Journal of Computational Information Systems 8: 20 (2012) 8637–8644
Available at http://www.Jofcis.com

A Linear Regression Based News Topic Hotness Calculation Approach

Hui LI*

Xihua University Archives, Chengdu 610039, China

Abstract

News topic rank is an increasing important issue in the field of web information retrieval. News topic hotness quantitatively reflects the degree of attention to a news topic. Users want to know about "what is hot" or "what is happening"; Governments hope to learn about "what are people paying attention to?" However, the vast amount of news are created and updated all the time, and it is almost impossible for users to view them all. So a problem arises: how do we rank news topics so as to return the top ones with high hotness to users? In this paper, we propose a linear regression based news topic hotness calculation approach, in which we compute news topic hotness with linear regression in order to rank news topics according to their hotness. Media focus, user attention and timeliness are considered as news features. Media focus, user attention and timeliness are dependent variables while news topic hotness is independent variable in the linear regression. This method is validated by correlation coefficient test (r test), determination coefficient test (R^2 test), regression coefficient significance test (T test) and regression equation significance test (F test).

Keywords: News Topic Hotness; Linear Regression; Media Focus; User Attention; Timeliness

1 Introduction

Network news has become one of the important channels through which people can learn about social events timely. Users want to know about "what is hot" or "what is happening"; Governments hope to learn about "what are people paying attention to?" However, the vast amount of news are created and updated all the time, and it is almost impossible for users to view them all. So a problem arises: how do we rank news topics so as to return the top ones with high hotness to users? Some search engines companies and others are realizing that news topic ranking automatically may have significant commercial implications. Jike (http://news.jike.com/) search engine displays the importance of news by means of the concept of hotness, which resembles the thermometer. The higher the temperature, the hotter the news is. For example, the hotness of "Wen Jiabao deploys economic work" was 89^0C at 17:35 Oct 30, 2011, which ranked in the first place. How to calculate news topic hotness quantitatively is the focus for our research.

*Corresponding author.
Email address: duyajun@mail.xhu.edu.cn (Hui LI).

October 15, 2012

News should be firstly clustered into topics through some technology, which has been done by means of the method of Formal Concept Analysis (FCA) in our previous work [1]. In this paper, we discuss the problem of computing news topic hotness. A linear regression based news topic hotness calculation approach is proposed, in which news topic hotness is computed with linear regression in order to rank news topics according their hotness. Media focus, user attention and timeliness are considered as news features. Media focus, user attention and timeliness are dependent variables while news topic hotness is independent variable in the linear regression. This method is validated by correlation coefficient test (r test), determination coefficient test (R^2 test), regression coefficient significance test (T test) and regression equation significance test (F test).

The remainder of this paper is organized as follows. We review prior work in Section 2, and continue by introducing three features, media focus, user attention and timeliness in Section 3. News topic hotness calculation model with the method of linear regression is presented in Section 4. Experimental results are described in Section 5. Section 6 concludes the paper and suggests directions for future research.

2 Related Works

There are some researches about news topic ranking [2, 3, 4]. Del Corso GM et al. [2] proposed a ranking news information algorithm, which took in account several criteria such as news freshness, news source authoritativeness and replication of pieces of news. Yang et al. [3] established a quantitative analysis model for calculating the network news force with information retrieval methods, where some factors (such as reply times) were analyzed. Wang et al. [4] ranked news topics using both media focus and user attention based on aging theory, in which the value of energy function indicates the popularity of a news topic.

The recency of news topics can and should influence topic ranking. Dong et al. [5] proposed a recency demotion grading method, which was a subjective judgment by a human about recency, where five judgment grades were applied on query-url pairs. It also used page age as an objective measurement of recency, which is employed in our paper as well. Dai et al. [6] proposed a machine-learning framework for simultaneously optimizing freshness and relevance.

Linear regression is a statistical approach that models the relationship between an dependent variable and one or more independent variables. It has been widely used in many fields [7, 8, 9, 10]. Linear regression has been used to signal processing [7], biomedical research [8], and so on. Mirjana Golusin et al. [9] researched the relationship between ecological and economic subsystems of sustainable development in the region of South Eastern Europe with the method of linear regression. Ye et al. [10] proposed and evaluated a regression-based document re-ranking approach that took into account rich features for boosting information retrieval performance, in which SVM regression model was used to learn a re-ranking function automatically.

User log, an important part of an information retrieval system, records the user's many operations and represents a person's interest to some degree. It is used in many applications, such as query recommendation [11], duplicate web documents detection [12], recommendation of similar users [13], and so on. Bernard J. Jansen et al. [14] investigated their research questions using a large amount of search log data collected from an operational real time search engine.

3 Feature Selection

News topic hotness quantitatively reflects the degree of attention to a news topic. To calculate the value of news topic hotness, we should analyze which features will influence on news topic hotness. Referring to the paper [4], those features mainly include media focus, user attention and timeliness. In this paper, we give different contents to these features.

3.1 Media focus

Media focus (M) reflects the degree of attention to a special news topic from the media perspective. As is known to all, the higher the degree of media focus is, the greater the news topics hotness is. In our paper, M can be got from the number of news stories (N_1) and the number of news comments (N_2) on a special topic.

News stories and news comments, as two different news genres, play distinguishing but important roles in news field [15]. They have both difference and relationship. The difference is that news stories report news event objectively without personal views, while news comments mainly express personal opinions on a special news event subjectively. The relationship between news stories and news comments is that news stories reflect phenomenon of news event, and news comments mine the deep meaning hidden behind phenomenon through expanding and deepening of news stories. They both are on the basis of facts. Additionally, news comments have certain ability on guiding public opinion as well.

Then, how do we distinguish news stories from news comments by computer programs? In our system, we collect some news websites, including central news websites, local news websites and portal websites. Through the analysis and statistics of the dataset, we find that the URLs of a lot of news comments have one of the strings "comment, pl, view, opinionreview". Hence, news stories and news comments can be simply separated by their URLs. In our system, we use this method to obtain the number of news stories (N_1) and the number of news comments (N_2) on a special topic and then get the value of media focus (M). M will increase when either or increases. So, we can get

$$M = N_1 + N_2 \tag{1}$$

where N_1 and N_2 are the number of news stories and the number of news comments respectively.

3.2 User attention

User attention (U) reflects the degree of attention to a special topic from users' perspective. The higher the degree of user attention is, the greater the news topics hotness is. In our paper, U is got from the user log. User log which records the user's many operations represents a person's interest to some degree. In our system, user log is mainly used to record user's search queries because user's search queries contain a great variety of human goals [16]. It is a text file named userLog.txt. In this log, each record has four fields:

- User identification: username of current user.
- Date: the date of the interaction.

- Search time: measured in hours, minutes, and seconds as recorded by the server on the date of the interaction.

- Search querie: the query that entered by the user.

For example, "cpfryy 2011-7-27 15:46:00 bullet trains rear-end accident" in a line.

If a term is frequently searched by users among a few days, it is mostly possible that the news topic to which the term belongs is a hot issue. Consequently, U can be got from the Eq. (2)

$$U = N_{userlog} \tag{2}$$

where $N_{userlog}$ is the number of terms which belong to a special news topic in the user log.

3.3 Timeliness

Timeliness (T) is the third feature in this paper. It is also called freshness or recency in some literatures [5, 6]. As is known to all, the importance of news topic will change over time. A topic which is being discussed by people at this second would be quickly replaced by another new emerging topic at next second. Timeliness is the nature of news. Hence, fresh news topic should be considered more important than an old one.

A number of news about a special news topic will emerge each day when the topic is being discussed by people. These news should be firstly clustered into news topics using clustering method. We order these news in time sequence, and then the freshness of a news topic can be reflected by the latest news. How do we find the latest news about a special news topic? Referring to the paper [5], the freshness of a piece of news can be represented by the elapsed time, which is defined as Eq. (3)

$$\Delta t = t_{now} - t_{news} \tag{3}$$

where t_{now} is the current time of our system. t_{news} is the news publication date which is extracted by regular expressions. This elapsed time is also called *page age* which is an objective measurement of timeliness [5].

The publication date t_{news} of the latest news is the latest and its page age Δt is the smallest. So, the freshness of a piece of news can be got from Δt. For example, t_{now} =2011-3-15 11:24, t_{news} =2011-3-15 11:04, and then $\Delta t = t_{now} - t_{news} = 20$. The range of Δt is $[0, 1439]$. In this paper, we let the timeliness of a news topic equal the freshness of the latest news about this topic. Its mean is that the topic is still being discussed if a piece of news about it emerges. The timeliness of a news topic in a day is calculated by Eq. (4). The purpose of adding one is to insure that T does not equal zero.

$$T = \frac{\min\{\Delta t\} + 1}{1440} \tag{4}$$

News topic hotness quantitatively reflects the degree of attention to a news topic. This value can not be watched directly, so should be set manually according to the fact of the topic on Internet. It is fell into ten grades and the value is the integer one to ten. One indicates that the hotness is the lowest, and ten is the highest.

4 A Linear Regression Based News Topic Hotness Calculation Approach

4.1 Linear regression

Linear regression is one of the most widely used methods for relating a dependent variable (normally referred to as Y) to a set of independent variables referred to as $X_1, X_2, \cdots, X_n$ [17]. According to the number of independent variables, linear regression can be divided into two categories, "simple linear regression" and "multiple linear regression".

The simple linear regression equation is of the form:

$$Y = aX + b \tag{5}$$

where Y is dependent variable; X is independent variable; a is the coefficient; b is constant term.

Multiple linear regression model is a generalization of simple linear regression where there are more than one independent variable. The aim of multiple linear regression is to explore and quantify the relationship between a dependent variable and one or more independent variables [17]. A multiple linear regression model has the following structure:

$$Y = a_1X_1 + a_2X_2 + \cdots + a_nX_n + b \tag{6}$$

where Y is dependent variable; $X_1, X_2, \cdots, X_n$ are independent variables; $a_1, a_2, \cdots, a_n$ are coefficients; and b is constant term.

Then, how do we obtain these unknown parameters? The most common method is the least squares method. It minimizes the sum of squared differences between the data values and their corresponding modeled values to seek the best linear equation fitting to the dataset.

When the dataset is large, it is fairly complex to compute these parameters manually and hence some tools emerge. LINEST function in Microsoft Excel 2003 has implemented least squares method, which accepts a dataset and returns an array that includes those parameters what we need [18]. LINEST function is used in our experiments.

4.2 The calculation of news topic hotness

There are three factors in our paper, media focus (M), user attention (U) and timeliness (T), so it belongs to the issue of multiple linear regression. On the basis of multiple linear regression, we calculate news topic hotness in three steps.

Step 1 Confirm dependent variable and independent variables.

News features which determine news topic hotness have been described in Section 3. In this paper, dependent variable is news topic hotness H; independent variables are media focus M, user attention U and timeliness T. Correlation coefficient test (r test) is used to test the independence of M, U and T.

$r_{0.05} < |r| < r_{0.01} \Longrightarrow$ The linear relationship between two variables is significant;

$|r| > r_{0.01} \Longrightarrow$ The linear relationship between two variables is extremely significant;

$|r| < r_{0.05} \Longrightarrow$ The linear relationship between two variables is not obvious.

Step 2 Confirm multiple linear regression equation.

A dependent variable H can be written as a function of independent variables M, U and T, with a constant value b in the form of linear equation such as below:

$$H = a_1 M + a_2 U + a_3 T + b \tag{7}$$

where b = a constant value; a_1 = the effect of media focus on news topic hotness; a_2 = the effect of user attention on news topic hotness; a_3 = the effect of timeliness on news topic hotness. LINEST function in Microsoft Excel 2003 is used to get a_1, a_2, a_3 and b.

Step 3 Test regression equation. The tests of multiple linear regression include Determination coefficient test (R^2 test), Regression coefficient significance test (T test) and Regression equation significance test (F test).

- Determination coefficient test (R^2 test): this test is used to test the close degree of linear relationship between dependent variable and all independent variables.
- Regression coefficient significance test (T test): this test is used to test whether the linear relationship between dependent variable and each independent variable is significant. The t absolute value of an independent variable should be greater than t critical value so that the independent variable is available.
- Regression equation significance test (F test): this test is used to test whether the linear relationship between dependent variable and the whole of all independent variable is significant.

5 Experiment

We collect webpages about the news topic "ractopamine" as our dataset. The date is from March 15, 2011 to March 31, 2011. The total number is 154. And then calculate the values of M, U, T. Set the value of H manually. The total days is 17 from March 15, 2011 to March 31, 2011.

(1) Correlation coefficient test (r test)

The number of the sample is 17 and the degree of freedom is 15. The CORREL function in EXCEL is used to calculate correlation coefficient r. $r_{0.05}(15) = 0.4821$, $r_{0.01}(15) = 0.6055$

$|correl(M,U)| = |0.43554896| < 0.4821 \Longrightarrow$ The linear relationship between M and U is not obvious.

$|correl(M,T)| = |-0.38751251| < 0.4821 \Longrightarrow$ The linear relationship between M and T is not obvious.

$|correl(U,T)| = |-0.12768247| < 0.4821 \Longrightarrow$ The linear relationship between U and T is not obvious.

Therefore, M, U and T are independent.

(2) Confirm multiple linear regression equation.

$a_1 = 9$; $a_2 = 4$; $a_3 = 4$; $b = -1.19481927$;

$$H = 9M + 4U + 4T - 1.19481927 \tag{8}$$

(3) Determination coefficient test (R^2 test)

$R^2 = 0.79618491 \Longrightarrow$ the linear relationship between dependent variable and each independent variable is significant.

(4) Regression coefficient significance test (T test)

The number of variable is 4 (H,M,U,T), the number of the sample is 17, and the degree of freedom is 13. $T_{0.025}(13) = 2.16$. The t absolute values of a_1, a_2, a_3 are 5.008912977, 2.474560591, 2.23775501 respectively. The three values are all greater than 2.16, so they are available.

(5) Regression equation significance test (F test)

$F_{0.05}(3, 13) = 3.41$; $F_{0.01}(3, 13) = 5.74$; $F = 9277678 > 5.74 \Longrightarrow$ the linear relationship between H and the whole of M, U, T is extremely significant.

All above tests indicate that Eq. (8) is reasonably. With Eq. (8), we can predict the hotness of a special news topic. Similarly, the hotness of other news topics can also be calculated in the same way. And then news topics can be ranked by their hotness to get hot news topics.

6 Conclusion and Future Research

In this paper, we compute news topic hotness with linear regression in order to rank news topics according their hotness. Media focus, user attention and timeliness are considered as news features. Media focus, user attention and timeliness are independent variables while news topic hotness is dependent variable. This method is validated by correlation coefficient test (r test), determination coefficient test (R^2 test), regression coefficient significance test (T test) and regression equation significance test (F test).

There are several areas of further works. On the media side, we can enlarge the range from news to microblog, bbs and so on. We can also add the celebrity effect to comments. On the user side, it would be interesting to investigate what are attractive to each user through query log and then we can research personal recommendation system. On the system side, we can integrate this module to search engines and other information retrieval systems.

Acknowledgement

This work is supported by the National Natural Science Foundation of China (Grant No. 60872089), the Fund of Key Discipline of Computer Software and Theory, Xihua University (Grant NO. SED0802-09-1), the Science and Technology Innovation Seedling Project of Sichuan (Grant No. 2010-019), the Innovation Fund of Postgraduate, Xihua University (Grant No. YCJJ201148), and the Cup of Xihua University in 2011.

References

[1] Y. Y. Ren, Y. J. Du, X. P. Huang, Y. Xu. Topic Detection of News Stories with Formal Concept Analysis [J], Journal of Information and Computational Science 8: 9 (2011), 1675 – 1682.

[2] Del Corso GM, Gulli A, Romani F. Ranking a stream of news [C]. In: Ellis A, Hagino T, eds, Proc. of the WWW 2005. New York: ACM, 2005, pages 97 – 106.

[3] W. J. Yang, R. W. Dai, X. Cui. Model for Internet News Force Evaluation Based on Information Retrieval Technologies [J], Journal of Software, vol. 20, no. 9, September 2009, pages 2397 – 2406.

[4] C. H. Wang, M. Zhang, L. Y. Ru, S. P. Ma. Automatic Online News Topic Ranking Using Media Focus and User Attention Based on Aging Theory [C], in: Proceedings of the 17th ACM conference on Information and knowledge management (CIKM), New York, USA, 2008, pages 1033 – 1042.

[5] A. Dong, Y. Chang et al. Towards recency ranking in web search [C], in: WSDM'10, New York, USA, 2010, pages 11 – 20.

[6] N. Dai, M. Shokouhi, B. Davision. Learning to rank for freshness and relevance [C], in: SIGIR'11, July 24 – 28, 2011, Beijing, China.

[7] Ciprian Doru Giurc?neanu, Seyed Alireza Razavi, Antti Liski. Variable selection in linear regression: Several approaches based on normalized maximum likelihood [J], Signal Processing, Volume 91, Issue 8, August 2011, pages 1671 – 1692.

[8] Adam G. Polak. Analysis of multiple linear regression algorithms used for respiratory mechanics monitoring during artificial ventilation [J], Computer Methods and Programs in Biomedicine, Volume 101, Issue 2, February 2011, pages 126 – 134.

[9] M. Golusin, O. M. Ivanovic, N. Teodorovic. The review of the achieved degree of sustainable development in South Eastern Europe-The use of linear regression method [J], Renewable and Sustainable Energy Reviews, Volume 15, Issue 1, January 2011, pages 766 – 772.

[10] Z. Ye, Jimmy X. J. Huang, H. F. Lin. Incorporating rich features to boost information retrieval performance: A SVM-regression based re-ranking approach [J], Expert Systems with Applications, 38 (2011), 7569 – 7574.

[11] Y. Q. Liu, J. W. Miao, M. Zhang, S. P. Ma, L. Y. Ru. How do users describe their information need: Query recommendation based on snippet click model [J], Expert Systems with Applications, Volume 38, Issue 11, October 2011, pages 13847 – 13856.

[12] F. Radlinski, P. N. Bennett, E. Yilmaz. Detecting duplicate web documents using clickthrough data [C], in: WSDM'11, February 9 – 12, 2011, Hong Kong, China.

[13] Pasquale De Meo, Antonino Nocera, Giorgio Terracina, Domenico Ursino. Recommendation of similar users, resources and social networks in a Social Internetworking Scenario [J] Information Sciences, Volume 181, Issue 7, 1 April 2011, pages 1285 – 1305.

[14] Bernard J. Jansen, Zhe Liu, Courtney Weaver, Gerry Campbell, Matthew Gregg. Real time search on the web: Queries, topics, and economic value [J], Information Processing & Management, Volume 47, Issue 4, July 2011, pages 491 – 506.

[15] J. J. Wang. Differences and relationship between news stories and news comments [J], Legal System And Society. 2010 (14), page 238.

[16] Markus Strohmaier, Mark Kröll. Acquiring knowledge about human goals from Search Query Logs [J], Information Processing & Management, Available online 3 June 2011.

[17] M. M. Rodríguez del águila, N. Benítez-Parejo. Simple linear and multivariate regression models [J], Allergol Immunopathol (Madr). 2011, 39 (3), 159 – 173.

[18] LINEST function in Excel [EB/OL], http://office.microsoft.com/zh-cn/excel-help/HP010069838.aspx.