low $\overline{w\text{CTR}(w)}$	
Amanda Bynes	0.171	Rise	0.031
Bynes	0.164	Natural Gas	0.033
Dress	0.120	Gas	0.033
Selena Gomez	0.111	Energy	0.034
Selena	0.111	Kings	0.034
Bump	0.111	Power	0.034
Bomb Suspect	0.108	Dow	0.034
Lindsay Lohan	0.106	Ratings	0.034
Kardashian	0.106	Shares	0.034
Kanye	0.105	Sales	0.035

Table 2: Words with highest (left) and lowest (right) $\overline{w\text{CTR}(w)}$. Words with high mean wCTR are related to celebrity names, and words with low mean wCTR are related to economic issues.

	$\Delta(w)$	$\overline{w\text{CTR}}(w)$
for	0.0038	0.0546
s	0.0043	0.0593
of	0.0043	0.0582
in	0.0044	0.0590
to	0.0045	0.0548
the	0.0050	0.0581
a	0.0071	0.0588
with	0.0074	0.0581

Table 3: Words in ascending order of daily variability of wCTR, and their respective mean value $\overline{w\text{CTR}(w)}$. Words with low daily variability of wCTR value are function words. Their mean wCTR values are close to global average (0.0571).

$$\overline{w\text{CTR}(w)} = \frac{1}{T}\sum_{i=1}^{T} w\text{CTR}(w,i)$$

$$\Delta(w) = \frac{1}{T-1}\sum_{i=1}^{T-1}(w\text{CTR}(w,i+1) - w\text{CTR}(w,i))^2$$

where T is the number of total days, and $w\text{CTR}(w,i)$ is the wCTR of the term w computed exclusively on data published the day i.

By ranking words based on their mean and daily variability of their wCTR value, we observe clusters of words with similar patterns (Table 2, 3). For example, celebrity-related words have high mean click value, whereas business-related words have low mean click value. This suggests that celebrity names attract more clicks when shown to users words related to economic issues. This finding on our dataset confirms recent study on news forecasting where the authors

POS	Tag information	$\overline{w\text{CTR}(w)}$
WP$	Possessive Wh-pronoun	0.0597
WP	Wh-pronoun	0.0513
PRP$	Possessive Pronoun	0.0495
VBD	Verb, past tense	0.0474
VBN	Verb, past participle	0.0468
PRP	Personal Pronoun	0.0467
RBS	Adverb, superlative	0.0454
RB	Adverb	0.0448
JJS	Adjective, superlative	0.0437
WRB	Wh-Adverb	0.0430
DT	Determiner	0.0428
RP	Particle	0.0422
MD	Modal	0.0419
NNP	Proper Noun, singular	0.0415
JJ	Adjective	0.0413
NN	Noun, singular or mass	0.0411
VB	Verb, base form	0.0411
NNPS	Proper Noun, plural	0.0408
NNS	Noun, plural	0.0403
FW	Foreign word	0.0398
VBP	Verb, non-3rd person singular present	0.0393
CD	Cardinal number	0.0390
RBR	Adverb, comparative	0.0382
JJR	Adjective, comparative	0.0368

Table 4: $\overline{w\text{CTR}(w)}$ value of different lexical categories computed on Yahoo news corpus across four months peroiod.

showed that named entities help in predicting popular news articles [3]. Furthermore, interestingly, function words such as preposition, and determiner have very low $\Delta(w)$ value, and their $\overline{w\text{CTR}(w)}$ value is very close to the global average (0.0571), which means that their click value does not change over time, and they have little impact on the headline regardless of the time period or the context they are used in.

Afterwards, we analyze the click value of different lexical categories using part-of-speech tags. The result (Table 4) confirms conventional wisdom on journalism that verbs are more effective than nouns, and adverbs are more effective than adjectives [17]. Another discovery is that superlative adverbs, and adjectives are much more effective than comparative ones in generating clicks.

4. HCTM

In this section, we propose headline click-based topic model (HCTM), a novel generative model that extends latent Dirichlet allocation (LDA) [6]. The rationale for approaching this problem with a topic model is that a close analysis of the data (detailed in Section 3) reveals that each word has an intrinsic click-value (i.e., how likely users clicks on a headline containing that word), and that the click-value is dependent on the context in which the word is used. For example, celebrity names such as "Lionel Messi" or "Cristiano Ronaldo" are certainly more important in a "Sports" article than in a "Business" or "Politics" article.

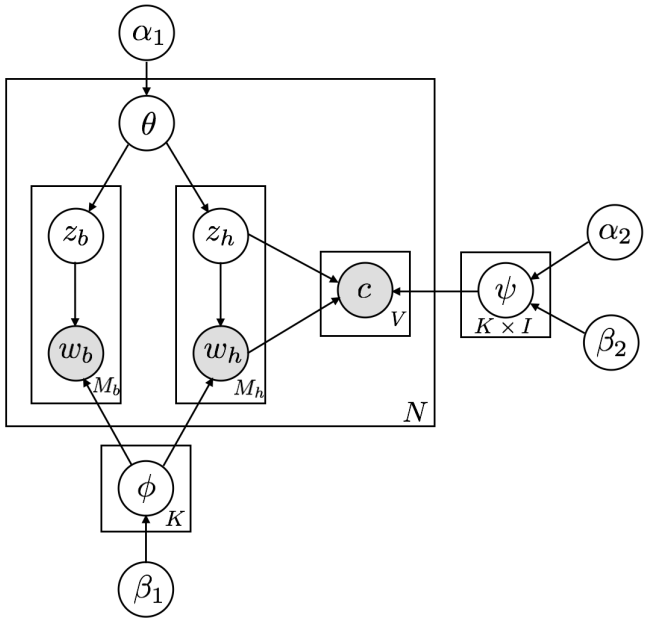

Figure 2: Graphical representation of Headline Click-based Topic Model (HCTM). HCTM is an extension of LDA with two significant changes: 1) an additional observable variable c is introduced to account for whether user click behavior, and 2) z (topic indicator) is split into z_h for headlines and z_b for bodies of articles. This enables realistic modeling of user click behavior, where clicks are generated solely from words and topics of the headline.

4.1 Model Description

HCTM jointly models headline and contents of news article as well as the click data. To analyze user clicks, we consider user *views* of headlines, where a *view* of a headline occurs when the user is presented with the headline on the Webpage, and a *click* occurs if the user actually clicks on the headline. More specifically, as Figure 2 shows, HCTM includes an observable variable c for user clicks where $c_v^d = 1$ if vth user who views headline d clicks on it, and $c_v^d = 0$ otherwise. The model also introduces a latent variable ψ for the topic-specific click-value of each word. Note that in our model the latent indicator variable for topics, typically a single set z, is separated into two, z_h for headline of the news, and z_b for its content (i.e., body of article). Only the former latent variable, z_h, guides the generation of clicks.

HCTM models the distribution of click value, ψ, as a Beta distribution, and the distribution of clicks, c, as a Binomial distribution and take advantage of the Beta-Binomial conjugacy in posterior inference. More precisely, the probability of each word w_h to be clicked in a specific context z_h is modeled by a latent variable c representing a Bernoulli trial. Our prior belief on the probability of a success of this trial is given by a Beta distribution. This completes a full generative Bayesian model.

Figure 2 depicts the graphical model of HCTM, and Table 5 describes the notation of the variables used in the model.

θ	topic distribution of the news article (a multinomial distribution over topics)
ϕ	word distribution of topics (a multinomial distribution over words)
ψ	topic-specific click value of words (a real number between 0 and 1)
z_h	topic of a word in headline
w_h	a word in headline
z_b	topic of a word in body
w_b	a word in body
c	a click (1 for clicked, 0 for not clicked)
N	the number of news articles
M_h	the number of words in headline of the articles
M_b	the number of words in body of the articles
V_d	the number of times the article d is shown to users (number of views)
K	the number of topics
I	the number of unique words

Table 5: List of variables used in the generative model.

A formal description of the generative process is as follows:

1. For each topic $k \in K$,
   (a) Draw word distribution $\phi_k \sim Dir(\beta_1)$
2. For each topic-word pair $(z, w) \in K \times I$,
   (a) Draw click value $\psi_{z,w} \sim Beta(\alpha_2, \beta_2)$
3. For each document $d \in N$,
   (a) Draw topic distribution $\theta_d \sim Dir(\alpha_1)$
   (b) For each word j in headline,
      i. Draw topic $z_h^{jd} \sim Mult(\theta_d)$
      ii. Draw word $w_h^{jd} \sim Mult(\phi_{z_h^{jd}})$
   (c) For each word i in body,
      i. Draw topic $z_b^{id} \sim Mult(\theta_d)$
      ii. Draw word $w_b^{id} \sim Mult(\phi_{z_b^{id}})$
   (d) For each user view $v \in [1, V_d]$,
      i. Draw word w_v^d from headline
      ii. Draw click $c_v^d \sim Bin(1, \psi_{z_v^d, w_v^d})$

4.2 Posterior Inference

In this section, we propose a Markov Chain Monte Carlo algorithm for posterior sampling. More precisely, we use the collapsed Gibbs sampling approach introduced in [10].

The joint probability of the model can be written as the following

$$p(\mathbf{w_b}, \mathbf{w_h}, \mathbf{z}, \mathbf{c}, \theta, \psi, \phi) = p(\phi|\beta_1)p(\psi|\alpha_2, \beta_2) \prod_{d=1}^{N} p(\theta_d|\alpha_1)\Big(\prod_{i=1}^{M_b} p(w_b^{id}|z_b^{id})p(z_b^{id}|\theta_d)\Big) \Big(\prod_{j=1}^{M_h} p(w_h^{jd}|z_h^{jd})p(z_h^{jd}|\theta_d)\Big)\Big(\prod_{v=1}^{V} p(c_v^d|\psi_{z_v^d, w_v^d})p(w_v^d|\psi)\Big)$$

where $p(w_v^d|\psi)$ indicates the probability of a word from the headline to be associated with v'th click. This process is discussed more in detail as we describe the estimation of ψ.

We use Dirichlet-Multinomial conjugacy to write out the conditional distribution of w_b and w_h.

$$p(w_b^{id}|\mathbf{w_b}^{-id}, \mathbf{w_h}, z_b^{id}) = \frac{\sum_{i'd' \neq id} 1[z_b^{i'd'} = z_b^{id}, w_b^{i'd'} = w_b^{id}] + \beta_1}{\sum_{i'd' \neq id} 1[z_b^{i'd'} = z_b^{id}] + I\beta_1}$$

$$p(w_h^{jd}|\mathbf{w_h}^{-jd}, \mathbf{w_b}, z_h^{jd}) = \frac{\sum_{j'd' \neq jd} 1[z_h^{j'd'} = z_h^{jd}, w_h^{j'd'} = w_h^{jd}] + \beta_1}{\sum_{j'd' \neq jd} 1[z_h^{j'd'} = z_h^{jd}] + I\beta_1}$$

Sampling $\mathbf{z_b}$ The conditional distribution of z_b^{id} given word w_b^{id} is proportional to the number of times the topic is used in the document d multiplied by the conditional probability of w_b^{id} given the topic.

$$p(z_b^{id} = z|rest) \propto (n_{zd}^{-id} + \alpha_1) \times p(w_b^{id}|\mathbf{w_b}^{-id}, \mathbf{w_h}, z_b^{id} = z)$$

where n_{zd}^{-id} indicates the number of times topic z is assigned in document d without counting z_b^{id}.

Sampling $\mathbf{z_h}$ The conditional distribution of z_h^{jd} given word w_h^{jd} and ψ is proportional to the multiplication of the number of times the topic is used in document d, the conditional probability of word w_h^{jd} given the topic and the likelihood of clicks associated with w_h^{jd}.

$$p(z_h^{jd} = z|rest) \propto (n_{zd}^{-jd} + \alpha_1) \times p(w_h^{jd}|\mathbf{w_h}^{-jd}, \mathbf{w_b}, z_h^{jd} = z) \times \prod_v p(c_v^d|\psi_{z, w_h^{jd}})$$

where the last multiplication is taken over each click v associated with the word w_h^{jd}.

Estimating ψ The posterior sampling of z_h^{jd} involves estimation of click value ψ. First we write out the probability distribution of click variable c_v^d.

$$p(c_v^d|w_v^d, \psi) \sim Bin(1, \psi_{z_v^d w_v^d})$$

where w_v^d is the headline word associated with the click c_v^d and z_v^d is the topic assigned to w_v^d. We associate click variable with a word in headline at each iteration of sampling. For each c_v^d, we draw a word w_v^d from the headline words set $\mathbf{w}_h^d$ with probability proportional to its click value $\psi_{z_v^d w_v^d}$.

We use Beta-Binomial conjugacy to write out the conditional distribution of ψ given observations on clicks, headline words and their topics.

$$\psi_{z,w}|\mathbf{z}, \mathbf{w}, \mathbf{c} \sim Beta(m_{z,w}^1 + \alpha_2, m_{z,w}^0 + \beta_2)$$

where $m_{z,w}^1$ is the number of times click variable c associated with topic z and word w is observed to be 1 (clicked), and $m_{z,w}^0$ is the number of times it is observed to be 0 (not clicked).

5. EXPERIMENTS

In this section, we present the results of our analysis at two different levels. On one hand, we show how HCTM can provide further insight on headline formulation by jointly modeling news contents, click information, and topic-specific

Topic 17 (Technology)				Topic 28 (Economic Issues)				Topic 42 (Sports)			
ϕ_{17}		ψ_{17}		ϕ_{28}		ψ_{28}		ϕ_{42}		ψ_{42}	
Week 1	Week 2	Week 1	Week 2	Week 1	Week 2	Week 1	Week 2	Week 1	Week 2	Week 1	Week 2
microsoft	appl	upgrad	ballmer	bank	bank	bloomberg	rio	season	game	angi	punch
appl	googl	siri	failur	ceo	debt	mike	jamaica	team	hit	robben	locker
googl	mobil	loop	laptop	fund	countri	eu	nigeria	final	win	6	reliev
game	microsoft	duty	familiar	board	euro	june	malaysian	coach	inning	covert	victori
xbox	amazon	io	chromebook	financi	bond	center	cite	leagu	seri	psych	resum
technolog	samsung	slate	smallest	mcttest	europ	form	tragic	player	season	castl	suspend
mobil	technolog	taxi	radic	firm	european	auditor	400	game	beat	matt	hamilton
comput	devic	fuxx	threat	sharehold	financi	herbalif	caico	sport	score	scoop	marvel
phone	intel	destroy	effort	capit	itali	iceland	guarante	nba	team	curri	phantom
smartphon	phone	array	malwar	jpmorgan	interest	faith	island	nbc	preview	goal	cam

Table 6: Examples of topics (ϕ) and their respective highest click-value words (ψ) tracked during two consecutive weeks. Topics (ϕ) account for general terms, and stay homogenous over time. On the other hand, high click-value words (ψ) include words that describe specific information (names of people, locations, or special events), and change more drastically over time as they reflect real-time interest of audience.

click-values: we present topics and their respective high-click value words inferred using HCTM, and discuss previously unseen patterns (Section 5.2); Then we evaluate the performance of HCTM in generating headlines and compare it to other approaches (Section 5.3).

5.1 Data Processing

We use the same data set as described in 3.1. In building bag-of-words of each news articles, we generate both unigrams and bigrams from headline and body of the article and perform simple linguistic filtering based on word frequency. Words that occur in more than 10% of the articles are removed as stop words, and words that occur less than five times are removed as noise. Note that bigrams are important in "picking up" important topics or entities that consist of two words (e.g., "Boston Bombing", celebrity names).

5.2 Headline Analysis with HCTM

As with LDA, our model can also be used in unsupervised data analysis. The difference is that our model discovers which headline words attract user clicks (in terms of topic-specific click-value of words) as well as topics from the corpus. We identify and present three topics (ϕ_k) related to *technology, economy, and sports* and their respective high click-value words (ψ_k) tracked during two consecutive weeks. The match between topics in consecutive time period is done by associating each topic of one week to the most similar one from the next week in terms of KL divergence. For topics, we illustrate the top ten words in terms of word likelihood given topic, $p(w_i|\phi_k)$. For click-value, we illustrate the top ten words in terms of topic-specific click value, $\psi_{k,i}$ (see Section 4.1).

The results show two interesting previously unseen patterns with topics and their respective click-values (Table 6). First, topics account for general terms that describe certain thematic category such as sports, or technology whereas high click-value words refer to more specific details such as names of people, locations, or special events. Second, high click-value words change more rapidly than topic words. For instance, company names such as *microsoft*, *apple*, *google* always appear as top words of the technology topic. However, its high click-value words vary significantly with no overlapping words. This illustrates how quickly user interests change over time even within the same topical domain. Our model is capable of accounting for both thematic groups of words and temporal trends of user interest as topics and click-values, respectively.

5.3 Headline Generation Experiments

In this section, we describe how we evaluate the performance of HCTM in generating headlines for given news articles. We compare HCTM with a wide range of methods discussed in Section 5.3.1. We report detailed results measured in terms of area under the ROC Curve (AUC), and mean average precision (MAP).

Data for training and testing are prepared using the following method. News articles and their click information of seven consecutive days are gathered as training data. After training, the model is tested on the data of the following day. Given a test set of news articles without their headlines and click information, each model predicts the words of the headline for each news article. For instance, we train our model on the data from March 1 to March 7, and test on the data of March 8.

5.3.1 Baselines for Comparison

To evaluate the models, trained models are provided with the test data, the news articles without headlines (i.e. the body). Each model measures the headline score of words in the body, and produces a rank-ordered list of words for the headline. Below we describe how each model is trained, and produces the headline score for each word.

- **Baseline (wCTR)** We measure the average wCTR score of each word based on the training data. When a word does not have the score (in case when the word does not appear in any headline from the previous week), an average wCTR score is given. Then we multiply each wCTR score by its term-frequency inverse-document-frequency (tf-idf) within the test news article. The resulting score is used as the headline score of the word for the test article.

- **Graph-based (TextRank)** is a widely used algorithm for text summarization [15]. For each document, we draw a graph of words where each word in the document is represented as a node. An edge is given between two nodes if the corresponding words are found within the window of seven words in the document. We measure the eigenvalue centrality for each node. Each word is given a headline score equal to its centrality.

- **Content-based (LDA)** Topic models such as LSA and LDA have also been widely used for document summarization as they excel in capturing thematic gist of documents [11]. LDA learns topic distributions of each document, and word distributions of each topic. After training, LDA infers topic distributions for the test documents. For each word in the document, we compute the headline score as its posterior probability based on the model. We fixed $T = 30$ as the number of topics and used $\beta = 0.1$ and $\alpha = 50/T$ as suggested in [10]. Formally, the posterior probability of a word j within document d is given as

$$p(w^j|\theta_d) = \prod_k p(w^j|z^k)p(z^k|\theta_d)$$

  where k is iterated over all topics, and θ_d is the topic distribution of document d.

- **Click-based (ALDA)** Adapted Latent Semantic Analysis (ALSA) augments LSA-based document summarization algorithm using user query and click information [22]. Specifically, they update weights of a word in each web-page by the number of times users click on the page after making a query containing the word. We apply this method to LDA such that it fits our experimental setting. When building bag-of-words for a news article, we boost frequency of words that appear in the headline by the number of times the article is clicked. When calculating the headline score of words, we use the same method as in LDA above. As shown in Figure 3, this results in significant improvements over conventional LDA approach.

For HCTM, we used the same value for parameters as suggested for LDA, $\beta_1 = \beta_2 = 0.1$ and $\alpha_1 = \alpha_2 = 50/T$. We also conducted experiments with various number of topics (between 5 and 100) getting similar results. For consistency, we stick to results from using $T = 30$. Headline score of each word is calculated as the posterior probability of the word given each test news article as in LDA. Formally, we compute the headline score for each word j in document d as

$$p(w^j|\theta_d) \propto \prod_k p(w^j|z^k)p(z^k|\theta_d)\psi_{z^k,w^j}$$

5.3.2 Evaluation Metrics

Each model produces a rank-ordered list of headline score of words for each test document. We evaluate its predictive power based on the following two measures. In summing up the result, we take macro average over daily average scores.

- **MAP@k** Mean Average Precision @k is the mean of the average precision computed on the top k words predicted. We computed MAP@5, @10 and @20 as headlines have rarely more than 20 words. For cases where the headline have less than 5, 10 or 20 words, the average is calculated on the number of words in the headline [2]. In all cases, we found same patterns in comparative performances. Thus we only report the result with MAP@10 due to lack of space.

- **AUC** Area Under ROC Curve is widely used to measure performance on binary classification. It takes into account of both true positive rate and false positive rate.

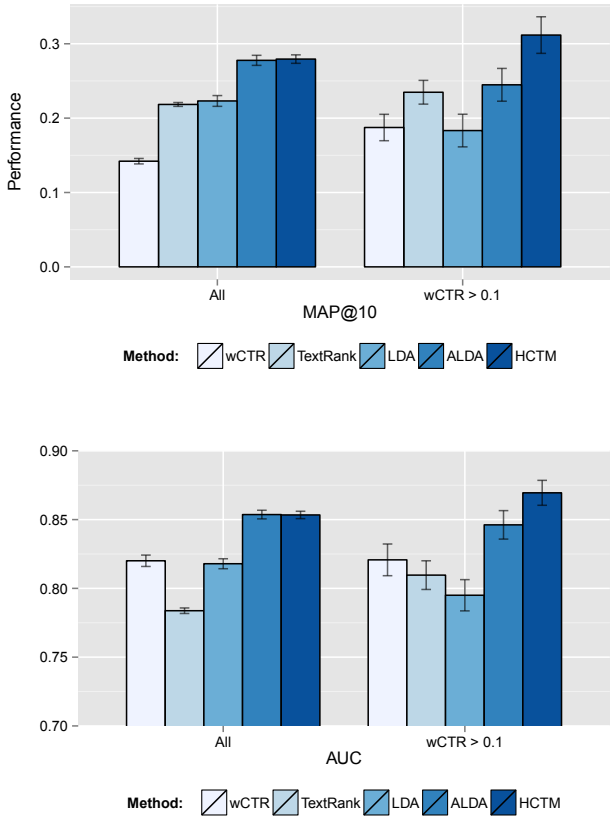

Figure 3: Performance of wCTR, TextRank, LDA, ALDA, and HCTM on the headline prediction task in terms of MAP@10 and AUC. Evaluation is performed with two experimental settings. The first task is to retrieve all headline words (All, left side). HCTM performs on par with the best comparative method, ALDA. The second task is to identify high click-value words in the headline (wCTR > 0.1, right side). HCTM significantly outperforms all comparative methods.

5.3.3 Experiment 1 (Full Headline - All)

The evaluation is performed with two different experimental settings. In the first test, we measure how well each model predicts all words in the headline given only the contents of a news article. Figure 3 summarizes the performance of each method. For this experiment (All, Figure 3

Headline	Suggested words
Obama: Shame on us if we've forgotten Newtown	senate, gun, vote, democrat, support, bill, propose, check, president, republican
Trial over Gulf oil spill set to resume Tuesday	BP, case, drill, rig, district, kill, jury, judge, offshore, defense, Halliburton
Cyprus banks reopen with strict limits on transactions	euro, country, deposit, deal, financial, bailout, official, cash, Laiki, loss, loan

Table 7: Headline suggestion example. Given news content and current headline, HCTM suggests words that could be useful for generating more clicks considering topical context of the article as well as temporal trends of latest click-through behavior. This could potentially be used in online journalism.

left side), performance of our model is on par with ALDA, a state-of-the-art summarization algorithm that is able to utilize click-through information.

5.3.4 Experiment 2 (Good Headline - wCTR > 0.1)

Even within a single headline, some words are more eye-catching than others. Identifying headline words that have high click-value is of greater importance as they attract more clicks. Therefore, in the second test, the objective is to identify headline words that have high wCTR value for current news audience. Specifically, we evaluate each model based on how well it predicts headline words whose wCTR value is higher than 0.1 (measured within unseen test data) which is approximately equivalent to top 10% of all vocabulary.

In this experiment (wCTR > 0.1, Figure 3 right side), our model significantly outperforms ALDA as well as all other comparative methods in terms of both MAP and AUC. This illustrates that our model is able to jointly model topics and click information of news articles in addition to identify topic-specific click-value of each word in the corpus. As a result, our model predicts headline words of a given news article that not only well represents thematic gist of the contents, but also triggers user clicks.

5.4 Discussion

Towards a read-world application The work presented here is a nice example of how social community preferences can be automatically used to suggest better headlines. In practice, the proposed model will be used to suggest new words for a news article for which editors have already proposed an headline. In that scenario, we can suggest to the editor the top words not already in the headline ranked by their posterior probability given by the model as shown in Table 7. Editors will have better understanding of real-time interests of news audience, and learn click-inducing words that are contextually appropriate. Also, we can assess how good the words used by the editors are for the headline given the model trained on last week data, henceforth capturing the current trend. In this way, we believe that such a tool may be very useful in any editorial platform such as WordPress that integrates for instance an A/B testing package [19].

Using more user generated content The study conducted in this paper have been restricted to study the impact of words (unigrams and bigrams) on the CTR, and how the user implicit feedback on the news platform can be used to improve the headline. However, other related studies, on news popularity forecasting have shown that other signals, mostly extracted from user generated content, can be used as well [3, 1, 23]. As discussed in the related work section, the task of headline generation using the CTR, is closely related to predicting the CTR of news articles, and therefore we strongly believe that enriching our model with input signals such as: comments, shares on Facebook, shares on twitter, could improve significantly the quality of the suggested headline. We leave this task as further improvement of our model.

Influence of personalization algorithms It is worth to notice that the user feedback information suffers from a *personalization bias*. Indeed, on the Yahoo front page, a personalization algorithm is used to display the most relevant articles for the user. This ranking depends on user preferences and therefore different users may have different ranking which can lead to a *position bias*. However, the ranking is mostly influenced by the time of publication (i.e. *recency*) and the popularity of the news article (CTR) which does not depend on the user. Furthermore, even if there is a personalization algorithm introducing a position bias, it remains that a click indicates a positive feedback. This is confirmed by our experiment where we show the superiority of using our model over LDA not exploiting the click information.

6. CONCLUSION

In this work, we introduced a novel generative click based topic model for headline generation (HCTM). It extends LDA by taking into account clicks generated by users when presented with a list of headlines on a online news portal. We conducted a large-scale study on a sample of 150K news articles published during four months, on which we showed that current articles' CTR is positively strongly correlated ($r = 882$) with average click-value of individual words in the headline. We also found that various aspects of words such as named-entities, and part-of-speech play important roles in click-through rates, confirming traditional wisdom in journalism, as well as finding novel patterns. We also observed that click value of words change rapidly even within the same domain of topics. By using HCTM, we showed that we can detect topics and their respective high click-value words. Finally, on a headline generation task, using HCTM, we obtain results as competitive as ALDA. More importantly, our model outperforms all other competing models (i.e., ALDA, LDA, and TextRank) in generating high click-value headline words for news articles.

7. REFERENCES

[1] M. Ahmed, S. Spagna, F. Huici, and S. Niccolini. A peek into the future: Predicting the evolution of popularity in user generated content. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, pages 607–616, New York, NY, USA, 2013.

[2] R. A. Baeza-Yates and B. A. Ribeiro-Neto. *Modern Information Retrieval - the concepts and technology behind search, Second edition*. Pearson Education Ltd., Harlow, England, 2011.

[3] R. Bandari, S. Asur, and B. A. Huberman. The pulse of news in social media: Forecasting popularity. In *ICWSM*, 2012.

[4] M. Banko, V. O. Mittal, and M. J. Witbrock. Headline generation based on statistical translation. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 318–325. Association for Computational Linguistics, 2000.

[5] N. Barbieri and G. Manco. An analysis of probabilistic methods for top-n recommendation in collaborative filtering. In *Machine Learning and Knowledge Discovery in Databases*, pages 172–187. Springer, 2011.

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

[7] C. Castillo, M. El-Haddad, J. Pfeffer, and M. Stempeck. Characterizing the life cycle of online news stories using social media reactions. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 211–223. ACM, 2014.

[8] B. Dorr, D. Zajic, and R. Schwartz. Hedge trimmer: A parse-and-trim approach to headline generation. In *Proceedings of the HLT-NAACL 03 on Text summarization workshop-Volume 5*, pages 1–8. Association for Computational Linguistics, 2003.

[9] T. Graepel, J. Q. Candela, T. Borchert, and R. Herbrich. Web-scale bayesian click-through rate prediction for sponsored search advertising in microsoft's bing search engine. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 13–20, 2010.

[10] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences of the United States of America*, 101(Suppl 1):5228–5235, 2004.

[11] A. Haghighi and L. Vanderwende. Exploring content models for multi-document summarization. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 362–370. Association for Computational Linguistics, 2009.

[12] E. Ifantidou. Newspaper headlines and relevance: Ad hoc concepts in ad hoc contexts. *Journal of Pragmatics*, 41(4):699–720, 2009.

[13] A. C. König, M. Gamon, and Q. Wu. Click-through prediction for news queries. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 347–354. ACM, 2009.

[14] Y. Liu, A. Niculescu-Mizil, and W. Gryc. Topic-link lda: joint models of topic and author community. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 665–672. ACM, 2009.

[15] R. Mihalcea and P. Tarau. Textrank: Bringing order into texts. In *Proceedings of EMNLP*, volume 4, pages 275–283. Barcelona, Spain, 2004.

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

[17] S. Saxena. *Headline writing*. Sage, 2006.

[18] Z. M. Seward. How the huffington post uses real-time testing to write better headlines. http://www.niemanlab.org/2009/10/how-the-huffington-post-uses-real-time-testing-to-write-better-headlines/.

[19] Z. M. Seward. A/b testing for headlines: Now available for wordpress. http://www.niemanlab.org/2010/11/ab-testing-for-headlines-now-available-for-wordpress/, Nov. 2010.

[20] J.-S. Shie. Metaphors and metonymies in new york times and times supplement news headlines. *Journal of Pragmatics*, 43(5):1318–1334, 2011.

[21] A. V. Stavros. *Advances in Communications and Media Research*, volume 2. Nova Publishers, 2002.

[22] J.-T. Sun, D. Shen, H.-J. Zeng, Q. Yang, Y. Lu, and Z. Chen. Web-page summarization using clickthrough data. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 194–201. ACM, 2005.

[23] G. Szabo and B. A. Huberman. Predicting the popularity of online content. *Commun. ACM*, 53(8), Aug. 2010.

[24] T. A. Van Dijk. *News as discourse*. Lawrence Erlbaum Associates, Inc, 1988.

[25] S. Wan, M. Dras, C. Paris, and R. Dale. Using thematic information in statistical headline generation. In *Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering-Volume 12*, pages 11–20. Association for Computational Linguistics, 2003.

[26] S. Xu, S. Yang, and F. C.-M. Lau. Keyword extraction and headline generation using novel word features. In *AAAI*, 2010.

[27] D. Zajic, B. Dorr, and R. Schwartz. Automatic headline generation for newspaper stories. In *Workshop on Automatic Summarization*, pages 78–85. Citeseer, 2002.