A Tweet-Centric Algorithm for News Ranking

Bo Zhang, Jinchuan Wang, Lei Zhang
Graduate School at Shenzhen,
Tsinghua University
{zhangbo0702, wjcthu}@gmail.com, zhanglei@sz.tsinghua.edu.cn

***Abstract*—Ranking news is helpful because of the information explosion which overloads readers. It is also a challenging task since news is now published by both news portals and microblogging platforms in real time. Most traditional news ranking algorithms consider two factors: media focus and user attention independently. While in the paper, we propose a news ranking framework to combine the two factors together. A better ranking algorithm is obtained via the following two parts: (1) the Influence Method (IM) to rank news and (2) the News Flow Graph Method (NFGM) to locate the most influential source for duplicated news from multiple news sources. We present four strategies to evaluate user attention. Experiments show that decay strategy based on Ebbinghaus forgetting curve is the best one. To the best of our knowledge, our paper is the first attempt to utilize microblogging data for news ranking.**

Keywords—News Ranking; Microblogging Platform; User Attention; Media Focus

I. Introduction

Nowadays, users generally have strong information needs. However, with the vast amount of news created and updated all the time, it is impossible for users to view them all. So, ranking news which is both timely and influential is a valuable research subject.

However, traditional link analysis techniques, like PageRank [11] or HITS [5], cannot work well for news ranking because they don't take time into account, which is exactly the most important characteristic for news spread. Although Google and other search engines have commercial search engines to index news feeds, little is known about their hidden algorithm, nor do we know how to evaluate their quality. Therefore, academic research on news ranking is still in great need.

Existing researches on news ranking have got two deficiencies. On the one hand, they never took microblogging data into account for news ranking, while they often assumed that users read news through web portals. So, they just utilize news portals data to rank news. However, the situation is being changed gradually when microblogging platforms are taking over. And related studies have shown that some microblogging platforms have both social feature and media feature. For example, [7] reveals Twitter is an Online Social Network Site (OSNS), as well as a news media, while [15] suggests Sina Weibo is the very first origin of some types of news through exploring the information flow patterns between Sina Weibo and several major news portals. Therefore, it will be possible to improve news ranking if we combine traditional web portals and microblogging platform. On the other hand, they never locate the most influential source for duplicated news. Obviously, there are a plenty of duplicated news so that it is highly possible there exists the same news among the several top news in the final sorted news list. So, it will be necessary for users to locate the most influential source so that all of the news in news list is distinct.

Influence of a piece of news is the relative importance of news over time, which is mainly determined by the following two factors:

- Media Focus
  It indicates the media report frequency.
- User Attention
  It indicates the user acceptability level.

In the research, we will discuss the problem of ranking news. Specifically, For dealing with the first deficiency which occurs in existing researches, we propose the Influence Method (IM) which not only comprehensively takes media focus and user attention into account, but also combines several major news portals and Sina Weibo, while, for settling the second deficiency, we present the News Flow Graph Method (NFGM) to locate the most influential source for duplicated news.

The contributions of this paper are as follows:

1. We propose a news ranking algorithm using microblogging data to rank news.
2. We present a news source locating method discovering the most influential source for duplicated news.

The rest of the paper is organized as follows: METHODS shows two methods we propose in detail. We describe the experimental data, evaluation on news ranking algorithm, and results in EXPERIMENTS. RELATED WORK gives a brief review of existing researches, followed by the conclusion and a discussion of future work in CONCLUSION.

II. Methods

The section consists of two parts: News Flow Graph Method and Influence Method. The former aims at locating the most influential source for duplicated news so that news in the final sorted news list is distinct, while the latter is intended to utilize microblogging data to perform news ranking. Fig. 1 shows the framework of our news ranking.

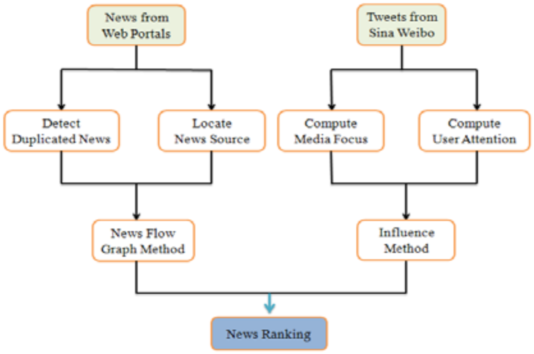

Fig. 1. News ranking framework.

A. News Flow Graph Method

Firstly, we compute the similarity of each pair of news to find clusters representing the same news. And then, for every discovered cluster of news, a news flow graph is created according to reference relationship among web portals publishing news in the cluster. Finally, a traditional link analysis technique, like PageRank or HITS, is used to locate the most influential source.

Given a list of news $N=[n_1,n_2,...,n_{|N|}]$, we can get clusters $C=[c_1,c_2,...,c_{|C|}]$ indicating the same news, where $c_i=[(s_{i_1},t_{i_1}),(s_{i_2},t_{i_2}),...,(s_{i_m},t_{i_m})]$ and $1\le i_1\le i_2\le...\le i_m\le|N|$. s_i means the referenced news media of news n_i, while t_i indicates the news media which cites news n_i from s_i. Namely, n_i flows from s_i to t_i. Then, for every discovered cluster c_i, a directed graph $G_i=(V_i,E_i)$ is produced to show the news flow process, where V_i stands for news media involved in cluster c_i and E_i reveals reference relationship among news media in cluster c_i. For example, if there exists $(P,Q)\in c_i$, there is an edge $<Q,P>\in E_i$.

In the work, we use HITS to estimate authority for each news media in graph G_i. HITS assumes each news media has a property of authority and a property of hub. Let $A(P)$ be authority for news media P and $H(P)$ be hub. It initializes authority and hub to 1 for each news media. And then, it recursively compute authority and hub for each news media according to the following four formulas:

$$A(P)=\sum_{Q\to P}H(Q) \tag{1}$$

$$H(P)=\sum_{P\to Q}A(Q) \tag{2}$$

$$A(P)=\frac{A(P)}{[\sum_{\forall P}(A(P))^2]^{\frac{1}{2}}} \tag{3}$$

$$H(P)=\frac{H(P)}{[\sum_{\forall P}(H(P))^2]^{\frac{1}{2}}} \tag{4}$$

It stops when $A(P)$ and $H(P)$ performs convergence. At the same time, news media with the biggest authority is regarded as the most influential source for cluster c_i.

B. Influence Method

We use tweets news media publish on Sina Weibo to rank news from traditional web portals. Specifically, we regard the number of tweets reporting some news as media focus of the news, while we present four strategies to model user attention.

If we directly use the number of comments and retweets to compute the influence of news, the value of influence is most likely large. So, it is necessary to design a function to map these original figures to a relatively small value. Generally, the function needs to meet the following conditions:

- $0\le f(x)\le 1$;
- $f(x)$ is a strictly increasing function of x ;
- $f(\infty)=1$ and $f(0)=0$.

In our work, we choose a sigmoid function which can meet the above mentioned conditions. Analogous to that used in [1], it is defined as followed:

$$f(x)=\begin{cases}\dfrac{\delta x}{1+\delta x} & ,x>0\\ 0 & ,otherwise\end{cases} \tag{5}$$

where δ is a smoothing parameter.

Let $T_d=[t_1,t_2,...,t_{|T_d|}]$ be a set of tweets which are published in date d. Intuitively, a tweet makes a contribution to influence of news if the tweet issues the news. Specifically, the level of contribution is related to user attention. Let IT_j^d indicate user attention of tweet j in a given day d, which is determined by the number of comments and retweets from users. An indicator variable X_{jk} means whether tweet j issues news k or not. It is defined as followed:

$$X_{jk}=\begin{cases}1 & ,Sim(t_j,n_k)\ge\varepsilon\\ 0 & ,otherwise\end{cases} \tag{6}$$

where ε is a threshold value.

In order to calculate $Sim(t_j,n_k)$, we define the vector space representation of t_j and n_k, namely, $t_j=n_k=[f_1,f_2,...,f_h]$ where $f_1,...,f_h$ is the common set of stemmed words by t_j and n_k after removing stops word. $w_T(f_i)$ and $w_N(f_i)$ represent the value of TF-IDF of feature f_i in t_j and n_k, respectively. Therefore, the similarity between t_j and n_k can be calculated as:

$$Sim(t_j,n_k)=\frac{\sum_{i=1}^{h}w_T(f_i)\cdot w_N(f_i)}{\sqrt{\sum_{i=1}^{h}w_T^2(f_i)}\cdot\sqrt{\sum_{i=1}^{h}w_N^2(f_i)}} \tag{7}$$

Finally, the influence of news k in a given day d can be defined as:

$$IN_k^d=\sum_{i=1}^{d}\sum_{j=1}^{|T_i|}X_{jk}\cdot IT_j^d \tag{8}$$

Concerning the computation of IT_j^d , we use four strategies. The first strategy (I) only considers the level of user attention of tweet j during the d th day, named u_j^d, and

$$IT_j^d = f(u_j^d) \tag{9}$$

Obviously, consecutive user attention will increase the influence of tweet. The second strategy (II) consider the accumulative user attention of j, named s_j^d , and

$$IT_j^d = f(s_j^d) = f(\sum_{i=1}^{d} u_j^i) \tag{10}$$

Clearly, the influence of tweet will decay over time. Unlike the first two strategies, the third strategy (III) considers decay. Analogous to decay value which [13] used, we use a constant β_c to show the decay value of the accumulative user attention, and

$$IT_j^d = f(s_j^d) = f((s_j^{d-1} - \beta_c) + u_j^d) \tag{11}$$

[6] proposed a dynamic decay scheme based on Ebbinghaus forgetting curve to calculate the influence of news topic and the scheme performed better in practical datasets. We try to reference the scheme directly and form the fourth strategy (IV). It is defined as:

$$IT_j^d = f(s_j^d) = f(\sum_{i=1}^{d} \alpha \cdot u_j^i \cdot \ln(\frac{\beta}{d-i}) + u_j^d) \tag{12}$$

where both α and β are fixed parameters.

III. EXPERIMENTS

Some information about experiments will be given in the part. First, we will show how we collect data in Data. Then, an evaluation scheme will be proposed to assess our news ranking algorithm in Evaluation, followed by some results of experiments in Results.

A. Data

To test the performance of IM, four types of data were collected: news from web portals, tweets from Sina Weibo, daily top ranked news for each category from Google News, and users oriented interest judgment from users.

1) News Collection

We collect news from seven web portals in a period of 22 days (from 2012/12/06 to 2012/12/27) and news dataset is classified in seven different categories.

2) Tweets Collection

Daily tweets published by news media are collected using Sina Weibo API from 2012/12/06 to 2012/12/27. Initially, we need to calculate media focus of specific news by exploring the number of tweets issuing the news so that it is necessary to store text segment of tweets. Then, if a tweet issues news, we need to use the number of retweets and comments to estimate user attention of the tweet. Therefore, for each tweet, we store time information, id information, retweet information, comment information and text information. Specifically, we collect tweets at 12:00pm every day. TABLE I shows the details.

TABLE I. Tweets statistics

# Media	#Tweets	#Retweets	#Comments
2298	479110	31953511	12241197

3) Top Ranked News Collection

Top twenty ranked news for each category, except for Military and Social, from Google News are collected every day at 12:00pm from 2012/12/06 to 2012/12/27. The ranking position of news is stored as well as news title, news source and date.

4) User Judgment Collection

Real time users' judgments are collected to serve as baseline to compare different ranking results for strategy I, II, III, IV and Google News. For each piece of news from Google News, we have five volunteers to judge it and the level of interest in news lies in the following four levels:

- Not Hear (1 Point)
  Users never hear the news.
- Not Interesting (2 Point)
  Users hear the news, but not interested in it.
- Interesting (3 Point)
  Users hear the news and interested in it.
- Very Interesting (4 Point)
  Users hear the news and very interested in it.

B. Evaluation

We utilize Normalized Discounted Cumulative Gain ($NDCG$) [4] to measure the effectiveness of different strategies. And, $NDCG$ value at position n on d th day is defined as:

$$NDCG_d @ n = Z_n \cdot \sum_{i=1}^{n} \frac{2^{r(i)} - 1}{\log_2(i+1)} \tag{13}$$

where Z_n is a normalization factor and $r(i)$ is rating of the news located at the i th position. Then, the average $NDCG$ of a ranking strategy at position n during some period of time is defined as:

$$NDCG @ n = \frac{\sum_{d=The\,First\,Day}^{The\,Last\,Day} NDCG_d @ n}{\# days} \tag{14}$$

C. Results

We perform a large number of experiments. For space reason, we just present the most important experiment results.

1) Method Evaluation

The first group of experiments is method evaluation. Firstly, we check $NDCG$ values at position n of strategy I, II, III, IV, and Google News for Domestic news and Entertainment news over a period of time (from 2012/12/06 to 2012/12/27). Fig. 2 proposes the trends of the $NDCG_d @10$ of strategy I, II, III, IV and Google News for Domestic News from 2012/12/06 to 2012/12/27. From that, we can see, clearly, the $NDCG_d @10$ of Google News is larger than that of others, except for a few days. And, for most days, strategy IV outperforms strategy I, II, and III. Analogous to

Fig. 2, we show that of Entertainment news in Fig. 3. Fig. 3 shows the $NDCG_d@10$ of strategy IV for most days is larger than that of others.

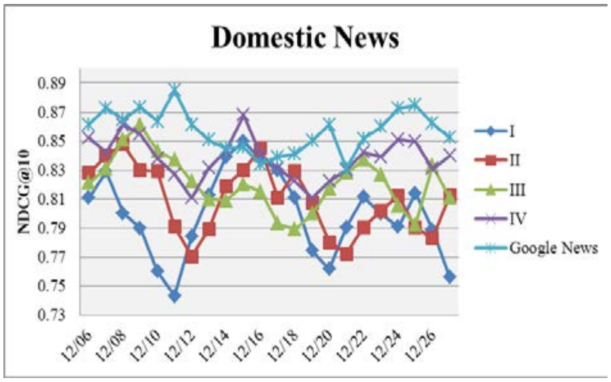

Fig. 2. Trends of the $NDCG_d@10$ of strategy I, II, III, IV and Google News for Domestic News from 2012/12/06 to 2012/12/27.

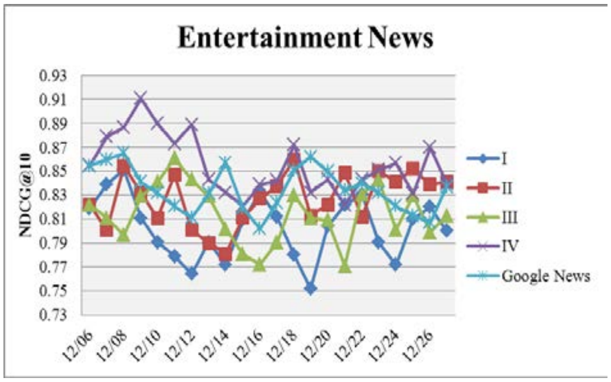

Fig. 3. Trends of the $NDCG_d@10$ of strategy I, II, III, IV and Google News for Entertainment News from 2012/12/06 to 2012/12/27.

TABLE II. Ranking performance comparison of Domestic news for strategy I, II, III, I V, and Google News. Except for the first column, the bold figure means the biggest value of corresponding $NDCG@n$ in each column, while * denotes the biggest value of that of strategy I, II, III, and IV.

Domestic News	**NDCG@3**	**NDCG@5**	**NDCG@10**
I	0.8293	0.8217	0.8031
II	0.8407	0.8337	0.8124
III	0.8571	0.8429	0.8209
IV	0.8804*	0.8691*	0.8367*
Google News	**0.8972**	**0.8724**	**0.8512**

TABLE III. Ranking performance comparisons of Entertainment news

Entertainment News	**NDCG@3**	**NDCG@5**	**NDCG@10**
I	0.8359	0.8207	0.8082
II	0.8528	0.8440	0.8316
III	0.8647	0.8497	0.8218
IV	**0.9098***	**0.8964***	**0.8548***
Google News	0.8837	0.8650	0.8369

Then, we calculate the average $NDCG_d@n$ of strategy I, II, III, IV, and Google News for Domestic news and Entertainment news over the observation period. Specifically, let n be 3, 5, and 10, respectively. We get the specific value shown in TABLE II and TABLE III.

On the one hand, from TABLE II, Google News outperforms others no matter what n is, while strategy IV is optimal in the four strategies proposed in the paper in terms of all of three $NDCG@n$ values, which means dynamic decay scheme based on Ebbinghaus forgetting curve is superior to accumulative user attention scheme and constant decay scheme. However, when it applies to Domestic news, its performance is inferior to Google News. On the other hand, TABLE III indicates strategy IV is not only optimal in the four strategies proposed in the paper, but also it is superior to that of Google News.

TABLE IV. Top ten news over the observation period of time (from 2012/12/06 to 2012/12/27) for Domestic news.

News Summary	**Source**	**Post Time**
Jinping Xi inspects Shenzhen	Sina	12/08
A man slashes 22 pupils in Henan Province	Chinanews	12/14
Jinping Xi inspects Guang Dong Province	Chinanews	12/11
Government building of Jinan costs 4billion	People	12/11
High speed rail of Jing-Guang opens today	The Beijing News	12/26
A shirt workshop in Shantou encounters a big fire	Ifeng	12/10
A satellite is launched in Jiuquan successfully	Chinanews	12/19
Chuncheng Li is involved in disciplinary violations	Xinhua	12/06
Ma Ying-jeou is hit by shoes	Sina	12/11
North Korea shoot a rocket	Chinanews	12/13

2) *Ranking News*

The second group of experiments deals with the principal goal of the paper, namely, ranking news. From the results discussed in the last section, strategy IV is optimal in the four strategies in terms of Domestic news and Entertainment news. Therefore, we use strategy IV to calculate the level of influence of each Domestic and Entertainment news. In addition, NFGM is used to locate the most influential source for duplicated news after we get the ranked news list in terms of the level of influence of a piece of news.

Once the most influential source of news is determined, we get a final ranked news list where there is not duplicated news. For space consideration, it is impossible for us to present the whole news list. Here, we just list top ten news over the observation of time for Domestic news and Entertainment news in TABLE IV and TABLE V, respectively.

TABLE V. Top ten news over the observation period of time (from 2012/12/06 to 2012/12/27) for Entertainment news.

News Summary	Source	Post Time
Haoming Yu makes an early return from burns	Xiaoxiang Morning	12/08
Thailand List is released all over the country	Sina	12/11
The new book Jing Cai Writes is released for the first time	Sina	12/15
Benshan Zhao's apprentice beats a person	Chinanews	12/06
Jackie Chan is involved in a money-laundering scandal	Chinanews	12/25
The Grand Master is unveiled in Berlin	Sina	12/20
Edison takes his girlfriend to meet godfather	Tencent	12/10
The box office of Thailand List is over 300 million	Sina	12/17
Life of Pi is selected in Oscar	Chinanews	12/19
CCTV News emerges switching error	Xinhua	12/09

IV. RELATED WORK

As we all know, media focus and user attention are two primary factors which determine the level of influence of a piece of news. Generally, researches on news ranking mainly perform around these two factors. Therefore, according to either just consider one of media focus and user attention, or take both into consideration, news ranking can be classified into the following three categories: news ranking based on media focus [2, 3, 9], news ranking based on user attention [12], and, news ranking based on media focus and user attention [6, 8, 13, 14].

A. News Ranking Based on Media Focus

The kind of researches just takes media focus into account. They think the level of influence of news is related to authority of news media reporting the news and the number of times that the news is referenced by another news media. Specifically, [2] is the first academic paper on news ranking. It proposes a ranking framework, which comprehensively takes reference relationship among news sources, mutual reinforcement between news and sources, and timeliness into consideration at, to model the process of generation of a stream of news and the process of clustering of news by topics. And, the complexity of its algorithms is linear so that online process of ranking news is allowed. Except for exploiting the mutual reinforcement relationship between news articles and news sources similar to that used in [2], [3] employs the visual layout information of news homepages to produce a quite effective ranking algorithm. Analogous to [2], [9] considers news and news sources as well. In addition to this, the concept of news topic is involved by clustering news from different sources. So, ranking algorithm in it fully exploits the reinforcement between news sources, topics and news.

However, some shortages still exist in the kind of researches. Firstly, they never consider influence from users when they calculate influence of news. Obviously, it is necessary to deal with user attention. Secondly, they never consider social network. Clearly, it is wise for us to utilize social network to news ranking.

B. News Ranking Based on User Attention

Just as the title implies, the kind of researches just consider influence from users when performing news ranking. Traditionally, it is difficult for researchers to use web portals data to model user attention since user interest is highly dynamic so that the number of researches just taking user attention into account is relatively less. However, the situation is being changed gradually when social network is taking over, considering real time property of social network. [12] is one of few researches using Twitter data to model user attention. Specifically, it proposes a community tweets voting model to re-rank Google and Yahoo news search results on the basis of vast amounts of Twitter community data. However, it deals with news ranking problems related to queries, while we wish to generate a ranked news list in terms of objective influence of news.

However, different from news ranking based on media focus, the kind of news ranking never considers media focus.

C. News Ranking Based on Media Focus and User Attention

Clearly, the kind of news ranking fully takes influence from media and users into consideration. The goal of [13] is to rank news topic. It references aging theory [1] to model a news topic's life span and considers a news topic as a life form with stages of birth, growth, decay and death. Based on [13], [6] proposes several ranking schemes according to whether decay function is constant or not and whether considering media focus and user attention or not. Final, it draws two meaningful conclusions: decay scheme based on Ebbinghaus forgetting curve outperform others when considering media focus only; fusion scheme based on interaction between media focus and user attention is optical when fully considering media focus and user attention. [8] regards media focus, user attention and timeliness as news features and uses a linear regression to perform news ranking. Like [2], [14] considers mutual reinforcement relationship between news and news sources as well. In addition, it regards response rate from users as user attention. Final, its ranking results is very close to people's judgment.

However, the kind of news ranking algorithms never considers social network data to model media focus and user attention.

V. CONCLUSION

In the paper, we propose a news ranking framework which consists of two parts: the one is Influence Method

(IM) which ranks news, while the other is News Flow Graph Method (NFGM) which locates the most influential source for duplicated news. And, we present four strategies (I, II, III and IV) to model user attention. Our main findings are listed as follows:

1. Decay strategy based on Ebbinghaus forgetting curve (IV) is the best one among the four strategies we propose to model user attention in the paper.
2. Strategy IV is inferior to Google News when applying to Domestic news.
3. Strategy IV is superior to Google News when applying to Entertainment news.

Future work lies in two directions: we will take different ranking schemes for different types of news; machine learning approaches will be used to rank news, since it has been used in [10].

VI. REFERENCES

[1] Chen, C., Chen, Y.T., Sun, Y., and Chen, M., "Life Cycle Modeling of News Events Using Aging Theory," In *Proc. ECML 2003*, Springer (2003), 47-59.

[2] Corso, G.M.D., Gulli, A., and Romani, F., "Ranking a Stream of News," In *Proc. WWW 2005*, ACM (2005), 97-106.

[3] Hu, Y., Li, M., Li, Z., and Ma, W., "Discovering authoritative news sources and top news stories," *Information Retrieval Technology*, Springer (2006), 230-243.

[4] Järvelin, K. and Kekäläinen, J., "Cumulated Gain-based Evaluation of IR Techniques," *ACM Trans. Inf. Syst.*, ACM (2002), 20(4): 422-446.

[5] Kleinberg, J.M., "Authoritative Sources in a Hyperlinked Environment," *Journal of the ACM*, ACM (1999), 46(5): 604-632.

[6] Kong, L., Jiang, S., Yan, R., Xu, S., and Zhang, Y., "Ranking News Events by Influence Decay and Information Fusion for Media and Users," In *Proc. CIKM 2012*, ACM (2012), 1849-1853.

[7] Kwak, H., Lee, C., Park, H., and Moon, S., ”What is Twitter, a Social Network or a News Media?" In *Proc. WWW 2010*, ACM (2010), 591-600.

[8] Li, H., "A Linear Regression Based News Topic Hotness Calculation Approach," *Journal of Computational Information Systems*, (2012), 8(20): 8637-8644.

[9] Mao, X. and Chen, W., "A method for ranking news sources, topics and articles," In *Proc. ICCET 2010*, IEEE (2010), 4:170-174.

[10] McCreadie, R., Macdonald, C., and Ounis, I., "A Learned Approach for Ranking News in Real-Time Using the Blogosphere," In *Proc. SPIRE 2011*, Springer (2011), 104-116.

[11] Page, L., Brin, S., Motwani, R., and Winograd, T., "The PageRank Citation Ranking: Bringing Order to the Web," *Technical Report*, Standard InfoLab (1999).

[12] Shuai, X., Liu, X., and Bollen, J., "Improving News Ranking by Community Tweets," In *Proc. WWW 2012*, ACM (2012), 1227-1232.

[13] Wang, C., Zhang, M., Ru, L., and Ma, S., "Automatic Online News Topic Ranking Using Media Focus and User Attention Based on Aging Theory," In *Proc. CIKM 2008*, ACM (2008), 1033-1042.

[14] Yang, W., Dai, R., and Cui, X., "Model for Internet News Force Evaluation Based on Information Retrieval Technologies," *Journal of Software*, (2009), 20(9): 2397-2406.

[15] Zhang, B., Wang, J., and Zhang, L., "Exploring Information Flow Patterns between News Portals and Micro-blogging Platforms," In *Proc. SWWS 2012*, Springer (2012).