

# Automatic Online News Topic Ranking Using Media Focus and User Attention Based on Aging Theory

Canhui Wang, Min Zhang, Liyun Ru, Shaoping Ma

State Key Lab of Intelligent technology & systems,  
Tsinghua National Laboratory for Information Science and Technology,  
CS&T Department, Tsinghua University, Beijing, 100084, China P.R.

{wangcanhui, mzhang.bj}@gmail.com

## ABSTRACT

News topics, which are constructed from news stories using the techniques of Topic Detection and Tracking (TDT), bring convenience to users who intend to see what is going on through the Internet. However, it is almost impossible to view all the generated topics, because of the large amount. So it will be helpful if all topics are ranked and the top ones, which are both timely and important, can be viewed with high priority. Generally, topic ranking is determined by two primary factors. One is how frequently and recently a topic is reported by the media; the other is how much attention users pay to it. Both media focus and user attention varies as time goes on, so the effect of time on topic ranking has already been included. However, inconsistency exists between both factors. In this paper, an automatic online news topic ranking algorithm is proposed based on inconsistency analysis between media focus and user attention. News stories are organized into topics, which are ranked in terms of both media focus and user attention. Experiments performed on practical Web datasets show that the topic ranking result reflects the influence of time, the media and users. The main contributions of this paper are as follows. First, we present the quantitative measure of the inconsistency between media focus and user attention, which provides a basis for topic ranking and an experimental evidence to show that there is a gap between what the media provide and what users view. Second, to the best of our knowledge, it is the first attempt to synthesize the two factors into one algorithm for automatic online topic ranking.

## Categories and Subject Descriptors

H.3.5 [Information Storage and Retrieval]: Online Information Services – *Web-based services*; H.1.2 [Models and Principles]: User/Machine Systems – *Human factors*; H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval – *Clustering*.

## General Terms

Algorithms, Experimentation, Human Factors

## Keywords

Topic Ranking, Media Focus, User Attention, Page View

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CIKM'08, October 26–30, 2008, Napa Valley, California, USA.  
Copyright 2008 ACM 978-1-59593-991-3/08/10...\$5.00.

## 1. INTRODUCTION

News stories are gathered from many Websites and organized into news topics by practical Web applications like *Google News* (<http://news.google.com/>). Topics are constructed and updated online automatically, using the techniques of Topic Detection and Tracking (TDT) [1]. News TDT results have greatly facilitated users who want to know about “what’s new” or “what’s going on”. However, with the vast amount of news topics created and updated all the time, it is almost impossible for users to view them all. So a new problem arises: how to rank the news topics to show the top ones with high priority, which are both timely and important?

Two simple rules are separately used for ranking news topics in [21]: ranking by the time when the topics are updated; ranking by the number of news stories included in the topics. The first rule focuses on returning timely results and the second one considers larger topics more important. Either rule involves only one aspect of the ranking problem. Besides them, more factors must be taken into consideration: (1) every news story of a topic contributes to its importance, while the contribution decays along the timeline; (2) topics that attract more users’ attention should be ranked higher.

On the whole, topic ranking is determined by two primary factors. First, how frequently and recently is the topic reported by news Websites? This factor is called *media focus*. Second, how much do users like to read news stories about the topic? This one is called *user attention*. Both media focus and user attention varies as time goes on, so the effect of time on topic ranking has already been included by the two factors. However, inconsistency exists between both factors. Detailed explanations are as follows:

A topic reported frequently by the media does not always attract close attention of users, because users may judge it as unimportant (although there are many news stories about it). There are also many other reasons, such as unattractive news titles, uninteresting subjects, unnoticed positions and so on. Therefore media focus itself is not enough for topic ranking.

In contrast, a topic reported poorly by the media might attract many users’ attention, such as, when news stories of the topic have interesting titles, or the topic is surprising or very close to users’ life and interests. There also exist other reasons, such as the effect of anchor texts. News stories are mainly from authority pages [13], which are pointed to by news index pages (hub pages, such as homepages of news Websites) with hyperlinks. Users often view the news stories through clicks on the hyperlinks, so anchor texts play an important role in the decision of viewing or not. Attractive or even odd anchor texts that have nothing to do with the real content will help stories win much user attention, but

in fact users are just cheated by the anchor texts. Hence only user attention is not enough for topic ranking, either.

The inconsistency between media focus and user attention is studied and calculated quantitatively in this paper. Media focus and user attention of topics is calculated online automatically based on an aging theory. Topics with high values of both media focus and user attention are ranked high.

There are two problems to be investigated in news topic ranking: (1) How to calculate media focus and user attention of topics and the inconsistency between both? (2) How to rank topics, taking into account all mentioned factors?

Motivated by these problems, we propose a news topic ranking algorithm in this paper. The state-of-the-art TDT techniques are used to organize news pages from a lot of news Websites into topics. An aging theory is added in the TDT process. In the meantime, media focus and user attention of topics is calculated. Finally, topics are ranked online automatically.

The contributions of this paper are twofold: (1) we define and analyze the quantitative measure of the inconsistency between media focus and user attention, which provides a basis for topic ranking. (2) We propose a news topic ranking algorithm using both media focus and user attention, which has not been studied, to the best of our knowledge.

The rest of the paper is organized as follows: Section 2 gives a brief review of related work. Section 3 describes the topic detection and tracking algorithm based on aging theory and burstiness of terms. The topic ranking algorithm based on inconsistency analysis between media focus and user attention is proposed in section 4. We describe the experimental data and results, and discuss on the results in Section 5, followed by the conclusion and a discussion of future work in Section 6.

## 2. RELATED WORK

Topic detection and tracking (TDT) are intended to structure news stories from newswires and broadcasts into topics [1]. Approaches in TDT were mainly variants and improvements of the single pass method and agglomerative clustering algorithms [2, 3, 7, 14, 15, 16, 19, 21, 22, 23]. Although [3] concluded that time information “did not help” improve the new event detection results, some recent work has utilized the aging theory or timeline analysis, and achieved good performance in TDT and hot topic extraction [4, 5]. The state-of-the-art TDT techniques are used to generate topics from news stories in our work. We also apply the aging theory both in the TDT process and the calculation of media focus and user attention. However, traditional TDT tasks [1] are not the main focus of our work.

Other work involving both topic detection and timeline analysis is mostly in the retrospective way: topic hierarchy construction based on identification of bursty periods of features [8]; topic sentence extraction along a timeline given a query [6]; topic detection based on identification of both aperiodic and periodic features’ bursts [10]; finding top bursty topics by identifying bursty words [11]; and so on. Previous approaches listed above analyzed the characteristics of features from a fixed corpus on the whole timeline, and hence have to be adjusted to suit to online use. Our system deals with dynamic-increasing news data online, and makes use of an aging theory [4] in topic detection and tracking.

Fung *et al* proposed a feature-pivot clustering approach, using “the time information to determine a set of bursty features which may occur in different time windows” [9]. Bursty topics were

detected by grouping bursty features based on the feature distributions, and the hot periods of bursty topics were identified finally.

Chen *et al* presented a hot topic extraction approach based on timeline analysis and multi-dimensional sentence modeling [5]. Hot terms were firstly “extracted by mapping their distribution over time” and key sentences were then identified and “grouped into clusters that represent hot topics”. The following characteristics of hot topics were considered in [5]: appearing in many news stories and many news channels; strong continuity; popularity variation over time.

Our work is different from [9] and [5]: (1) all topics are ranked. Topic ranking is actually a different problem from bursty/hot topic detection. (2) News stories about the same topic are clustered for users’ quick access. (3) The approaches in [9] and [5] required information from a fixed corpus on the whole timeline, while our system deals with dynamic-increasing data online.

He *et al* proposed a semi-automatic hot topic detection algorithm in [12]. They detected new topics “by using incremental TF-IDF model and incremental clustering algorithm”. Topics were finally ranked by the frequencies of topic reported in time units and the number of consecutive effective time units. However, their approach is heuristics, not particularly well justified.

A distinct feature of our approach from all previous work lies in the consideration of users’ effect on topic ranking. The proposed algorithm performs an inconsistency analysis between media focus and user attention, which is novel, to the best of our knowledge. The information of users’ attention to topics is introduced and the topic ranking result reflects the influence of time, the media and users.

## 3. TOPIC DETECTION AND TRACKING BASED ON AGING THEORY AND BURSTINESS OF TERMS

The first stage of topic ranking is to organize news stories, published by various Websites, into topics. We utilize our topic detection and tracking algorithm based on clustering [21], considering the characteristics of news appearance. A topic is considered as a life form and an aging theory is used to model its life span [4]. The burstiness of terms is calculated and added in the story representation. Newly appearing stories are firstly clustered into new topic candidates, which are used to track previous “alive” topics. A new topic is generated if the candidate is not combined with any previous topic. Previous topics become popular with the insertion of new stories, and fade away with time. The whole topic detection and tracking process is performed online automatically.

### 3.1 Aging Theory

Chen *et al* applied an aging theory to model a news topic’s life span and considered a news topic as “a life form with stages of birth, growth, decay and death” [4]. They used the concept of energy function to track the life cycles of topics. The value of energy function indicates the liveliness of a news topic in its life span. The energy of a topic increases when it becomes popular and decreases as its popularity decays. Like in the nutriology, things that contribute to the energy of topics (e.g. new stories inserted into the topics) are called *nutrition*.

The timeline is divided into time slots of equal length. We make use of three functions from [4] to calculate and update the energy of topics in every time slot:

- *getNutrition()* calculates the nutrition that a topic receives from a story;
- *energyFunction()* converts a topic nutrition value into an energy value. It's a strictly monotone increasing function, so *energyFunction<sup>-1</sup>()* converts an energy value into a topic nutrition value;
- *energyDecay()* carries out the energy decrease in each time slot.

### 3.2 Burstiness of Terms

Topics are related to a set of keywords [8]. New topics usually appear along with some bursty terms, and stories with the same bursty terms are more probably from the same topic. The relationship between topics and corresponding bursty keywords has been utilized previously [8, 9, 10, 11, 17, 18]. We make use of the burstiness of terms to obtain better TDT results, analogous to what [11] did in topic clustering.

The calculation of term burstiness is in a similar way to [18]. If we model stories of topics as a random process with an unknown binomial distribution, we can check the associations between terms and time slots. For each term  $w$ , we calculate the contingency table, as shown in Table 1:

**Table 1. Contingency Table for term  $w$  in time slot  $i$**

	$i$	$\bar{i}$
$w$	$A$	$B$
$\bar{w}$	$C$	$D$

where

- $A$  is the count of stories that contain term  $w$  in time slot  $i$ ;
- $B$  is the count of stories that contain term  $w$  outside time slot  $i$ ;
- $C$  is the count of stories that don't contain term  $w$  in time slot  $i$ ;
- $D$  is the count of stories that don't contain term  $w$  outside time slot  $i$ .

We use the  $\chi^2$  statistic formula to calculate the burstiness of term  $w$  during time slot  $i$ :

$$b_i(w) = \frac{(A+B+C+D)(AD-BC)^2}{(A+B)(C+D)(A+C)(B+D)} \quad (1)$$

### 3.3 Representations and Algorithms

#### 3.3.1 Pre-Processing and Story Representation Using B-VSM

News pages are gathered from news Websites by crawlers. Then the crawled pages are analyzed to get the main body. Their titles and content are extracted along with metadata such as publishing date, category, source and images. A news Web page's content part (including the title) is called a story.

We split stories into sentences and remove redundant sentences in a story because they usually bring in nothing useful but noises. Titles are weighted and added to the story representation. And then we tokenize words (word segmentation is performed when dealing with Chinese texts), perform part-of-speech tagging, recognize named entities and remove stopwords. Finally a term vector is created for each story.

Incremental TF-IDF model is widely applied to term weight calculation in TDT [2, 3, 22, 23]. We choose this model as a base to weight terms.  $DF$  (document frequency) of term  $w$  in time slot  $i$  is calculated as:

$$df_i(w) = df_{i-1}(w) + df_{S_i}(w) \quad (2)$$

where  $S_i$  means a set of stories coming during time slot  $i$ , and  $df_{S_i}(w)$  means the number of stories that term  $w$  appears in.  $df_{i-1}(w)$  represents the number of stories that term  $w$  appears in before time slot  $i-1$  (included). A training corpus comprised of a sufficient amount of stories is used for the calculation of  $DF$  initially. As shown in formula (2),  $DF$  is updated dynamically in each time slot  $i$ .

Then each story  $d$  in time slot  $i$  is represented as an  $n$ -dimension vector, where  $n$  is the number of distinct terms in story  $d$ . Each dimension is weighted using a combination of incremental TF-IDF model and B-VSM model [11], which considers the burstiness of terms. And the vector is normalized so that it is of unit length:

$$weight(d, w) = \frac{tf(d, w) \log((N_i + 1)/(df_i(w) + 0.5)) \cdot b_i(w)}{\sqrt{\sum_{w' \in d} (tf(d, w') \log((N_i + 1)/(df_i(w') + 0.5)) \cdot b_i(w'))^2}} \quad (3)$$

where  $tf(d, w)$  means how many times term  $w$  appears in story  $d$  and  $N_i$  represents the total number of stories before time slot  $i$  (included).

#### 3.3.2 Similarity Calculation

Cosine similarity is used to calculate the similarity between two stories. For story  $d$  and  $d'$ , their similarity is calculated as:

$$similarity(d, d') = \sum_{w \in d \cap d'} weight(d, w) * weight(d', w) \quad (4)$$

#### 3.3.3 Topic Detection, Tracking and Energy Calculation

We combine our TDT algorithm proposed in [21] with the aging theory described in Section 3.1 to perform online topic detection and tracking. Topics are generated and updated with stories coming in chronological order, one time slot after another. Topics are represented as term vectors, which are defined as the arithmetic average of term vectors of all stories within them. The algorithm is shown in Figure 1:

```

T = null;
For each time slot i
  New coming stories are clustered into new topic candidates;
  For each new topic candidate  $t_c$ 
     $t = \text{argmax}_{t \in T}(\text{sim}(t, t_c))$ ;
    If  $\text{sim}(t, t_c) > \text{threshold}_{\text{track}}$  then
      Nutrition of  $t$  currently:  $n_t = \text{energyFunction}^{-1}(e_t^{(i-1)})$ ;
      For each story  $d$  in  $t_c$ 
         $\text{getNutrition}()$  from  $d$ :  $n_d = \text{sim}(t, d)$ ;
        Add to the nutrition of  $t$ :  $n_t = n_t + \alpha_m * n_d$ ;
      End for
      Calculate the new energy of  $t$ :  $e_t^{(i)} = \text{energyFunction}(n_t)$ ;
      Combine  $t_c$  to  $t$ , and update the term vector of  $t$ ;
    Else
       $t_c$  is created as a new topic, and add  $t_c$  to  $T$ ;
      Assign  $t_c$  the initial nutrition:  $n_{t_c} = \alpha_m * |t_c|$ ;
      Calculate the initial energy:  $e_{t_c}^{(i)} = \text{energyFunction}(n_{t_c})$ ;
    End if
  End for
End for

```

**Figure 1. Topic detection, tracking and energy calculation algorithm**

New coming stories are clustered into new topic candidates, using the state-of-the-art clustering algorithms [24]. The similarity between topics  $t$  and  $t_c$  ( $\text{sim}(t, t_c)$ ) is calculated as the arithmetic average of pair-wise similarities between the stories in  $t$  and the stories in  $t_c$ . As pointed out in [24], the similarity calculated equals to the inner product between term vectors of topics  $t$  and  $t_c$ , so the topic similarity calculation is the same as the story similarity calculation.

$e_t^{(i)}$  represents the energy value of topic  $t$  in time slot  $i$ . The  $\text{energyFunction}()$  used in this paper is a sigmoid function, analogous to that used in [4], defined as:

$$\text{energyFunction}(x) = \begin{cases} \frac{x}{1+x} & x > 0 \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

It is easy to see that  $0 \leq \text{energyFunction}(x) < 1$ .

$|t_c|$  is the number of stories contained in topic  $t_c$ . The  $\text{threshold}_{\text{track}}$  and the coefficient used in the increase of the nutrition ( $\alpha_m$ ) are determined by experiments with training data.

### 3.3.4 Topic Removal Based on Energy Decay

The energy of a topic increases when new stories are added to the topic, and decreases as time goes by. The energy value of every topic is reduced with a decay factor  $\beta_m$  in every time slot. When no or few stories are added to a topic, its energy value will gradually decline. If the energy value is below a threshold, the topic is considered “dead” and removed from the topic set  $T$  to keep all topics in the news reading system up to date. The topic

removal algorithm based on energy decay is analogous to that of [4], described in Figure 2:

```

For each time slot i
  For each topic  $t$  in  $T$ 
    If  $e_t^{(i)} > \beta_m$  then
       $\text{energyDecay}()$  on  $t$ :  $e_t^{(i)} = e_t^{(i)} - \beta_m$ ;
    Else
       $e_t^{(i)} = 0$ ;
      Remove  $t$  from  $T$ , and add  $t$  to  $T_{\text{remove}}$ ;
    End if
  End for
End for

```

**Figure 2. Topic removal algorithm based on energy decay**

“Dead” topics are put in  $T_{\text{remove}}$ . The decay factor  $\beta_m$  is determined by experiments with training data.

## 4. TOPIC RANKING USING BOTH MEDIA FOCUS AND USER ATTENTION

Through investigation of characteristics of topics, we have found out that topic ranking is determined by two primary factors: media focus and user attention. Media focus quantifies how frequently and recently the topic is reported by the media, and user attention measures how much and recently users like to read news stories about the topic. Both factors vary as time goes on, so the effect of time on topic ranking has already been included. However, inconsistency exists between both factors. We calculate media focus and user attention of topics and rank the topics by considering both factors.

### 4.1 Calculation of Media Focus and User Attention

#### 4.1.1 Media Focus Calculation

Media focus indicates how frequently and recently topics are reported by the media. The energy value of topics calculated in Section 3.3 shows the liveliness of topics from the perspective of the media. Therefore, we use the energy value of a topic in each time slot as the value of media focus on the topic directly. In this way, media focus values of topics are calculated online automatically:

$$mf_t^{(i)} = e_t^{(i)} \quad (6)$$

where  $mf_t^{(i)}$  represents the media focus value of topic  $t$  in time slot  $i$ .

#### 4.1.2 User Attention Calculation

User attention represents how much and recently users pay attention to topics. Anonymous Web page view information is collected for the user attention calculation, with the help of a Web browser toolbar from a major commercial Website in China (<http://www.sohu.com/>). Web browser toolbars become more and more popular recently. Many search engine companies (e.g. Google, Yahoo, Baidu...) have released toolbar software to collect user behavior, in order to provide better service based on the wisdom of the crowds. Therefore it's not difficult to collect

Web page view information of users, and our algorithm is full of practicability.

In this proposed system, news pages are gathered from many news Websites by crawlers, and users' page view behavior on these Websites can be extracted from the anonymous Web page view information that we have collected on the whole Web. In this way, we get a sample of users' page view information on news Websites. At present, this sample is directly used to calculate user attention values of news topics.

Actually, the system we proposed in this paper is also a news reading Website. We have not taken the user behavior on our system into consideration, because it is sort of a positive feedback, which is not our focus in this paper.

Similar to the media focus calculation, the user attention is also calculated based on the aging theory. Every page view of any story about topic  $t$  is considered as viewing  $t$  once. The page views contribute to the user attention of a topic, which decays as time goes by.

We make use of three functions from [4] to calculate and update the user attention of topics in every time slot. The nutrition here is called *view value*, and  $ua_t^{(i)}$  is used to represent the user attention value of topic  $t$  in time slot  $i$ . The user attention values of new topics are set as 0 initially. The algorithm is shown in Figure 3:

```

For each time slot  $i$ 
  For each topic  $t$  in  $T$  and  $T_{remove}$ 
    View value of  $t$  currently:  $v_t = energyFunction^{-1}(ua_t^{(i-1)})$ ;

    For each story  $d$  in  $t$ 
       $getNutrition()$  from  $d$ :  $v_d = \text{page view count of } d \text{ in } i$ ;
      Add to the view value of  $t$ :  $v_t = v_t + \alpha_u * v_d$ ;
    End for
    The new user attention of  $t$ :  $ua_t^{(i)} = energyFunction(v_t)$ ;
    If  $ua_t^{(i)} > \beta_u$  then
       $energyDecay()$  on  $t$ :  $ua_t^{(i)} = ua_t^{(i)} - \beta_u$ ;
    Else
       $ua_t^{(i)} = 0$ ;
    End if
  End for
End for

```

**Figure 3. User attention calculation algorithm**

$\alpha_u$  and  $\beta_u$  are determined by experiments with training data. Note that user attention values of topics in  $T_{remove}$  are also calculated, because users may continue their attention to a topic even if it is not reported by the media any more.

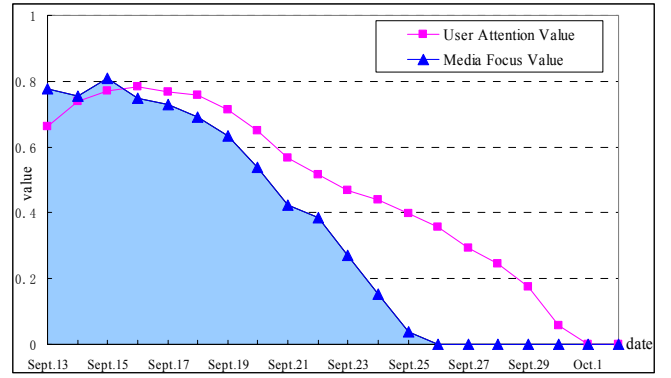
## 4.2 Quantitative Measure of Inconsistency between Media Focus and User Attention

For each topic, the values of media focus and user attention are varying in each time slot. Moreover, the user attention is usually delayed a little bit than the media focus, because users read news stories after they have been issued. Therefore, it is difficult and

unnecessary to put media focus and user attention of a topic in alignment, and measure the inconsistency between both in every time slot.

The inconsistency between media focus and user attention is quantitatively measured, in consideration of the following factors: (1) the accumulative media focus and user attention values are used, instead of the values in a certain time slot. The former is more reasonable and avoids the problem of aligning both factors; (2) only topics in  $T_{remove}$  with  $ua_t^{(i)}=0$  are considered, because the accumulative media focus and user attention values of these topics are steadier; (3) topics, not a single topic, are used in the calculation, in order to make the result of statistical significance.

Figure 4 below shows the media focus and user attention values of topic  $t$  along its lifetime. The length of time slots in actual use is 15 minutes, in order to provide timely results online. Only the values in the last time slot of each day are shown in Figure 4, for the sake of clear demonstration.



**Figure 4. Media focus and user attention values of topic  $t$  along its lifetime**

The area of the shaded part in Figure 4 represents the accumulative media focus value of topic  $t$ , which is calculated as:

$$amf_t = \sum_i mf_t^{(i)} \quad (7)$$

Similarly, the accumulative user attention value of topic  $t$  is calculated as:

$$aau_t = \sum_i ua_t^{(i)} \quad (8)$$

The inconsistency between media focus and user attention is quantitatively measured using the Pearson correlation coefficient:

$$r_{xy} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}} \quad (9)$$

Using  $amf_t$  and  $aau_t$  of all topics in  $T_{remove}$  with  $ua_t^{(i)}=0$  as  $x_i$  and  $y_i$  in equation (9), the correlation between media focus and user attention can be calculated in each time slot and the inconsistency value is calculated as  $1-r_{xy}$ .

## 4.3 Topic Ranking

Topics are ranked by considering both media focus and user attention values in each time slot online automatically. Therefore we calculate the values of topics for ranking by the combination

of the media focus and user attention values, in the most intuitive way:

$$value_i^{(t)} = mf_i^{(t)\alpha} ua_i^{(t)1-\alpha} \quad (10)$$

where  $\alpha$  ( $0 \leq \alpha \leq 1$ ) can be set different values according to actual needs.

The user attention is usually delayed than the media focus, because news stories are read after the issuance. The question is how soon users' attention is attracted to the news stories issued by news Websites. Our previous work found out that on the day of birth, most news pages accumulate more than 80% of the total page views in their whole lifetime [20]. Anonymous Web page view information on news Websites is collected online all the time, therefore we consider the user attention values calculated are timely enough for topic ranking.

Topics seldom reported by the media or poorly viewed by users recently get very small ranking values according to equation (10). Topics with large values of both media focus and user attention will be ranked as top ones. Since  $mf_i^{(t)}$  and  $ua_i^{(t)}$  of all topics are varying in each time slot, the topics are ranked online automatically.

## 5. EXPERIMENTS AND DISCUSSIONS

Preliminary experiments are firstly performed on a training dataset to find proper values for parameters. Then the analysis on results of calculating inconsistency between media focus and user attention is demonstrated. The discussions of topic ranking results are presented finally.

### 5.1 Dataset and Experimental Setup

Experiments are performed on datasets constructed from practical Web environment. Crawlers are gathering news pages from dozens of Chinese news Websites all the time. The collected news stories are filtered using a keyword list to get only stories about search engine related companies, such as Google (<http://www.google.com/>), Yahoo! (<http://www.yahoo.com/>) and etc. We focus on the search engine related domain because: (1) we're familiar with the domain; (2) the proposed system is initially designed for watching topics of the search engine related domain online. In fact, it is not necessary to consider datasets with multi-domain news because news stories of various categories are usually put in corresponding channels of news Websites.

News stories published from Jan 1 to Oct 31, 2007 are studied in this paper. There are 53,369 stories in total, divided into two parts:

- TrainingSet: contains 14,602 news stories, published from Jan 1 to Mar 31, 2007.
- TestingSet: contains 38,767 news pages, published from Apr 1 to Oct 31, 2007.

Anonymous Web page view information is collected for user attention calculation, from Jan 1 to Dec 31, 2007, which contains over 10.64 billion page views in 1.51 billion Web pages and 78.2 million user sessions. Page view information of news stories in TrainingSet and TestingSet is extracted (0.6 million user page views) to calculate the values of user attention to the generated topics.

### 5.2 Parameter Settings

Preliminary experiments are performed on TrainingSet to find proper values for parameters:  $threshold_{track}$ ,  $\alpha_m$ ,  $\beta_m$ ,  $\alpha_u$  and  $\beta_u$ .

Topics are generated and updated using traditional TDT algorithms from [21].

50 topics are labeled by assessors. The TDT results on the labeled topics are best with  $threshold_{track} = 0.182$ .  $\alpha_m$ ,  $\beta_m$ ,  $\alpha_u$  and  $\beta_u$  are calculated as the method used in [4]: For each topic, a proportion  $r_l$  of the total nutrition corresponds to a proportion  $s_l$  of the total energy. By using two points  $(r_1, s_1)$  and  $(r_2, s_2)$ ,  $\alpha$  and  $\beta$  can be solved. We use the averages of  $\alpha_m$ ,  $\beta_m$ ,  $\alpha_u$  and  $\beta_u$  of 50 topics as the final parameter values:  $\alpha_m = 0.251031$ ,  $\beta_m = 0.001793$ ,  $\alpha_u = 0.008478$  and  $\beta_u = 0.001846$ .

### 5.3 Results of Measuring Inconsistency between Media Focus and User Attention

TestingSet is used to perform the automatic online news topic ranking experiment. The length of time slots is 15 minutes. The topic detection and tracking algorithm described in Section 3 is performed, with 2,517 topics generated totally. It's time consuming to give an overall evaluation on the TDT results because the dataset is from practical Web and in lack of complete annotation. We just trust in the state-of-the-art TDT techniques adopted and the results of preliminary experiments. Some evaluation on TDT results can be found in our previous work [21]. Furthermore, we had 5 assessors (no author is among them) view the results and their feedbacks are summarized as follows: most stories are clustered correctly; a few topics had better be combined as one.

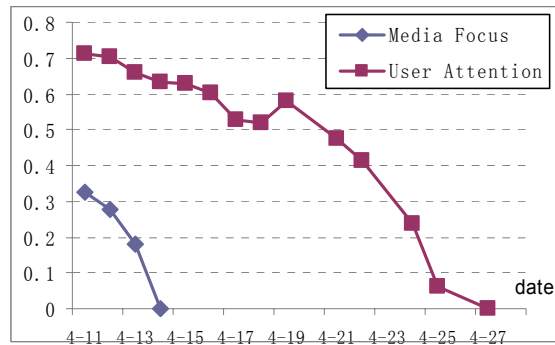
The correlation between media focus and user attention is calculated in the last time slot of each day, as Section 4.2 described. The correlation coefficient on each day is shown as Figure 5:



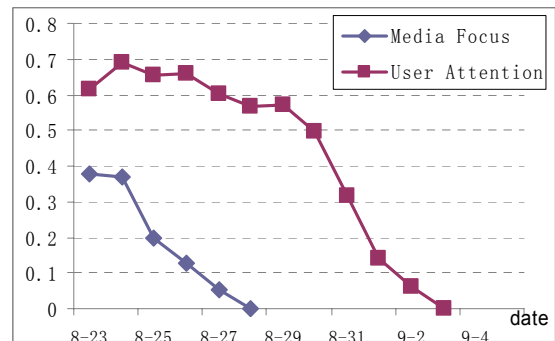
**Figure 5. Pearson correlation coefficient between media focus and user attention on each day from Apr 1 to Oct 31, 2007**

Figure 5 shows that the correlation coefficient between media focus and user attention is unstable at the beginning, but soon gets steadier when more and more topics are included in the calculation. In most cases, the value is between 0.4 and 0.6. Therefore it can be deduced that the inconsistency value, which is calculated as  $1 - r_{xy}$ , is between 0.4 and 0.6. This result confirms that there is indeed inconsistency between media focus and user attention.

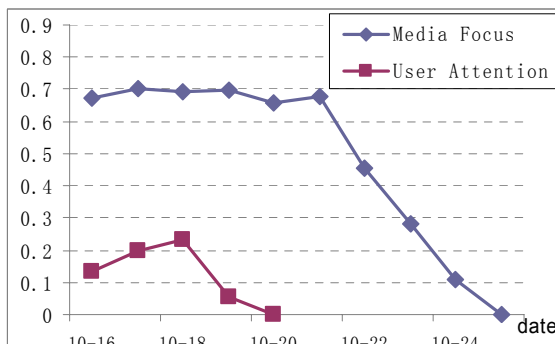
Some topics with discrepant values of accumulative media focus and user attention are shown in Figure 6. Only the values in the last time slot of each day are displayed:



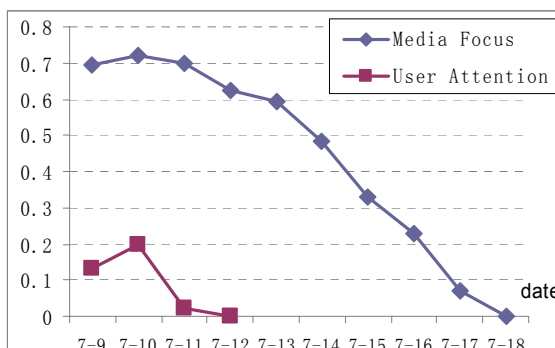
(a) Competition of Chinese input method software: Google, Sogou and Ziguang pinyin



(b) Google's Mayer talks iPhone, Facebook



(c) EU to press Google on DoubleClick deal



(d) Plaintiff in China's first AdSense arbitration case makes second claim

**Figure 6. Some topics with discrepant values of accumulative media focus and user attention**

As show in Figure 6, corresponding topic names are put below the figures. The reasons for inconsistent media focus and user

attention have been discussed in Section 1, and we guess the actual reasons for cases in Figure 6 are due to users' interests in the topics.

## 5.4 Topic Ranking Results

Topics are ranked in every time slot online automatically, using the method described in Section 4.3. The length of time slots is 15 minutes.  $\alpha$  in equation (10) is set as 0.5, treating media focus and user attention equally. Part of the screenshot of the result page of our topic ranking system is shown as Figure 7:



**Figure 7. Top 3 topics on search engine related companies from the topic ranking result page at 8:00 a.m., Oct 26, 2007<sup>1</sup>**

Figure 7 shows the top 3 topics on search engine related companies at 8:00 a. m., Oct 26, 2007, which talk about “the upcoming Alibaba IPO in Hong Kong”, “Baidu to enter C2C E-commerce Market” and “Microsoft invests \$240 million in Facebook”, respectively.

As shown in Figure 7, the titles of the latest news are used as the anchor texts of the hyperlinks linking to topics and news stories currently. Snippets of the latest news story are shown as the summaries of topics. In this way, the up-to-date status of a topic will be viewed. Users will read more about the current progress, background and history through a hyperlink like “all 367 news articles”. Images of related news are added and shown automatically. It is also worth noting that media focus and user attention curves are provided for users to know the topic trends.

The result of topic ranking using both media focus and user attention is compared with that of using only media focus, as shown in Table 2. The names of top 10 topics are demonstrated:

<sup>1</sup> The result of Figure 7 has been translated into English at APPENDIX.



**Table 2. Top 10 topics on search engine related companies at 8:00 a.m., Oct 26, 2007**

Rank	Media focus only	Mixture of media focus and user attention
1	The upcoming Alibaba IPO in Hong Kong	The upcoming Alibaba IPO in Hong Kong
2	Microsoft invests \$240 million in Facebook	Baidu to enter C2C E-commerce Market ( ↑ )
3	Baidu to enter C2C E-commerce Market	Microsoft invests \$240 million in Facebook ( ↓ )
4	Google to announce gPhone in two weeks	Baidu shares continue to go up after a setback ( ↑ )
5	Google shares exceed \$700	Google to announce gPhone in two weeks ( ↓ )
6	Baidu shares continue to go up after a setback	Google shares exceed \$700 ( ↓ )
7	Baidu reports Q3 earnings	Baidu reports Q3 earnings
8	Google takes aim at China's mobile market	Sogou Pinyin 3.0 released ( <i>new</i> ↑ )
9	Google CEO Eric Schmid feels confident of DoubleClick deal	Google denies tax evasion in China ( ↑ )
10	Google denies tax evasion in China	Google takes aim at China's mobile market ( ↓ )

As shown in Table 2, the topic ranking results of using both factors and that of using only media focus are compared. The same topics are in the same colors. The symbols ↑ and ↓ indicate a change of rank by introducing the user focus to the ranking purely by the media focus, and *new* means the topic is not included in the top 10 of the media focus based ranking result, but in the top 10 of the ranking result based on both media focus and user attention. “Sogou Pinyin 3.0 released” is a topic of this kind, which attracts many users’ attention.

The changes shown in Table 2 reflect users’ special interests in some kinds of topics:

- Indigenous companies. For example, users paid special attention to topics on Baidu.
- Things in relation to users. Users gave much attention to Sogou Pinyin, a Chinese input method software, which is useful for them.

It’s difficult to give an accurate evaluation on the topic ranking result. We had 22 users (no author is among them) with various backgrounds (undergraduate students, graduate students and engineers) view the topic ranking result page, part of which is shown as Figure 7. Most of the users agree that the result of the proposed topic ranking algorithm reflects the influence of time, the media and users, which is more reasonable compared to the result of using media focus only.

## 6. CONCLUSIONS AND FUTURE WORK

In this paper, we propose a novel automatic online algorithm for news topic ranking based on an aging theory, using both media focus and user attention. Both media focus and user attention varies as time goes on, so the effect of time on topic ranking has already been included. Inconsistency exists between media focus and user attention, which is analyzed and quantitatively measured in this paper. Topics are ranked by the combination of their media

focus and user attention values online automatically. Related news stories of topics are provided for users’ quick access. Empirical evaluation on the topic ranking result indicates that the proposed topic ranking algorithm reflects the influence of time, the media and users.

The main contributions of this paper are twofold: (1) we present the quantitative measure of the inconsistency between media focus and user attention, which provides a basis for topic ranking and an experimental evidence to show that there is a gap between what the media provide and what users view. (2) To the best of our knowledge, it is the first attempt to synthesize both media focus and user attention into one algorithm for automatic online topic ranking.

In the future, we hope to analyze the topic trends in continuous time slots. We also plan to discover more user behaviors that can be utilized for topic ranking.

## 7. ACKNOWLEDGEMENTS

This work is supported by the Chinese National Key Foundation Research & Development Plan (2004CB318108), Natural Science Foundation (60621062, 60503064, 60736044) and National 863 High Technology Project (2006AA01Z141). The authors would like to thank Bin Liang and Zaihong Qu for their work on evaluation and demonstration. They also thank the anonymous reviewers for their useful comments.

## 8. REFERENCES

- [1] <http://www.nist.gov/speech/tests/tdt/>
- [2] J. Allan, R. Papka, and V. Lavrenko. On-line new event detection and tracking. In Proceedings of the 21st Annual International ACM SIGIR Conference, Melbourne, Australia. ACM Press. 1998, 37-45.
- [3] T. Brants, F. Chen, and A. Farahat. A System for New Event Detection. In Proceedings of the 26th Annual International ACM SIGIR Conference, New York, NY, USA. ACM Press. 2003, 330-337.
- [4] C.C. Chen, Y.T. Chen, Y. Sun and M.C. Chen. Life Cycle Modeling of News Events Using Aging Theory. In Proceedings of 14th European Conference of Machine Learning (ECML ’03), pp. 47-59, 2003.
- [5] K. Y. Chen, L. Luesukprasert and S. T. Chou. Hot topic extraction based on timeline analysis and multi-dimensional sentence modeling. IEEE Trans. on Knowledge and Data Engineering, 2007, 19(8):1016-1025.
- [6] H. L. Chieu and Y. K. Lee. Query Based Event Extraction along a Timeline. In Proceedings of the 27th Annual International ACM SIGIR Conference, Sheffield, UK, ACM Press. 2004, 425-432.
- [7] M. Connell, A. Feng, G. Kumaran, H. Raghavan, C. Shah, and J. Allan. UMass at tdt 2004. In 2004 Topic Detection and Tracking Workshop (TDT’04), 2004.
- [8] G.P.C. Fung, J.X. Yu, H. Liu and P.S. Yu. Time-Dependent Event Hierarchy Construction. In Proceedings of KDD2007, pages 300-309, California, USA, 2007.
- [9] G.P.C. Fung, J.X. Yu, P.S. Yu and H. Liu. Parameter free bursty events detection in text streams. In Proceedings of the 31st VLDB Conference, pages 181-192, Trondheim, Norway, 2005.



- [10] Q. He, K. Chang, and E. P. Lim. Analyzing Feature Trajectories for Event Detection. In Proceedings of the 30th Annual International ACM SIGIR Conference, Amsterdam, the Netherlands. ACM Press. 2007, 207-214.
- [11] Q. He, K. Chang and E. P. Lim. Using Burstiness to Improve Clustering of Topics in News Streams. In Proceedings of the 7th IEEE International Conference on Data Mining, pp. 493-498, 2007.
- [12] T. He, G. Qu, S. Li, and et al. Semi-automatic Hot Event Detection. In Proceedings of the 2nd International Conference on Advanced Data Mining and Applications. 2006, LNAI4093, 1008-1016.
- [13] J. Kleinberg. Authoritative sources in a hyperlinked environment. Proc. 9th ACM-SIAM Symposium on Discrete Algorithms, 1998.
- [14] G. Kumaran and J. Allan. Text Classification and Named Entities for New Event Detection. In Proceedings of the 27th Annual International ACM SIGIR Conference, Sheffield, UK, ACM Press. 2004, 297-304.
- [15] M. Spitters and W. Kraaij. TNO at TDT2001: Language Model-Based Topic Detection. Topic Detection and Tracking Workshop Report, 2001.
- [16] N. Stokes and J. Carthy. Combining Semantic and Syntactic Document Classifiers to Improve First Story Detection. In Proceedings of the 24th Annual International ACM SIGIR Conference, New Orleans. ACM Press. 2001, 424-425.
- [17] R. Swan and J. Allan. Extracting Significant Time Varying Features from Text. In Proceedings of the 8th Conference on Information and Knowledge Management, pages 38-45, 1999.
- [18] R. Swan and J. Allan. Automatic Generation of Overview Timelines. In Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 49-56, 2000.
- [19] D. Trieschnigg and W. Kraaij. Hierarchical topic detection in large digital news archives. In Proceedings of the 5th Dutch Belgian Information Retrieval workshop, 2005.
- [20] Y. Wang, Y. Liu, M. Zhang, S. Ma, Identify Temporal Websites Based on User Behavior Analysis, In Proceedings of 3rd International Joint Conference on Natural Language Processing, Hyderabad, India, 2008.
- [21] C. Wang, M. Zhang, S. Ma and L. Ru. Automatic online news issue construction in Web environment. In proceedings of the 17th international conference on World Wide Web, 2008, 457-466.
- [22] Y. Yang, T. Pierce, and J. Carbonell. A Study of Retrospective and On-line Event Detection. In Proceedings of the 21st Annual International ACM SIGIR Conference, Melbourne, Australia. ACM Press. 1998, 28-36.
- [23] K. Zhang, J. Li, and G. Wu. New Event Detection Based on Indexing-tree and Named Entity. In Proceedings of the 30th Annual International ACM SIGIR Conference, Amsterdam, the Netherlands. ACM Press. 2007, 215-222.
- [24] Y. Zhao and G. Karypis. Criterion Functions for Document Clustering. Technical Report, 2005.

## APPENDIX

### Translation for Figure 7

<b>1</b> <b><u>Frozen capital of Alibaba IPO exceeds 400 billion HK dollars</u></b> Jinyang - 2007-10-26 07:38 Reported by Hong Kong media, Alibaba (1688.HK) IPO receives strong reaction. Besides providing a big draw for investing organizations, the Alibaba IPO attracts passion from private investors (including investors in the mainland). It is said that the subscribe for cash loan from 12 Brokerage Houses is as much as 101.4 billion... <a href="#">It is said that hidden trouble exists behind Alibaba IPO</a> –IT World- (12 same stories) <a href="#">Trouble exists behind Alibaba IPO : subscribe from private investors is over 240 times as much</a> –East Wealth- (2 same stories) <a href="#">Internal placement of Alibaba IPO exceeds \$160 billion</a> –NETEASE- (6 same stories) <a href="#">all 367 news articles &gt;&gt;</a>	<a href="#">More details</a>
<b>2</b> <b><u>Baidu, a titan of search engines, enters C2C</u></b> Tianfu Online – 2007-10-26 07:26 On Oct 17, Baidu announces to enter the electronic commerce market, beginning with C2C, after it entered the news field. It is said that Baidu has established the division of electronic commerce. The division is in charge of the business deals on C2C platform, which is supposed to launch next year. Until now, Alibaba, Tencent, Baidu... <a href="#">Baidu deals with C2C with the help of search techniques</a> –Financial Community - (2 same stories) <a href="#">Baidu announces to enter C2C electronic commerce field</a> –HC360- (2 same stories) <a href="#">Baidu enters C2C: difficulties and ambitions</a> –SINA- (2 same stories) <a href="#">all 153 news articles &gt;&gt;</a>	<a href="#">More details</a>
<b>3</b> <b><u>Microsoft invests in Facebook</u></b> SOHU – 2007-10-26 07:47 Yesterday, Microsoft invests \$240 million for 1.6% of Facebook shares, defeating two major competitors: Google and Yahoo!. Until now, the fight for Facebook between the titans of the Internet industry comes to a conclusion. So why does Facebook catch the titans' eyes?... <a href="#">Google: don't mind losing Facebook, for there are plenty of others</a> –SINA TECH- (8 same stories) <a href="#">Microsoft invests: Facebook options become less attractive</a> –Dagong- (3 same stories) <a href="#">Microsoft invests in Facebook: a threat to Google?</a> –21CN- (5 same stories) <a href="#">all 192 news articles &gt;&gt;</a>	<a href="#">More details</a>