Noname manuscript No.
(will be inserted by the editor)

From Popularity Prediction to Ranking Online News

Alexandru Tatar · Panayotis Antoniadis · Marcelo Dias de Amorim · Serge Fdida

Received: date / Accepted: date

Abstract News articles are an engaging type of online content that captures the attention of a significant amount of Internet users. They are particularly enjoyed by mobile users and massively spread through online social platforms. As a result, there is an increased interest in discovering the articles that will become popular among users. This objective falls under the broad scope of content popularity prediction and has direct implications in the development of new services for online advertisement and content distribution. In this paper, we address the problem of predicting the popularity of news articles based on user comments. We formulate the prediction task as a ranking problem, where the goal is not to infer the precise attention that a content will receive but to accurately rank articles based on their predicted popularity. Using data obtained from two important news sites in France and Netherlands, we analyze the ranking effectiveness of two prediction models. Our results indicate that popularity prediction methods are adequate solutions for this ranking task and could be considered as a valuable alternative for automatic online news ranking.

Keywords Online news · User comments · Ranking · Predictions

Alexandru Tatar
LIP6/CNRS – UPMC Sorbonne Universités
4 Place Jussieu, 75005, Paris, France
E-mail: tatar@npa.lip6.fr

Panayotis Antoniadis
Communication Systems Group – ETH Zurich
35 Gloriastrasse, 8092, Zurich, Switzerland
E-mail: antoniadis@tik.ee.ethz.ch

Marcelo Dias de Amorim
LIP6/CNRS – UPMC Sorbonne Universités
4 Place Jussieu, 75005, Paris, France
E-mail: amorim@npa.lip6.fr

Serge Fdida
LIP6/CNRS – UPMC Sorbonne Universités
4 Place Jussieu, 75005, Paris, France
E-mail: sf@npa.lip6.fr

1 Introduction

The widespread adoption of smart phones and the rise of social networking sites has accelerated the consumption of online news in the latest years. This is a type of content that can be easily produced and has a small size, short lifespan, and low cost – properties that places it in a top position to be consumed on mobile or social sharing platforms. As a result, a significant amount of research has been centered on understanding the world of online news and many research problems have been addressed within this space, such as: tracking the propagation of topics across the web (Leskovec et al. 2009), describing the decay of interest over time (Dezso et al. 2006), detecting online communities (Adamic et al. 2005), and prediction of popularity (Tsagkias et al. 2010). It is the last one, however, that gained most of the research focus both because this problem is very challenging and because of its immediate practical implications. Indeed, predicting the popularity of online content is valuable for different stakeholders: news sites and news aggregators can better highlight the most popular content, online advertisers can propose more profitable monetization strategies, and online readers can filter the huge amount of information more easily.

There are different ways of expressing the notion of popularity. For example, the classical way of defin-

This paper is a significant extension of our previous work, "Ranking news articles based on popularity prediction", published at ASONAM 2012.

ing it is through the click-through rate. However, this information is seldom available to external observers and, when available, it is difficult to estimate the actual number of times that a page was requested by users or due to web crawlers and search engines. Nevertheless, as reading news has become a social experience, there are other metrics that capture readers' interest. These metrics are based on user participation activities such as comments, votes, or shares through social media or email services. In this paper we focus solely on one dimension of the content popularity – comments – and consider the *number of comments* as an implicit indicator of the interest generated by a news article.

Predicting the popularity of news articles is a complex and difficult task and different prediction methods and strategies have been proposed in several recent studies (Lee et al. 2010; Lerman et al. 2010; Szabo et al. 2008; Tsagkias et al. 2010). Most previous efforts have focused on predicting the exact amount of attention that online content will generate at a future moment in time. This information can indeed prove valuable in online advertising, where new revenue models could be designed to charge advertisers for the (future) amount of attention that a content will generate. However, in another practical situation, a news platform may want to use this information to rank news stories in real-time and highlight the most popular ones. For example, imagine an online newspaper that publishes news stories and at random moments of the day it promotes some of its articles on the social networks accounts. The decision of which content to promote can be done by human raters or through an automatic operation that ranks news stories and selects the most important ones.

In this paper, we focus on the latter option, by studying the feasibility of using popularity prediction methods for automatic online news ranking. To this end, we compare the ranking effectiveness of two prediction methods: a linear model on a logarithmic scale and constant scaling model. In order to properly evaluate the ranking performance, we propose a general setting that takes into consideration two important properties of the articles: lifetime and distribution of popularity. We validate the effectiveness of these methods by using two news sources and compare them with various baseline methods and dedicated learning to rank algorithms. As a summary, the main contributions of this work are:

- We analyze two important online news platforms from France and Netherlands and provide valuable insights on how users post comments on news articles. By exploring these data sets we observe that news stories have a very short lifespan and that the volume of comments per article can be described by a power-law distribution.
- In the context of automatic online news ranking we evaluate the ranking effectiveness of two popularity prediction methods and show that a linear model on a logarithmic scale is an effective method for online news ranking.
- We compare the performance of these methods with learning to rank algorithms and show that for this ranking problem, popularity prediction methods could successfully replace more complex ranking algorithms.

The rest of the paper is organized as follows. In Section 2 we give a brief overview of our ranking approach and present the two data collections. We explore the properties of these data sets in Section 3 and describe the evaluation strategy in Section 4. We evaluate the ranking performance of our proposed methods and compare them with several baseline methods (Section 5) and learning to rank algorithms (Section 6). We conclude with a presentation of the related work in Section 7 and present future perspectives in Section 8.

2 Methodology and data sets

2.1 Methodology

We tackle the problem of ranking online news by using a two-phase procedure. The first step consists in understanding two underlying properties of news articles that are relevant to our ranking problem: articles' lifetime and distribution of popularity. Then, based on these observations, we recommend a more rigorous evaluation strategy adapted to the characteristics of news.

The entire ranking process, through model training, article scoring, and actual ordering can add a significant overhead to a system if the number of items is large. News platforms are particularly affected by this problem as a large corpus of articles accumulates over time. But news stories have a short lifetime and only few items continue to catch readers attention over a longer period of time. It is thus important to study the lifetime of articles and use this information to reduce the pool of articles considered for ranking.

The second relevant information for this ranking problem is the distribution of popularity. News articles, similar to most of the content found over the Internet, depict a very skewed distribution of interest. Understanding how readers' attention is distributed between articles can be exploited to conduct a more focused ranking evaluation where a ranking method should be particularly accurate in identifying the top most important articles.

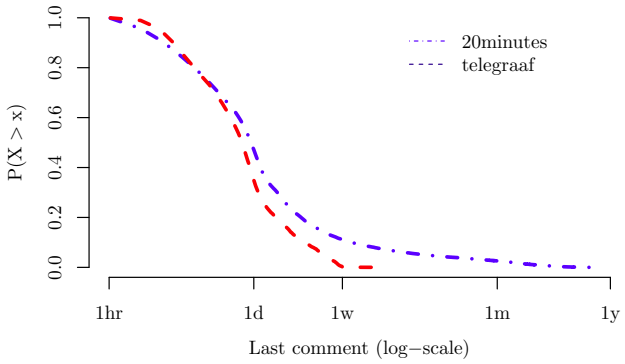
Table 1: Summary of the data sets analyzed in this paper.

Data set	20minutes	telegraaf
Lifespan:		
- start	3/2/2007	18/8/2008
- end	6/5/2011	21/4/2009
Total articles	231,120	40,287
Total comments	2,635,489	731,395
Articles per day		
- mean	157	176
- median	136	153
Comments per day		
- mean	1,255	3,086
- median	1,231	3,052

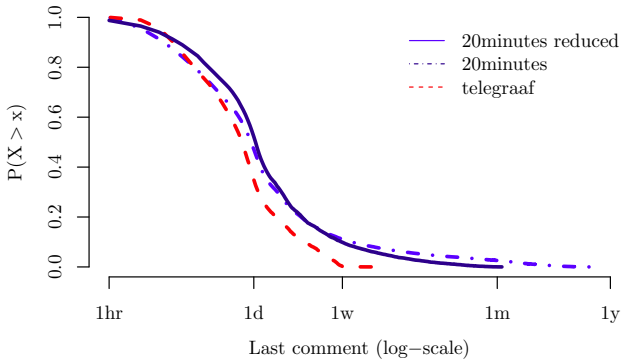
2.2 Data collections

In this study we use data from two news platforms, `20minutes`[1] and `telegraaf`[2]. Both news sources are popular daily newspapers that complement the hard copy editions with online sites that allow users to read news stories and express their opinions through comments. The sites' content is news oriented, starting with the main articles from the printed version and being periodically updated with the latest news. These newspapers target a broad audience and cover diverse topics from national and international politics, sports, economy, or lifestyle.

The two data collections differ in size and lifespan: `20minutes` contains 231,120 articles and 2,635,489 comments published from February 2007 until May 2011 (Tatar et al. 2011); `telegraaf` data set contains 40,287 articles and 731,395 comments published from August 2008 until April 2009 (Tsagkias et al. 2010). We present a summary of the data sets in Table 1.



(a)



(b)

Fig. 1: Complementary cumulative distribution function corresponding to the articles' lifetime (time elapsed between article publication time and the last comment time). The labels on the x-axis correspond to one **h**our, **d**ay, **w**eek, **m**onth, and **y**ear. We represent two versions of `20minutes` data: one over the entire data set and a reduced version that covers the same period of time as `telegraaf` data set.

3 News properties

3.1 Lifetime of an article

A common characteristic of online content is that it suffers from a decay of interest over time and, depending on the type of content, this decay can be steep or gradual. News articles incur a very steep decay compared to videos (Cha et al. 2007) or photos (Cha et al. 2009), as they refer to a recent type of information that by its nature has a very short life cycle (Dezso et al. 2006).

We provide a coarse representation of articles' lifetime by analyzing the timestamp of the last comment received by an article.[3]. The results are presented in Figure 1 by means of a complementary cumulative distribution function of the duration between the publication time of an article and its last comment. For both news sources we observe that the majority of articles (72% for `telegraaf` and 61% for `20minutes`) acquire all comments within the first day after the publication. There are indeed articles that stimulate user interest for a longer period of time, but this interest is sparse and not constant as observed for other type of online content (Cha et al. 2007). This can be seen in Figure 2 by means of a probability density function of the comments publication time relative to articles publication time. As it can be observed, users react very fast to the publication of news articles, but their interest drops quickly after six hours and only a negligible amount of comments are received after one day.

[1] http://www.20minutes.fr/

[2] http://www.telegraaf.nl/

[3] We are aware that there are other fine-grained methods of evaluating the decay of attention over time (Lee et al. 2010; Simkin et al. 2012; Wu et al. 2007), but for the scope of our work, this coarse characterization provides us with sufficient information.

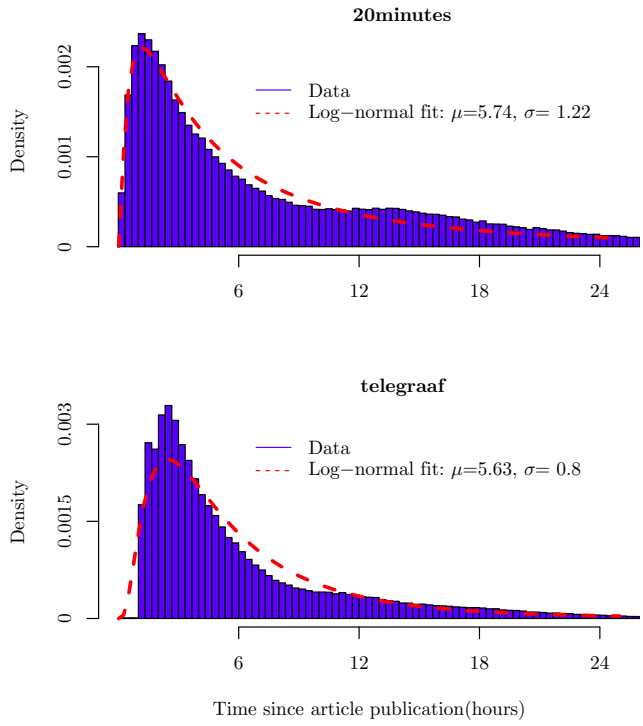


Fig. 2: Probability distribution function of the comments time relative to the articles publication time. We represent the histogram covering a one-day period along with the best probability fit, which in our case is described by a log-normal distribution.

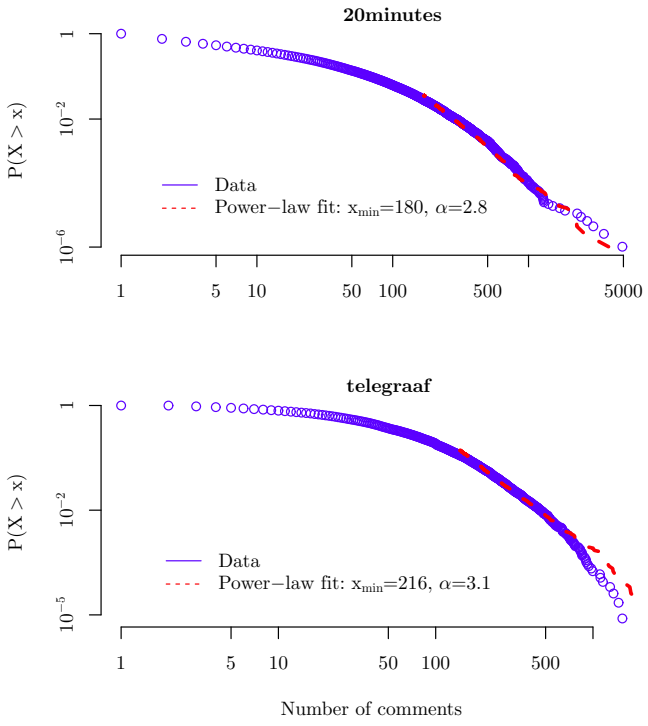


Fig. 3: The complementary cumulative distribution function of the articles' popularity and the corresponding power-law fit.

Comparing the two news sources, we observe that, while the drop of interest over time is similar in the first day for both sites, articles published on `20minutes` engage users in a commenting activity for a longer period of time than those published on `telegraaf`. This difference can be explained by the different lifespan of the data sets, one covering more than four years and the other one only eight months. To isolate this effect we analyze a reduced version of the `20minutes` data set, one that covers the same period of time as `telegraaf` (Figure 1b). Even after this adjustment we can observe that, in general, `20minutes` articles receive comments for a longer period of time than `telegraaf`. There are several factors that could explain this difference. One of them is that `20minutes` news have a greater exposure than `telegraaf` news, as indicated by the traffic statistics of the two web sites (5.5 million unique visitors per month for 20minutes.fr compared to 3.8 million for telegraaf.nl[4]). The result is that `20minutes` articles may seize a greater amount of attention in the early stages after the publication, which could further impact the popularity and smoothen the decay of interest over time. Other explanations, which unfortunately cannot be deduced from the information found in our data sets, could be related to the tone of the articles (a more personal and subjective voice may be more captivating to online readers) or the topic of the news (it has been observed that certain topics have a longer life cycle (Leskovec et al. 2009)).

[4] According to the latest statistics of the two sites: http://corporate.tmg.nl/en/result-second-quarter-2012 (`telegraaf`); http://www.mediametrie.fr (`20minutes`)

3.2 Distribution of popularity

A common question addressed by scientists that study the properties of online content is whether the data under observation exhibits heavy-tail characteristics or not. While this is interesting from a scientific point of view, where a mathematical model can summarize empirical data, this information also has practical implications. For example, it has been shown that understanding the underlying distribution of popularity for web content can have important consequences in the design of caching algorithms (Breslau et al. 1999; Guo et al. 2008) or in the improvement of search engines (Fortunato et al. 2006).

Table 2: Comparing the power-law fit against other alternative distributions. For each alternative distribution, we provide the *p-value* and the likelihood ratio test (LR). We consider a significance level of 0.1 for the *p-value* and display the significant values in bold. Positive values of the log-likelihood indicate that the power-law is a better fit model than the alternative distributions.

Data set	Exponential		Power + cut-off		Log-normal	
	LR	p	LR	p	LR	p
20minutes	34.42	**0.07**	-1.24	0.11	-2.5	0.31
telegraaf	13.40	0.12	-5.6	**0.00**	-4.6	**0.05**

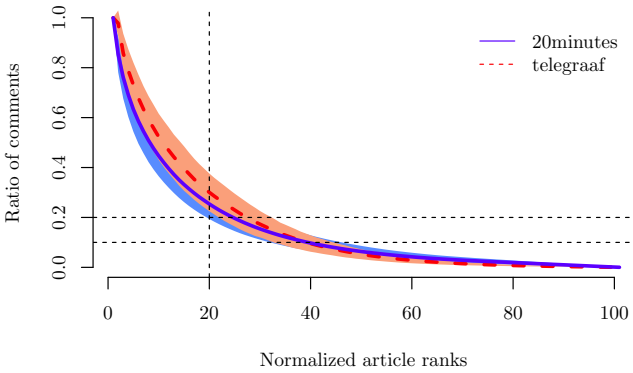


Fig. 4: Normalized article ranks and the cumulative of proportion of comments received on a daily basis. We present the average value and one standard deviation (shaded area).

In the case of social media content, recent work, on different sources of online content and using various popularity metrics, indicates that content popularity can be described by heavy-tail distributions and the log-normal distribution appears to give the most consistent description (Tsagkias et al. 2010; Van Mieghem et al. 2011; Wu et al. 2007). Our data sets make no exceptions from this observation. This can visually be observed in Figure 3, where we illustrate the complementary cumulative distribution of the number of comments per article and the power-law fit. The power-law behavior appears in the tail of the distribution and has been confirmed by rigorous power-law tests proposed by Clauset et al. (Clauset et al. 2007).[5] There is, however, a difference between the two news sources as observed in Table 2. Our results indicate that while a power-law provides the most accurate description for `20minutes` articles, a power-law with exponential cut-off gives a more precise solution for `telegraaf` data set.

It is out of the scope of this paper to debate over which distribution is the most adequate one for describing the popularity of online news and we encourage the reader to follow the enriching discussion presented in (Mitzenmacher et al. 2004). One possible explanation of why the power-law provides a more precise description for `20minutes` articles can be given by the web site recommendation strategy. The site highlights the most commented articles in a dedicated section and twice a day it delivers to its subscribers a short electronic edition with the most commented articles. This creates a *rich-get-richer* effect, which is one of the reasons why power-law appears so often on the Internet (Easley et al. 2010). The recommendation mechanism can also explain why the power-law fails to appear in the beginning of the distribution and could also account for the difference in articles' lifetime observed in Section 3.1. Articles that are unpopular in the beginning do not benefit from any recommendation mechanism and the probability of receiving any kind of attention drops even more as they lose their position on the web site (Simkin et al. 2012).

The heavy-tail property has important implications in the ranking evaluation. Indeed, given that the distribution is so heavily skewed, a ranking algorithm should perform particularly well in identifying the top most important articles. We explore in Figure 4 the daily distribution of comments for the top most commented articles. On the x-axis we order articles based on their popularity (in a decreasing way) and normalize the ranks from 0 to 100. On the y-axis we consider the proportion of daily comments received by the top-k most important articles. As we can observe in Figure 4, for both data sets, on a daily basis the top 10% most commented articles gather 50% of the total number of comments and around 20% of the articles receive 80% of all the comments published that day.

4 Experimental setting

Methodology. To evaluate the ranking performance we propose the following methodology:

1. We break the corpus of articles of each data set in small subsets, where each subset contains all articles published during a certain *period* of time before a specific reference hour h. We set the duration of the *period* to one day given our previous observations of how readers significantly lose their interest in articles after one day.
2. We rank each subset of articles based on the number of comments that articles receive after the reference hour and consider this ranking as the ground truth. We then apply the different methods (heuristics, popularity prediction methods, and learning to rank algorithms) to estimate the ranking of articles

[5] Statistical techniques based on maximum-likelihood methods and Kolmogorov-Smirnov statistics.

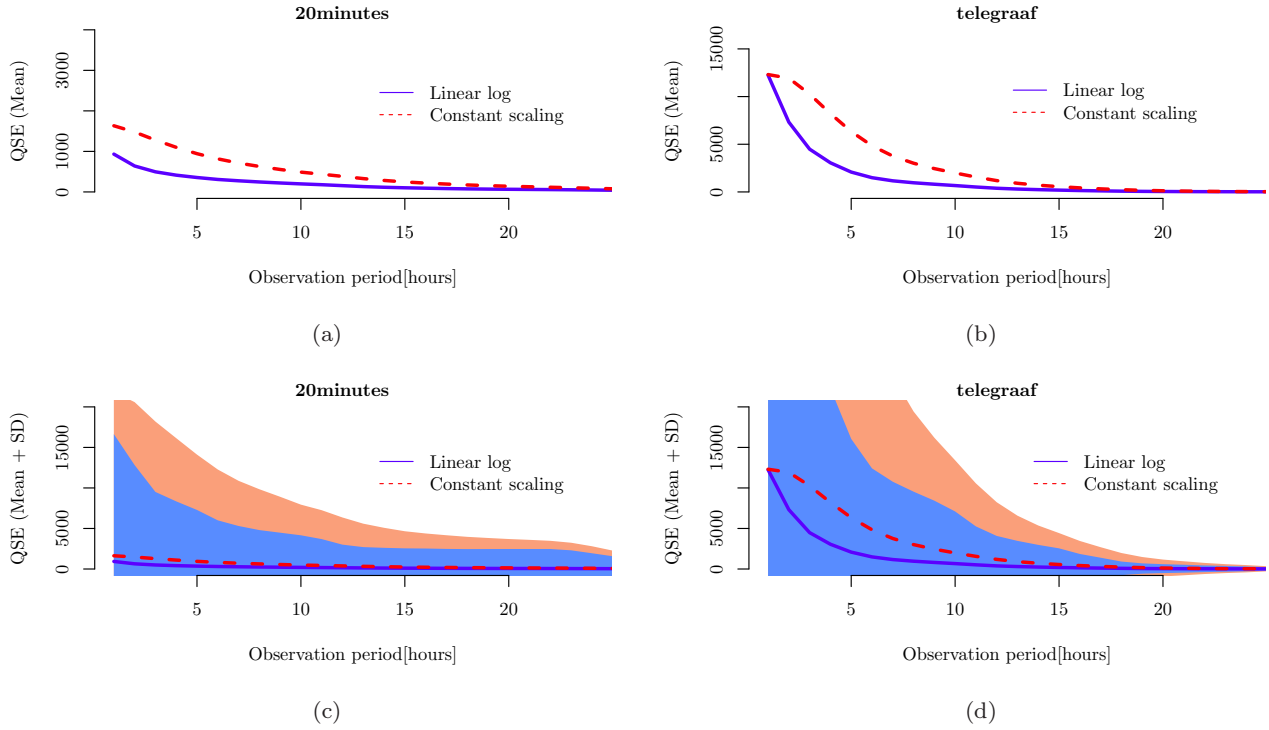


Fig. 5: The prediction error in terms of QSE for the two popularity prediction methods. On the x-axis we vary the observation period from 1 to 24 hours. On the y-axis we represent the mean error - depicted in the top figures - and the mean along with one standard deviation (SD) represented by the shaded area in the bottom figures.

and assess the ranking effectiveness using NDCG evaluation measure.

In the following, we explain the ranking and the evaluation strategies in more detail.

Ranking strategy. Let A be the corpus of articles published by a news platform during a period of time T, with $a \in A$ being one specific article. We discretize time on an hourly basis and consider h a precise hour of the day according to a 24-hour clock. Let t_h be the absolute time in hours and denote d as a one-day period. According to this time description and relative to an hour h we split A in k subsets, with $k = \lceil T/d \rceil$. Denote A_h^i the ith subset of articles created relative to an hour h, with $A = \bigcup_{i=1}^{k} A_h^i$. Please note that as h varies from 0 to 23 there are 24 ways of separating the corpus of articles. This separation allows us to further measure how the ranking performance is influenced by the hour we perform the ranking.

For every article a we refer to a_{t_0} as the article's publication time and define $N_a(t)$ the number of comments received by article a from a_{t_0} to certain time t. We also consider $N_a(t_h, t_r)$ the number of comments received by an article from t_h to t_r.

For our specific ranking task, given a set of articles A_h^i and a ranking time t_h, our goal is to accurately rank articles by the number of comments they will receive from t_h until a future time t_r, with $t_r > t_h$. We set t_r to 30 days to catch only the relevant comments and remove possible sources of spam. Under this description the ground truth ranking for A_h^i is given by $N_a(t_h, t_r)$. We consider this value the relevance of an article, and note

$$rel(a_{t_h,t_r}) = N_a(t_h, t_r). \quad (1)$$

Evaluation measure. We assess the ranking performance of the different strategies using the normal discounted cumulative gain (NDCG) (Jarvelin et al. 2002). To compute NDCG for a set of q articles we first determine DCG as

$$\mathrm{DCG} = rel_1 + \sum_{i=2}^{q} \frac{2^{rel_i} - 1}{\log_2(i+1)}, \quad (2)$$

where rel_i is the relevance of an article found at position i in the ranked list. From this value we compute NDCG as

$$\mathrm{NDCG} = \frac{\mathrm{DCG}}{\mathrm{IDCG}}, \tag{3}$$

where IDCG is the ideal DCG, the DCG of the perfectly ranked list of articles (ground truth ranking). We report the results using 10-fold cross-validation. That is, after splitting the corpus of articles in k subsets we randomly divide these subsets in 10 folds. We use 9 folds to train the models and assess their performance on the remaining fold; we repeat the process 10 times, using a different fold at each step, and report the average value.

5 Ranking methods

Each ranking method rates the relevance of an article using a certain criterion and one method is considered adequate if the estimated ranked list is close to the ground truth ranking. We analyze the ranking effectiveness of two methods based on content popularity prediction and compare them with several baselines.

5.1 Popularity predictions methods

We consider the following two popularity prediction methods:

- Linear regression on a logarithmic scale (**linear log**) model proposed by Szabo and Huberman (Szabo et al. 2008) and previously evaluated on Digg news, YouTube videos, and Dutch news articles (Tsagkias et al. 2010).
- **constant scaling** model also described by Szabo and Huberman and evaluated on Digg news and YouTube videos (Szabo et al. 2008).

The choice of the prediction model is justified by the properties of our data, where the linear model on a logarithmic scale is particularly well adapted to data with heavy-tail characteristics. We also consider the constant scaling model in our analysis following the observations that this model outperforms the *linear log* model when minimizing the relative squared error (Szabo et al. 2008).

These two models are regression functions where the dependent variable is the total number of comments received by an article until time t_r and the independent variable is the number of comments received t_i hours after its publication. The goal of the prediction method is thus to estimate the number of comments t_r hours after an article a is published using the information received in the first t_i hours.

The estimated popularity for the *linear log* model is described by the following equation:

$$\widehat{N}_a^{\mathrm{LN}}(t_i, t_r) = \exp\left(\ln(N_a(t_i)) + \beta_0(t_i, t_r) + \frac{\sigma_0^2(t_i, t_r)}{2}\right). \tag{4}$$

For the parameters of Equation 4, β_0 is computed on the training set using maximum likelihood parameter estimation on the regression function $\ln N_a(t_r) = \beta_0(t_i, t_r) + \ln N_a(t_i)$ and σ_0^2 is the estimate of the variance of the residuals on a logarithmic scale.

The *constant scaling* model is expressed as

$$\widehat{N}_a^{\mathrm{CS}}(t_i, t_r) = \alpha_2(t_i, t_r) \times N_a(t_i), \tag{5}$$

where we estimate α_2 using the following expression:

$$\alpha(t_i, t_r) = \frac{\sum_a \frac{N_a(t_i)}{N_a(t_r)}}{\sum_a \left[\frac{N_a(t_i)}{N_a(t_r)}\right]^2}. \tag{6}$$

We assess the performance of these methods in predicting the exact popularity of the articles using the absolute squared error (QSE):

$$\mathrm{QSE}(a, t_i, t_r) = \left[\widehat{N}_a(t_i, t_r) - N_a(t_r)\right]^2. \tag{7}$$

We analyze the predictive performance of these models as a function of the observation period (t_i) in Figure 5. The results indicate that the prediction error for both models is significantly high for an observation period of less than 6 hours and it rapidly decreases after that. Comparing the two data sets, we observe that `telegraaf` articles have very low predictive performance in the beginning and a negligible one after 20 hours. On the other hand, `20minutes` articles show a better overall predictive performance but the error prevails even after one day. The different performance of these models can, however, be explained by the different dynamics of the comment arrival rate presented in Section 3. As observed in Figure 2, the most significant share of comments is received in the first 6 hours, which explains the high prediction error for short observation periods. Similar, the low error for `telegraaf` news stories after 20 hours is explained by the saturation of articles' popularity in less than one day.

5.2 Baselines

We compare the effectiveness of these methods with three baseline strategies:

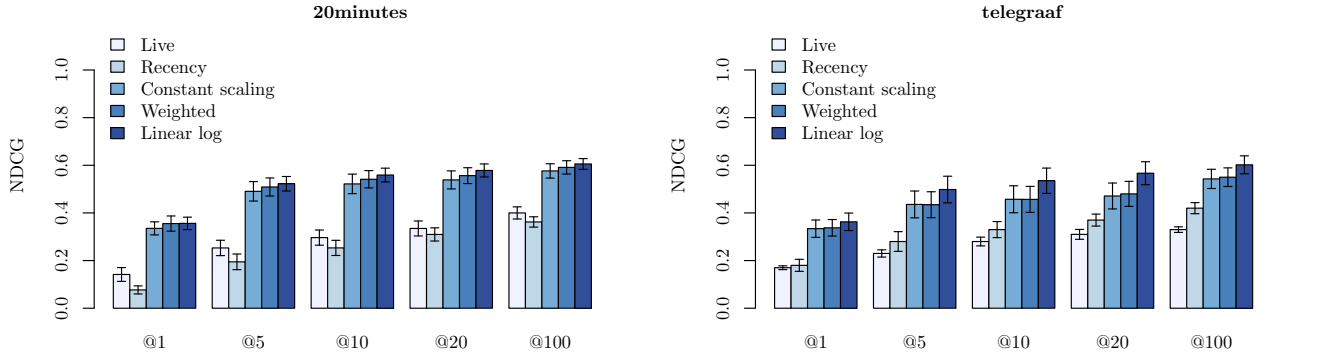


Fig. 6: NDCG at different levels of precision. *@n* corresponds to the NDCG score for the top most important n articles. We present the mean over all prediction hours h (n=24) along with a 95% confidence interval.

– *Live*: rank articles by the number of comments received until the prediction moment, $N_a(t_h)$.
– *Recency*: rank articles by the time of publication, a_{t_0}, with the most recent first.
– *Weighted*: rank articles by the number of comments but weight the volume of comments per hour giving importance to more recent information.

The first two methods are simple heuristics often used by news portals to highlight their popular content, where *live* is oblivious to the temporal information and *recency* considers the time of the publication as the only factor that matters in the ranking decision. The third baseline method is similar[6] to the algorithm proposed by McCreadie et al. that showed one of the most accurate performance on TREC 2009 blog collection (McCreadie et al. 2010). This method combines the partial popularity and recency of articles in the ranking decision by weighting the popularity relative to its closeness to t_h. By using this method, the score S assigned to an article a at time t_h is given by the following formula:

$$S(a_{t_h}) = \sum_{t=a_{t_0}}^{t_h} f(t_h - t) N_a(t) \cdot \tag{8}$$

where f is a probability density function that describes how much weight we should assign to past popularity on an hourly basis. In our case, we observed in Figure 2 that the decay of interest over time follows a log-normal behavior. As a result, we express f as log-normal probability density function:

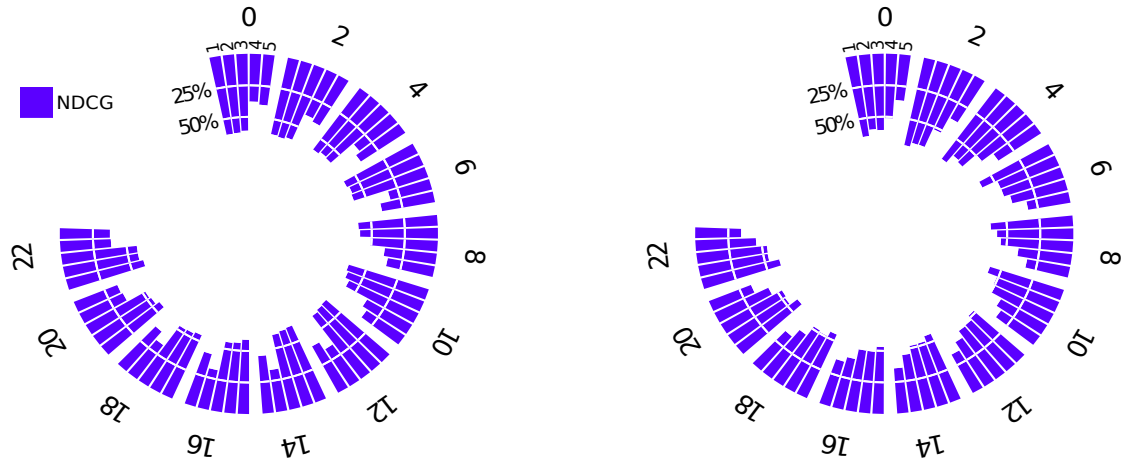
$$f(\delta; \mu, \sigma) = \frac{1}{\delta\sigma\sqrt{2\pi}} \exp\left(\frac{(-\ln(\delta) - \mu)^2}{2\sigma^2}\right), \tag{9}$$

where we obtain the values of μ and σ by fitting the log-normal distribution on the empirical data.

[6] The algorithm uses the number of blog posts to predict users' interest in articles.

5.3 Results

Using the experimental setting described in Section 4 we compare the ranking performance of the two popularity prediction models with the baseline strategies (Figure 6). We report the mean value and a 95% confidence interval over all prediction hours and for various levels of precision: NDCG@1, NDCG@5, NDCG@10, NDCG@20, and NDCG@100. One can observe from the results that the simplest baseline models, *live* and *recency*, have limited ranking capabilities. This suggests that news ranking based on the submission time – *recency* heuristic – or one based on static view of the popularity – *live* heuristic – are inefficient solutions for this ranking task. The performance can however be improved using popularity prediction methods or a *weighted* solution. For a precision level of NDCG@100 (that allows us to capture on average 98% of the daily comments - Figure 4) the *linear log* model shows 50% improvement compared to *live* solution (for both data sets) and a 40% improvement for `telegraaf` - and 75% for `20minutes` - compared to the *recency* solution. From the top three performing algorithms, the *linear log* model shows the overall highest performance; the only exception is observed for NDCG@1, where the weighted model is equally effective. The gain of *linear log* model, compared to the second best solution (*weighted model*) for NDCG@100, is of 2% for `20minutes` and 10% for `telegraaf`. If the benefit brought by the *linear log* model over the other top two models is important for `telegraaf` (with an increase between 10% and 14% for precision levels greater than NDCG@5), for `20minutes` the top three methods show a similar performance suggesting that they are equally fit for this ranking task.



(a) **20minutes** (b) **telegraaf**

Fig. 7: Ranking accuracy in terms of NDCG@100 per hourly basis. The outer numbers correspond to different reference hours h (only the even hours of the day). The inner numbers correspond to the different ranking methods, with 1 - linear log, 2 - weighted, 3 - constant scaling, 4 - recency, 5 - live.

These results depict the average performance over all hours of the day. However, in our previous analysis (Tatar et al. 2011) and other similar studies (Szabo et al. 2008; Tsagkias et al. 2010), it has been observed that articles and comments are published at a different rate during the day. As a consequence, articles may be more popular or exhaust their interest more quickly depending on the publication hour, an effect that can influence the ranking accuracy. To capture the impact of this observation, we illustrate in Figure 7 the ranking performance as a function of different prediction hours (to ease the presentation of the figure we report only the even hours of a day). We take as example the case of NDCG@100, but we observed that the relative performance of the ranking methods is equivalent for the other levels of precision. One can notice that, in general, the top three algorithms show a consistent improvement over the simple heuristics *live* and *recency*. The improvement of the *linear log* model over the other two methods is insignificant for `20minutes` – suggesting that the top three ranking solutions are equally effective – but has an important impact for `telegraaf` data set where the improvement is notable for some specific hours (e.g. the improvement for 10 a.m. is 12%.)

6 Comparison with learning to rank algorithms

A different approach to this ranking problem is to automatically construct a ranking model using learning to rank algorithms. These algorithms propose a straightforward approach to the ranking problem and provide higher adaptability to include more features into the ranking model. We compare our approach with several learning to rank algorithms.

Depending on how they address the ranking problem, there are three main classes of learning to rank algorithms: pointwise, pairwise, and listwise (Liu et al. 2009). We consider a representative model from each category:

- Multiple additive regression trees (MART) - **pointwise** approach based on the gradient boosting technique proposed in (Friedman 2001).
- RankBoost - **pairwise** approach based on a boosting algorithm and multiple weak rankers (Freund et al. 2003).
- LambdaMART - **pairwise** and **listwise** approach using boosted regression trees and designed to optimize NDCG (Wu et al. 2010).
- AdaRank - **listwise** approach also based on a boosting algorithm that minimizes an exponential loss function (Xu et al. 2007).

Using the same evaluation strategy (10-fold cross-validation) we deploy and assess the performance of these algorithms for our specific ranking task.[7] While the format of the previous models is not adapted to be used with a large number of features, this can easily be done using learning to rank algorithms. We thus compare the performance of dedicated learning to rank algorithms using the same amount of information as the previous models, with models that include other features into the ranking decision (e.g. section, author, mean inter-comment time). As a result, we train and

[7] We deploy these algorithms using RankLib open source library (Ranklib).

evaluate these algorithms using two different set of features:

- **basic** set of features: partial popularity, time since publication, publication hour.
- **enhanced** set of features: basic features + (section, author, time of the first comment, mean and median inter-comment time, weekday, and week).[8]

We report the performance of these models in Table 3 and compare them with the best performing model from our previous analysis, the *linear log* model. Overall, one can observe that the *linear log* method is more effective than most of the learning to rank solutions, being surpassed only by the MART model with an enhanced set of features for NDCG@100. From the learning to rank algorithms, MART exhibits effective performances (very close to *linear log* method) across all levels of prediction. This is likely due to the underlying structure of the model that solves the ranking problem through a set of regression trees. Using the basic set of features, the other learning to rank solutions generally do not perform as well as the previous two, which suggest that they are not able to solve the pairwise and listwise constrains for this ranking problem. In general, we observe that adding more features in the model improves the ranking performance except for AdaRank applied to `20minutes` data set, which shows a reduced performance. These results suggest that popularity prediction methods can accurately identify the top most commented articles and could be used as a valuable solution to automatic online news ranking.

Table 3: Ranking accuracy in terms of NDCG for different levels of precision. We compare the *linear log* model and the learning to rank algorithms using different set of features: **b**asic and **e**nhanced. The bold value indicates the best performing algorithm for a specific precision level.

(a) 20minutes

Method	NDCG				
	@1	@5	@10	@20	@100
Linear log	**0.35**	**0.52**	**0.56**	**0.58**	**0.61**
MART- b	0.3	0.48	0.52	0.54	0.57
MART - e	0.33	0.50	0.54	0.56	0.59
RankBoost - b	0.07	0.12	0.13	0.14	0.26
RankBoost - e	0.24	0.36	0.39	0.43	0.48
LambdaMART - b	0.06	0.17	0.22	0.25	0.32
LambdaMART - e	0.05	0.16	0.22	0.26	0.32
AdaRank - b	0.13	0.23	0.28	0.31	0.38
AdaRank - e	0.07	0.19	0.24	0.29	0.35

(b) telegraaf

Method	NDCG				
	@1	@5	@10	@20	@100
Linear log	**0.36**	**0.50**	**0.55**	**0.59**	0.60
MART- b	0.31	0.48	0.52	0.56	0.59
MART - e	0.32	0.49	0.54	**0.59**	**0.61**
RankBoost - b	0.19	0.24	0.28	0.32	0.40
RankBoost - e	0.27	0.46	0.49	0.52	0.56
LambdaMART - b	0.13	0.20	0.24	0.30	0.39
LambdaMART - e	0.14	0.21	0.25	0.29	0.40
AdaRank - b	0.15	0.22	0.24	0.24	0.37
AdaRank - e	0.16	0.26	0.41	0.41	0.51

7 Related Work

Several works have addressed the problem of predicting the popularity of online content. One of the first models, used to predict the popularity of Slashdot stories, was proposed by Kaltenbrunner et al. (2007). This solution considers that, depending on the publication hour, the popularity of news stories follows a constant growth. Szabo et al. proposed two other prediction methods that have shown good results in predicting the popularity of YouTube videos and Digg stories (Szabo et al. 2008). Tsagkias et al. showed that the *linear log* method is also reliable for predicting the popularity of news articles (Tsagkias et al. 2010). Lerman et al. propose a different approach to the prediction problem and present a model built on the social influence and web platform characteristics in the prediction process (Lerman et al. 2010). A different approach was proposed by Lee et al. where, instead of predicting the exact value, the authors are interested in predicting the probability that a content will continue to receive comments after a certain period of time (Lee et al. 2010). More recent results (Bandari et al. 2012), which use the number of tweets as the popularity metric, show that it is possible to classify articles in four classes of popularity, but that it is still difficult to predict the exact amount of attention. We place ourselves in this context of popularity prediction. In our work we analyze the predictive characteristics of news articles, on an unexplored data set (`20minutes`), using methods that have shown good results in previous works. We make a step further in our research and analyze the ranking capabilities of these methods by taking into consideration the dynamic nature of news generation.

The feasibility of ranking online news has been addressed in (Morales et al. 2012; McCreadie et al. 2010). McCreadie et al. propose a ranking method based on relevant blog posts and show that the blogosphere activity is a reliable indicator of news stories importance (McCreadie et al. 2010). A different approach was proposed by Morales et al. (2012) who use a learning to rank algorithm and Twitter posts to rank news articles based on

[8] Information about *section* and *author* are available only for `20minutes` data set.

the future number of clicks. The study shows that micro blogging activity can successfully be used to detect the important news stories. In our study we share the same general objective of ranking news articles, but our work differs both in the ranking technique, notion of article relevance, and input used for the ranking methods.

8 Summary and outlook

In this paper, we analyzed the efficiency of popularity prediction methods in the context of automatic online news ranking. We conducted our study using a large corpus of articles and comments from a French and a Dutch online news platforms and performed an evaluation centered on two fundamental characteristics of online content: the distribution of popularity and the lifetime of articles. In this context, we analyzed the ranking effectiveness of two content popularity prediction methods and compared them with several baselines methods and learning to rank algorithms. Our results indicate that a *linear log* popularity prediction model is an effective solution to online news ranking, with a performance that can evenly match more customized learning to rank algorithms.

The quality of the prediction, even under the most accurate ranking method, shows a moderate performance. One way to boost the ranking performance is to include more information in the ranking decision. During our analysis, we observed that some learning to rank algorithms improve their performance by including more features in the model. The improvement is, nevertheless, modest and other sources of information should be considered in the future work. We believe that an interesting direction would be to study how news articles spread in social networks (Li et al. 2013) or blogs (Cha et al. 2012) and to understand how this process influences articles' popularity. As it has already been observed that the blogosphere (McCreadie et al. 2010) and online social networks (Morales et al. 2012) provide reliable signals for content popularity, an appealing extension of this work is to create a ranking model that puts toghether evidence from all these sources. An additional source of information lies in the profile of the users that comment on news articles. Blogging is often a social activity and user communities may form around certain topics (Macskassy et al. 2011) which can animate the discussions and increase news articles popularity.

The popularity prediction models studied in this paper are designed to predict the popularity at a specific future time and are oblivious to how the popularity is spread over the entire lifetime of an article. While most articles share a similar temporal trend – fast decay of user interest – more complex temporal dynamics have been observed with online content (Crane and Sornette 2008, Leskovec et al. 2009). Uncovering the different temporal evolution patterns could refine the quality of the prediction and further improve the ranking accuracy.

Finally, in future work we will propose strategies that are more adequate for an online evaluation of the ranking methods. In our work, we use an offline evaluation strategy that makes abstraction of how the outcome of the ranking influences the commenting activity. In reality, the prediction outcome, used for content recommendation or front page ordering, may play a critical role in the future popularity of a content (Zhou et al. 2010). Future work should focus on the design of more adequate assessment tools based on an internal feedback loop between the web platform and user reaction to ranking outcome.

9 Acknowledgements

The work presented in this paper has been carried out at LINCS (www.lincs.fr) and was partially supported by the ANR project Crowd under contract ANR-08-VERS-006 and EINS, the Network of Excellence in Internet Science, FP7 grant 28802. The authors are grateful to Manos Tsagkias for sharing the `telegraaf` data set.

References

1. (2012) Crowd project. `http://anr-crowd.lip6.fr/`
2. Adamic LA, Glance N (2005) The political blogosphere and the 2004 us election: divided they blog. In: Proceedings of the 3rd international workshop on Link discovery, ACM, pp 36–43
3. Agarwal N, Liu H, Tang L, Yu PS (2008) Identifying the influential bloggers in a community. In: Proceedings of the international conference on Web search and web data mining, ACM, WSDM '08
4. Bandari R, Asur S, Huberman B (2012) The pulse of news in social media: Forecasting popularity. Arxiv preprint arXiv:12020332
5. Breslau L, Cao P, Fan L, Phillips G, Shenker S (1999) Web caching and zipf-like distributions: Evidence and implications. In: INFOCOM'99. Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE, IEEE, vol 1, pp 126–134
6. Cao Z, Qin T, Liu TY, Tsai MF, Li H (2007) Learning to rank: from pairwise approach to listwise ap-

proach. In: Proceedings of the 24th international conference on Machine learning, ACM, pp 129–136
7. Cha M, Kwak H, Rodriguez P, Ahn Y, Moon S (2007) I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement, ACM, pp 1–14
8. Cha M, Mislove A, Gummadi KP (2009) A measurement-driven analysis of information propagation in the flickr social network. In: Proceedings of the 18th international conference on World wide web, ACM, pp 721–730
9. Cha M, Pérez JAN, Haddadi H (2012) The spread of media content through blogs. Social Network Analysis and Mining 2(3):249–264
10. Christensen LL (2007) The Hands-On Guide for Science Communicators: A Step-by-Step Approach to Public Outreach. Springer
11. Clauset A, Shalizi C, Newman M (2009) Power-law distributions in empirical data. SIAM Rev 51:661–703
12. Crane R, Sornette D (2008) Robust dynamic classes revealed by measuring the response function of a social system. In: Proceedings of the National Academy of Sciences, 105(41):15649–15653
13. Dang V (2012) Ranklib library. `http://people.cs.umass.edu/vdang/ranklib.html`
14. De Francisci Morales G, Gionis A, Lucchese C (2012) From chatter to headlines: harnessing the real-time web for personalized news recommendation. In: Proceedings of the fifth ACM international conference on Web search and data mining, ACM, pp 153–162
15. Dezsö Z, Almaas E, Lukács A, Rácz B, Szakadát I, Barabási AL (2006) Dynamics of information access on the web. Physical Review E 73(6):066,132
16. Easley D, Kleinberg J (2010) Networks, Crowds, and Markets: Reasoning About a Highly Connected World. Cambridge University Press
17. Fortunato S, Flammini A, Menczer F, Vespignani A (2006) Topical interests and the mitigation of search engine bias. Proceedings of the National Academy of Sciences 103(34):12,684–12,689
18. Freund Y, Iyer R, Schapire RE, Singer Y (2003) An efficient boosting algorithm for combining preferences. The Journal of Machine Learning Research 4:933–969
19. Friedman JH (2001) Greedy function approximation: a gradient boosting machine.(english summary). Ann Statist 29(5):1189–1232
20. Guo L, Tan E, Chen S, Xiao Z, Zhang X (2008) The stretched exponential distribution of internet media access patterns. In: Proceedings of the twenty-seventh ACM symposium on Principles of distributed computing, ACM, pp 283–294
21. Hsu C, Khabiri E, Caverlee J (2009) Ranking comments on the social web. In: Computational Science and Engineering, 2009. CSE'09. International Conference on, IEEE, vol 4, pp 90–97
22. Järvelin K, Kekäläinen J (2002) Cumulated gain-based evaluation of ir techniques. ACM Transactions on Information Systems (TOIS) 20(4):422–446
23. Kaltenbrunner A, Gomez V, Lopez V (2007) Description and prediction of slashdot activity. In: Proceedings of the 2007 Latin American Web Conference, IEEE Computer Society, Washington, DC, USA, pp 57–66
24. Kaltenbrunner A, Gómez V, Moghnieh A, Meza R, Blat J, López V (2007) Homogeneous temporal activity patterns in a large online communication space. CoRR URL `http://arxiv.org/abs/0708.1579v1`
25. Lee J, Moon S, Salamatian K (2010) An approach to model and predict the popularity of online contents with explanatory factors. In: Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01, IEEE Computer Society
26. Lerman K, Ghosh R (2010) Information Contagion: an Empirical Study of the Spread of News on Digg and Twitter Social Networks. In: Proceedings of 4th International Conference on Weblogs and Social Media (ICWSM), URL `http://arxiv.org/abs/1003.2664`
27. Lerman K, Hogg T (2010) Using a model of social dynamics to predict popularity of news. In: Proceedings of the 19th international conference on World wide web, ACM, New York, NY, USA, WWW '10, pp 621–630
28. Leskovec J, Backstrom L, Kleinberg J (2009) Meme-tracking and the dynamics of the news cycle. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, KDD '09
29. Li CT, Kuo TT, Ho CT, Hong SC, Lin WS, Lin SD (2013) Modeling and evaluating information propagation in a microblogging social network. Social Network Analysis and Mining pp 1–17
30. Liu TY (2009) Learning to rank for information retrieval. Foundations and Trends in Information Retrieval 3(3):225–331
31. Macskassy SA (2011) Contextual linking behavior of bloggers: leveraging text mining to enable topic-based analysis. Social Network Analysis and Mining

1(4):355–375
32. Manning CD, Raghavan P, Schtze H (2008) Introduction to Information Retrieval. Cambridge University Press, New York, NY, USA
33. McCreadie R, Macdonald C, Ounis I (2010) News article ranking: Leveraging the wisdom of bloggers. In: Adaptivity, Personalization and Fusion of Heterogeneous Information, pp 40–48
34. Metzler D, Bruce Croft W (2007) Linear feature-based models for information retrieval. Information Retrieval 10(3):257–274
35. Mitzenmacher M (2004) A brief history of generative models for power law and lognormal distributions. Internet mathematics 1(2):226–251
36. Paterson CA (2008) Making online news: The ethnography of new media production, vol 1. Peter Lang Pub Incorporated
37. Salganik M, Dodds P, Watts D (2006) Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market. Science 311(5762):854–856
38. Simkin M, Roychowdhury V (2012) Why does attention to web articles fall with time? Arxiv preprint arXiv:12023492
39. Szabo G, Huberman BA (2008) Predicting the popularity of online content. Communications of the ACM 53(8):80
40. Tatar A, Antoniadis P, Limbourg A, de Amorim MD, Leguay J, Fdida S (2011) Predicting the popularity of online articles based on user comments. In: WIMS'11, ACM, pp 67–75
41. Tsagkias M, Weerkamp W, De Rijke M (2009) Predicting the volume of comments on online news stories. In: Proceeding of the 18th ACM conference on Information and knowledge management, ACM, pp 1765–1768
42. Tsagkias M, Weerkamp W, De Rijke M (2010) News comments: Exploring, modeling, and online prediction. In: Proceedings of the 32nd European conference on Advances in Information Retrieval, Springer, ECIR2010
43. Tsagkias M, de Rijke M, Weerkamp W (2011) Linking online news and social media. In: Proceedings of the fourth ACM international conference on Web search and data mining, ACM, WSDM '11
44. Van Mieghem P, Blenn N, Doerr C (2011) Lognormal distribution in the digg online social network. Eur Phys J B 83:251–261
45. Wu F, Huberman B (2007) Novelty and collective attention. Proceedings of the National Academy of Sciences 104(45):17,599
46. Wu Q, Burges CJ, Svore KM, Gao J (2010) Adapting boosting for information retrieval measures. Information Retrieval 13(3):254–270
47. Xu J, Li H (2007) Adarank: a boosting algorithm for information retrieval. In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, pp 391–398
48. Yin P, Luo P, Wang M, Lee WC (2012) A straw shows which way the wind blows: ranking potentially popular items from early votes. In: Proceedings of the fifth ACM international conference on Web search and data mining, ACM, WSDM '12, pp 623–632
49. Zhou R, Khemmarat S, Gao L (2010) The impact of youtube recommendation system on video views. In: Proceedings of the 10th ACM SIGCOMM conference on Internet measurement, ACM, IMC '10, pp 404–410