A Novel Combine Forecasting Method for Predicting News Update Time

Mengmeng Wang
College of Computer Science and Technology, Jilin University, Changchun, China
Key Laboratory of Symbolic Computation and Knowledge Engineering attached to the Ministry of Education, Jilin Unversity, Changchun, China
Email: wmmwwlh@126.com

Xianglin Zuo
College of Computer Science and Technology, Jilin University, Changchun, China
Email: 295228473@qq.com

Wanli Zuo and Ying Wang
College of Computer Science and Technology, Jilin University, Changchun, China
Key Laboratory of Symbolic Computation and Knowledge Engineering attached to the Ministry of Education, Jilin Unversity, Changchun, China
Email: { zuowl, wangying2010}@ jlu.edu.cn

***Abstract*—With the rapid development of Internet, information provided by the Internet has shown explosive growth. In the face of massive and constantly updated information on the Internet, how the user can fast access to more valuable and more information has become one of the hot spots. The time of Web Page update appears to be erratic, so forecasting the update time of news reports is even more difficult. From the view of application, we can use mathematical models to maximize the approximation of variation, although it cannot be completely accurate. So is the predicting the update time of news which helps in improving the news crawler's scheduling policy. In this paper, we proposed a combined predict algorithm for news update. In order to predict the update time of news, firstly, we applied the Exponential Smoothing method to our dataset, and we also have selected the optimal parameters. Secondly, we leveraged the Naive Bayes Model for prediction. Finally, we combined two methods for Combination Forecasting, as well as made a compare with former methods. Through the experiments on Sohu News, we show that Combination Forecasting method outperforms other methods while estimating localized rate of updates.**

***Index Terms*—Exponential Smoothing Method, Naive Bayes Model, Combination Forecasting, News Update Time**

I. INTRODUCTION

News provides information on recent events, and therefore timeliness is very important to the news.

Timeliness means that report and the fact are taking place synchronized in order to meet the needs of audience. The rapid development of Internet technology make demands of the real-time grow geometrically. The network news media has been an unprecedented development. Nicholas Negroponte once said, "Network media is the traditional media's grave digger". A recent survey shows that about 90% of decision information can be acquired from the web[1].

Web is growing explosively[2]. And it is almost impossible to download all novel pages. The update time of web page appears to be erratic, news are extremely time sensitive by nature, so forecasting the update time of news reports is even more difficult. From the view of application, we can use mathematical models to maximize the approximation to variation, although it cannot be completely accurate. In the face of massive and constantly updated information on the Internet, predicting the time of news page update helps in improving the news crawler's scheduling policy. Fetterly et al.[3] analyzed several million of pages with the aim of measuring the rate and the degree of changes to web pages. The statistical observations of the measurements showed that page size was a strong predictor of both frequency and degree of change. Real-time Web crawling system leveraged the active crawling, the updated time of news pages is unknown, but in order to maximize the approximation to news page update frequency, fixed cycle crawling obviously does not work, but to crawl in the dynamic frequency which should continue to be adjusted in the application environment. Focused crawlers only download pages related to a given topic[4][5][6][7]. These works are similar with ours. We also make several kinds of focused crawler, and each crawler is in charge of one kind of news.

In this paper, we proposed a combined predict algorithm about news update. In order to predict the update time of news, firstly, we applied the Exponential

Corresponding author: Wanli Zuo
Tel.: +1-359-608-5187
E-mail address: zuowl@jlu.edu.cn

doi:10.4304/jsw.7.12.2787-2793

Smoothing method to our dataset, and we also have selected the optimal parameters. Secondly, we leveraged the Naive Bayes Model for prediction. Finally, we combined two methods for Combination Forecasting, as well as made a compare with former methods. Through the experiments on Sohu News, we show that Combination Forecasting achieve better compared to the other two methods. Our study differs from previous studies in that we derived Combination Forecasting which combined Naive Bayes Model and Exponential Smoothing to predict the time of news update.

Roadmap for rest of this paper is organized as follows: Some related work is discussed in section 2;Section 3 briefly describes the problem formulation; Section 4 introduces the approach proposed; experiment and the result is analyzed in Section 5; Section 6 is the conclusion of this paper.

II. Related Work

The rapid development of the Web2.0 technology put forward higher requirements on the timeliness issue. The flood of information in news, blogs and micro blogs is explosive, undergoing rapid changes and changes over time. Explosive events which happen in the morning may demise at noon, if do not pass this information to user until afternoon, the user is already not interested.

The time-sensitive information can be divided into two forms: static time-stamp information and dynamic time-stamp information. For instance, news belongs to static time-stamp information. The information which does not change over time is only related with a point in time or time period. Once generated, it is tightly bounded with time, and with time variation of it only changes at the moment it is produced, namely, the process from scratch over time. Although such information is no longer changed after generation, when and where they are generated is random. It is an almost impossible task that fully grasps the precise information[8][9].

Forecasting page update frequency is a very difficult task and related works have been published very early. Their research can be traced back to the seventies of the last century which mainly focus on forecasting update frequency. There are two main approaches in the study of the variation of the page: A kind of method is based on the experimental method of the Web page for Web sampling. Through collecting and checking sample to study the change rule of the Web so as to estimate the change rule, such as the work in[10][11][12][13][14][15][16][17]; Another kind of method is a more classic algorithm, Establish poisson mathematical model, then carry on the analysis and argument, and verify the model by experiments and estimate related parameters, so as to predict the time of page changes, such as the work in [10][11][13][14][18][19][20]. Poisson distribution is often used for modeling a series of the probability of stochastic time series which happened independent at a fixed speed.

Nowadays, it is widely to simulate the behavior of the website update through the poisson distribution model, from the thought put forward, it is a long time researching on it, but it is difficult to have substantial breakthrough and progress, and web behavior fitting accuracy and stability is not very high through the inornate mathematical model forecast.

Ashutosh Dixit and A.K. Sharma[21] have proposed architecture of incremental web crawler which manages the process of revisiting of a web site with a view to maintain fairly fresh documents at the search engine site. The computation of update time helps in improving the effectiveness of the crawling system by efficiently managing the revisiting frequency of a website.

Niraj Singhal, Ashutosh Dixit, and Dr. A. K. Sharma[22] has developed a novel method for computing the revisiting frequency that helps crawler to remove its deficiencies by dynamically adjusting the revising frequency thereby improving the efficiency. The proposed mechanism not only reduces the network traffic but also increases the quality by providing the updated pages. In our work, we continue to adjust the news update time in the application environment by leveraging Combination Forecasting method.

III. Problem Formulation

We present Combination Forecasting to predict the time of news update, a novel method which combined Naive Bayes Model and Exponential Smoothing.

Fig. 1 shows the architecture of Combination Forecasting. The input items are the news files extracted from the Sohu News pages. These files have been preprocessed and converted to XML files. The first component of the system is predictor 1 that predicts the time of news update in Exponential Smoothing method. The second component of the system is predictor 2 that predicts the time of news update in Naive Bayes Model. The outputs of predictor 1and predictor 2 are sent to the final predictor to compute the combined result. Then crawlers can download pages according to the result.

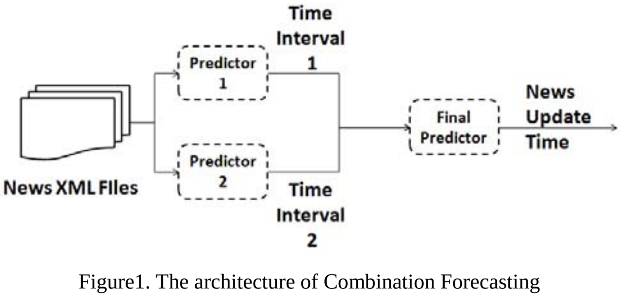


Figure1. The architecture of Combination Forecasting

IV. Combination Forecasting Method

A.Exponential Smoothing Method

The Exponential Smoothing method is proposed by Robert G. Brown. Brown believes that the time series trend has stability or regularity, hence the time series can be reasonably homeopathic postponed; He holds the opinion that the trend of the recent past situation will continue in the future in a way, thus the larger weight should be put on recent data.

The exponential smoothing method is a common method of production forecasts. Also used for short-term economic trends. Of all prediction methods, exponential smoothing method is most used. The full-period average method used time series data equivalently left out of all of them; the moving average rule does not consider the longer-term data, and given the recent data more weight in the weighted moving average method. Exponential smoothing method is compatible with the full-period average and moving average methods for it does not abandon the past data, but gives it a diminishing extent.

B. Naive Bayes Model

Naïve Bayes model is consisted of a tree Bayesian network, which contains a root node and a number of leaf nodes. Naïve Bayes model uses probability to represent all forms of uncertainty, dose learning and reasoning processes by the probabilistic rules. Naïve Bayes model is based on Bayes' theorem, reducing the computational cost through the conditional independence assumptions. Predicting unknown sample data belongs to the highest posterior probability of class standard.

C.Combination Forecasting Method

Combination Forecasting uses two or more different prediction methods for the same problem. It can be a combination of several quantitative or qualitative methods, however, a combination of qualitative and quantitative methods is often used. The main purpose of the combination is leveraging the information provided by various methods for the sake of improving the prediction accuracy as much as possible.

The combination forecast has two basic forms:

(a) Equivalent Weigh Combination Forecasting, namely combine into a new predictive value of the predictive value of each prediction method according to the same weights;

(b) Non-equivalent Weigh Combination Forecasting, that is, the weight given to the predictive value of different prediction methods is not the same.

The principles and application of these two forms are equal, but the different weights taken. In our work, we combined Exponential Smoothing method and Naive Bayes Model for Combination Forecasting.

$$S_t = \alpha y_t + (1-\alpha)S_{t-1} \tag{1}$$

Exponential smoothing method, in its simplest form such as (1), involves successive applications of the formula, where, in our work, y_t is the value of a time interval between two latest news update time and S_t is a 'smoothed' value representing the next time interval of news update and $0 \le \alpha \le 1$.

The calculation of the news updates also in line with the Naive Bayes model which is for the type of news and the size of news site to decide, where the news update time interval can be seen as the root node in the Naive Bayes model, the type of news and the size of news site can be seen as a leaf node in the Naïve Bayes model. Therefore, the news updates time interval computing such as (2) below:

$$\begin{aligned} P(N_i = i \mid N_t = j, W_s = h) &= \frac{P(N_i = i, N_t = j, W_s = h)}{P(N_i = i)P(N_t = j)P(W_s = h)} \\ &= \frac{P(N_i = i)P(N_t = j \mid N_i = i)P(W_s = h \mid N_i = i)}{P(N_i = i)P(N_t = j)P(W_s = h)} \end{aligned} \tag{2}$$

Where i, j, h are the value of corresponding discrete interval, Ni, Nt and Ws stand for the news update time interval, the type of news and the size of news site respectively. And they are defined as follows:

Ni, is the interval between the news next update time and the news latest update time;

Nt, is the type of news, including IT news, house news, health news, education news, economic news, travel news, media news, auto news, fashion news, sports news, culture news, game news and entertainment news;

Ws, is the number of occurrences of news sites in the training dataset. And the model is presented in Fig. 2.

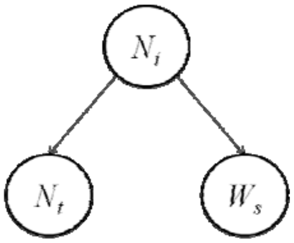


Figure2. The Naïve Bayes model of the news update time interval

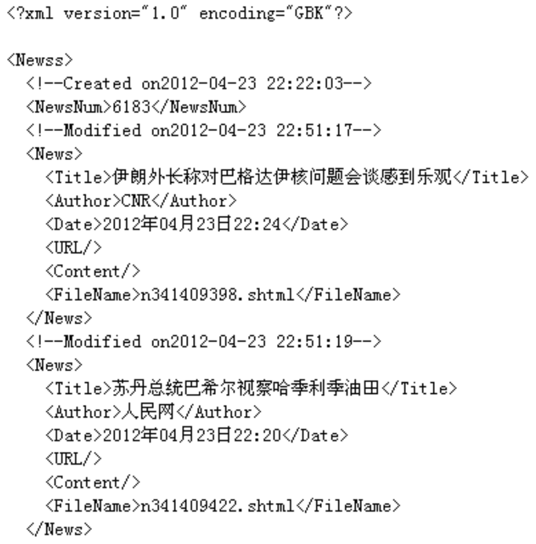
We will discuss the value of α, discrete interval and the parameter of Combination Forecasting in the next section.

V. Experiment and Evaluation

A.Dataset

Since the updates rate of the sources may vary with time, localized estimation provides more detail, useful and accurate information. We have chosen the Sohu news from April 23, 2012 to June 23, 2012, including two months data, as our data set. Among these data, we have chosen the news from April 23, 2012 to May 16, 2012 as the training dataset and chosen news from May 16, 2012 to June 23, 2012 as the test dataset.

The news is sorted out into 13 categories. Each kind of news has a XML file, file format is presented in Fig. 3.All kinds of news have the same file format.



```xml
<?xml version="1.0" encoding="GBK"?>

<Newss>
  <!--Created on2012-04-23 22:22:03-->
  <NewsNum>6183</NewsNum>
  <!--Modified on2012-04-23 22:51:17-->
  <News>
    <Title>伊朗外长称对巴格达伊核问题会谈感到乐观</Title>
    <Author>CNR</Author>
    <Date>2012年04月23日22:24</Date>
    <URL/>
    <Content/>
    <FileName>n341409398.shtml</FileName>
  </News>
  <!--Modified on2012-04-23 22:51:19-->
  <News>
    <Title>苏丹总统巴希尔视察哈季利季油田</Title>
    <Author>人民网</Author>
    <Date>2012年04月23日22:20</Date>
    <URL/>
    <Content/>
    <FileName>n341409422.shtml</FileName>
  </News>
```

Figure3. XML file of the IT news

B.Experiments

1) Exponential Smoothing

For the exponential smoothing forecasting methods, the smoothing constant determines the level of smooth and the response speed of the differences between predicted values and actual values.

The more the smoothing constant α is close to 1, the less impact the long-term actual value has on the decreasing rate of the current smoothed value. The more the smoothing constant α is close to 0, the more impact the long-term actual value has on the decreasing rate of the current smoothed value. Hence, when the time series is relatively stable, a lager α should be selected; A smaller α should be selected when the time series fluctuate, so as to not ignore the influence of the long-term value. Thus, to evaluate the effects of α , we compare the results by assigning different values to α : 0.5, 0.05, 0.005, 0.0005 and 0.00005. Table Ⅰ compares the MAD (Mean Absolute Difference) of different parameters, where the unit of MAD is minute. Experimental results demonstrate that 0.0005 was the optimal parameter.

2) Naive Bayes Model

For the Bayesian approach, News Type is a discrete attribute, but in view of some variables (NewsTime and WebSize) are continuous attributes, in order to calculate the conditional probability of continuous attributes, the Naive Bayes model provides two methods:

(a)Make the continuous attributes discrete and use of discrete intervals instead of continuous attributes;

(b)Leverage probability distribution function to calculate.

We intend to use the first method for the calculation of conditional probability. The discrete intervals of Ni are defined as fast, middle and slow, while large, middle and small are the intervals of Ws. The discrete interval of each attribute is defined as shown in Table Ⅱ. The parameter values are determined according to the experiment.

In line with probability distribution of the news updates, update frequency is divided into three levels: fast, middle and slow. Consequently achieve the forecast on the news updates.

TABLE I.

THE MAD (MEAN ABSOLUTE DIFFERENCE) OF DIFFERENT PARAMETERS

News Type	0.5	0.05	0.005	0.0005	0.00005
IT news	71	61	45	44	44
house news	54	45	34	34	34
health news	39	31	25	25	25
education news	44	36	28	28	28
economic news	44	38	30	30	30
travel news	64	56	44	44	44
media news	69	60	45	45	45
auto news	45	36	29	29	29
fashion news	58	50	38	38	38
sports news	167	155	117	117	117
culture news	68	62	48	48	48
game news	100	92	62	62	62
entertainment news	47	38	29	29	29

3) Combination Forecasting

Finally, we use Non-equivalent Weigh Combination forecasting, assigning weights to exponential smoothing, and Bayesian methods. Tuple $\omega = (1-\beta, \beta)$ is used to represent assignment of the weight, where β is the weight assigned to the exponential smoothing method, $1-\beta$ is the weight assigned to the Bayesian model. We compared the results obtained from setting different values to the parameter ω, where these values are (0.85, 0.15), (0.95, 0.05), (0.98, 0.02), (0.99, 0.01) , (0.995, 0.005). Table Ⅲ has shown the MAD (Mean Absolute Difference) of different parameters, where the unit of MAD is minute. Experiments showed that ω = (0.99, 0.01) was the optimal parameter.

C. Evaluation

The comparisons of different estimators are based on the same datasets. The following experiment compares our Combination Forecasting method to other baseline methods Exponential Smoothing method and Naive Bayes Model discussed above.

Table Ⅳ shows the the MAD (Mean Absolute Difference) for each method, where the unit of MAD is minute. As illustrated in Table Ⅳ, Combination Forecasting method, which leverages the information provided by various methods, achieved the best performance.

In Table Ⅳ, ES, B and CF stand for Exponential Smoothing method, Naive Bayes Model and Combination Forecasting method respectively. Due to space limitations, Fig. 4(a), (b) and (c) merely show the predict time and accurate time of media news for each method with the optimal parameter. We can make a conclusion that the Combination Forecasting method which we proposed outperforms other methods for most of the cases.

The follow diagram can be obtained with the statistical of variation on collecting pages, where the X axis is the number of the news, the Y axis is the value of the update time of news by the log function. '*' stands for predict news update time and '-' stands for accurate news update time. From Fig.4 (a), (b) and (c) we found that the update of the news has obvious temporal locality regular pattern, which is similar to Tao Meng et al[23].

VI. CONCLUSION

The update time of web page appears to be erratic, news are extremely time sensitive by nature. From the view of application, we can use mathematical models to forecast the update time of news reports, although it can not be completely accurate. Predicting the time of news page update helps in improving the news crawler's scheduling policy. In this paper, we proposed a new predict policy for news updates. In order to predict the time of news updates, firstly, we applied the Exponential Smoothing method to our dataset, and we also have

TABLE II.

THE OPTIMAL DISCRETE INTERVAL OF EACH ATTRIBUTE

Variable	Interval1	Interval2	Interval3
Ni	fast[0-134]	middle(134,853]	slow(853,+∞)
Ws	large(62,+∞)	middle(11,62]	small[0,11]

TABLE III.

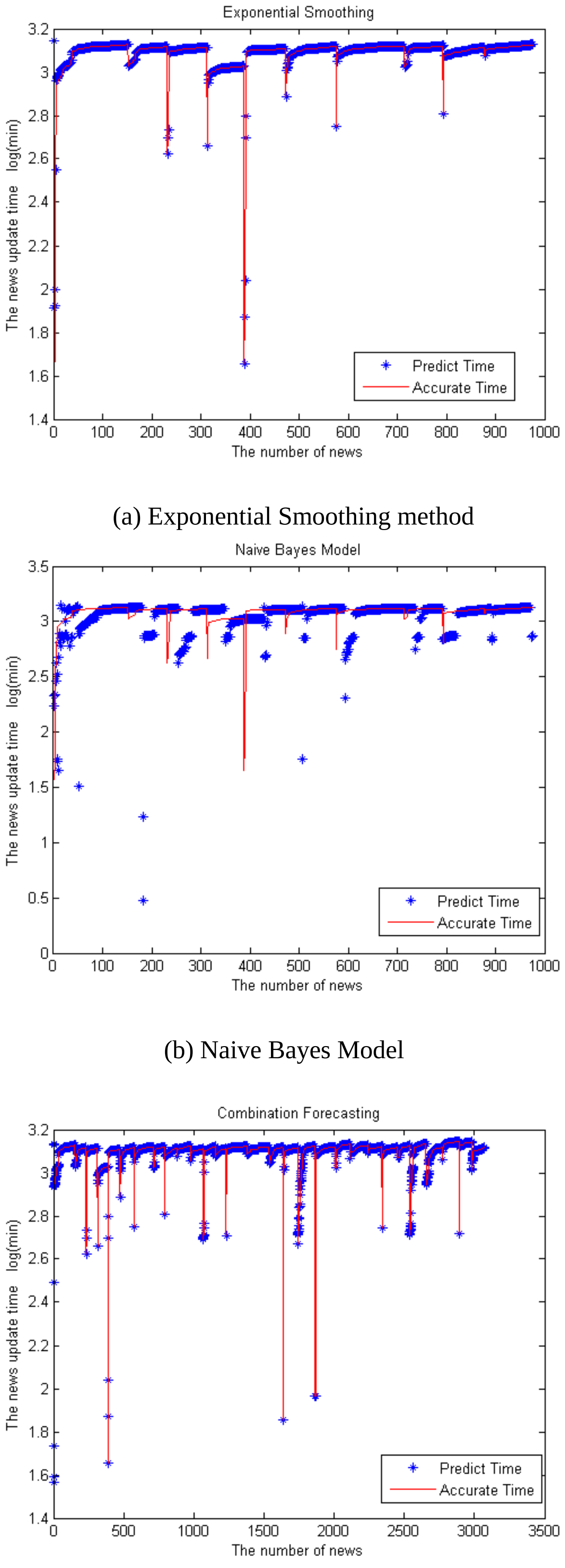
THE MAD (MEAN ABSOLUTE DIFFERENCE) OF DIFFERENT PARAMETERS

News Type	(0.85,0.15)	(0.95,0.05)	(0.98,0.02)	(0.99,0.01)	(0.995,0.005)
IT news	26	21	20	20	20
house news	19	19	19	19	19
health news	21	17	17	17	18
education news	22	21	21	21	21
economic news	11	9	9	8	8
travel news	22	21	21	21	21
media news	36	21	17	15	15
auto news	48	25	19	18	18
fashion news	35	25	23	23	23
sports news	77	31	20	17	17
culture news	31	24	23	23	23
game news	60	60	60	60	60
entertainment news	22	19	18	17	17

TABLE IV.

THE BEST VALUE OF MAD (MEAN ABSOLUTE DIFFERENCE) OF DIFFERENT METHODS

Method	IT	house	health	education	economic	travel	media	auto	fashion	sports	culture	game	entertainment
ES	44	34	25	28	30	44	45	29	38	117	48	62	29
B	369	184	251	224	254	241	684	546	374	1045	459	147	425
CF	20	19	17	21	8	21	15	18	23	17	23	60	17



(a) Exponential Smoothing method

(b) Naive Bayes Model

(c) Combination Forecasting

Figure4. The predict time and accurate time of media news in three methods

selected the optimal parameters. Secondly, we leveraged the Naive Bayes Model for prediction. Finally, we proposed a new method, which combined the above methods for Combination Forecasting, as well as made a compare with the other two methods. In a scenario with inconsistent rate of updates, Combination Forecasting provides more detail and useful information compared to global estimation. Tests on datasets confirm that the proposed Combination Forecasting method outperforms the case in which uses Exponential Smoothing method or Naive Bayes Model only, while estimating localized rate of updates.

ACKNOWLEDGEMENTS

This work is supported by the National Natural Science Foundation of China under Grant No.60973040; the National Natural Science Foundation of China under Grant No.60903098; the basic scientific research foundation for the interdisciplinary research and innovation project of Jilin University under Grant No.201103129; the Science Foundation for China Postdoctor under Grant No.2012M510879.

REFERENCES

[1] S. Thompson, C. Y. Wing. Assessing the Impact of Using the Internet for Competitive Intelligence. *Information & Management*, 2001.

[2] Brewington, B. & Cybenko, G. How Dynamic is the Web. *Proceedings of WWW –9th International World Wide Web Conference*, 2000.

[3] D. Fetterly, M. Manasse, M. Najork, and J. Wiener. A Large-Scale Study of the Evolution of Web Pages. *Software: Practice & Experience*, 2004.

[4] Menczer F,Belew R. Adaptive Retrieval Agents:Internalizing Local Context and Scaling up to the Web. *Machine Learning* , 2000.

[5] Pant G, Menczer F. Topical Crawling for Business Intelligence. Proc 7th European Conference on Research and Advanced Technology for Digital Libraries, 2003.

[6] K. Stamatakis, V. Karkaletsis, G. Paliouras, J. Horlock, et al. Domain-specific Web site identification: the CROSSMARC focused Web crawler. *Proceedings of the 2nd International Workshop on Web Document Analysis*, 2003.

[7] Filippo Menczer, Gautam Pant, Padmini Srinivasan, Topical web crawlers: Evaluating adaptive algorithms. *ACM Transactions on Internet Technology*, 2004.

[8] Dennis Fetterly, Mark Manasse, Marc Najork, Janet L. Wiener: A large-scale study of the evolution of Web pages. *Software: Practice & Experience*, 2004.

[9] Judit Bar-Ilan: Search Engine Ability to Cope With the Changing Web. *Web Dynamics*, 2004.

[10] Junghoo Cho, Hector Garcia-Molina: The Evolution of the Web and Implications for an Incremental Crawler. *Very Large Data Base Endowment Inc.*, 2000.

[11] Fred Douglis, Anja Feldmann, Balachander Krishnamurthy, Jeffrey C. Mogul: Rate of Change and other Metrics: a Live Study of the World Wide Web. *USENIX Symposium on Internet Technologies and Systems*, 1997.

[12] Dennis Fetterly, Mark Manasse, Marc Najork: On The Evolution of Clusters of Near-Duplicate Web Pages. J. *Journal of Web Engineering*, 2004.

[13] Brian E. Brewington, George Cybenko: How dynamic is the Web? *Computer Networks*, 2000.

[14] Brian E. Brewington, George Cybenko: Keeping Up with the Changing Web. *IEEE Computer*, 2000.

[15] Luis Francisco-Revilla, Frank M. Shipman III, Richard Furuta, Unmil Karadkar, Avital Arora: Perception of content, structure, and presentation changes in Web-based hypertext. *Hypertext* , 2001.

[16] Tao Meng II, Hongfei Yan, Jimin Wang, Xiaoming Li: The Evolution of Link-Attributes for Pages and Its Implications on Web Crawling. *Web Intelligence*, 2004.

[17] Dennis Fetterly, Mark Manasse, Marc Najork, Janet L. Wiener: A large-scale study of the evolution of web pages. *Proceedings of WWW –12th International World Wide Web Conference*, 2003.

[18] Judit Bar-Ilan, Bluma C. Peritz: Evolution, continuity, and disappearance of documents on a specific topic on the Web: A longitudinal study of informetrics. *Journal of the American Society for Information Science and Technology*, 2004.

[19] XM Li. An Estimation of the Quantity of Web Pages Ever in China. *Journal of Peking University (Science and Technology)*, 2003. (in Chinese with English abstract)

[20] Sandeep Pandey, Christopher Olston: User-centric Web crawling. *Proceedings of WWW –15th International World Wide Web Conference*, 2005.

[21] Ashutosh Dixit and A.K. Sharma. A Mathematical Model for Crawler Revisit Frequency. *Proceedings of IEEE 2nd International Advance Computing Conference*, 2010.

[22] Niraj Singhal, Ashutosh Dixit, and Dr. A. K. Sharma. Design of a Priority Based Frequency Regulated Incremental Crawler. *International Journal of Computer Applications*, 2010.

[23] Tao Meng, Hongfei Yan, Jimin Wang, Characterizing Temporal Locality in Changes of Web Documents. *Journal of the China Society for Scientific and Technical Information*, 2005.

Mengmeng Wang was born in 1987 in the city of Changchun. She graduated from college of computer science and technology in Jilin university in the year of 2011.She is currently a graduate student with the computer software and theory at Jilin University, China, working in the fields of social computing and data mining. Her research interest includes data mining, information retrieval and social network.

Xianglin Zuo is a student at Jilin University, China. His major is Computer Science and Technology. His research interest includes data structure, data mining and web crawling.

Ying Wang was born in the year of 1981. She took her master degree(Computer application technology) from Jilin University in the year of 2007. She received her Ph.D(Computer application technology) from Jilin University in the year 2010.Currently she is working as an instructor in the colleage of computer science and technology at Jilin University, China. She has many international publications in Journals and conferences. Her research interest includes data mining, information retrieval and social network. In the year of 2009, she won the second prize of Jilin province scientific and technological progress.

Wanli Zuo received his Ph.D(computer software and theory) from Jilin University in the year of 1985. Currently he is working as professor in the colleage of computer science and technology at Jilin University, China. His research interest includes database theory, machine learning, data mining, web mining and internet search engine. He has guided several Ph.D students. He has published 5 materials and works and has more than 110 publications in international Journals and conferences.

Prof.Wanli is China's computer society system software professional committee, teaching commission committee member of college of computer science and technology in Jilin university, the first batch of Jilin province top-notch innovative talents, Jilin university teacher's morality pacesetter, Jilin university teaching demonstration teachers title. He has outstanding contributions of the young and middle-aged professional and technical personnel of Jilin province. Prof.Wanli has received several awards such as” The national teaching achievement prize”,” the baosteel education fund excellent teachers” and five a provincial-level award.