Web Scraping for Estimating new Record from Source Site

Warna Agung Cahyono
Department of Electrical Engineering
University of Brawijaya
Malang, East Java, Indonesia

Wijono
Department of Electrical Engineering
University of Brawijaya
Malang, East Java, Indonesia

Herman Tolle
Department of Information System
University of Brawijaya
Malang, East Java, Indonesia

Abstract: Study in the Competitive field of Intelligent, and studies in the field of Web Scraping, have a symbiotic relationship mutualism. In the information age today, the website serves as a main source. The research focus is on how to get data from websites and how to slow down the intensity of the download. The problem that arises is the website sources are autonomous so that vulnerable changes the structure of the content at any time. The next problem is the system intrusion detection snort installed on the server to detect bot crawler. So the researchers propose the use of the methods of Mining Data Records and the method of Exponential Smoothing so that adaptive to changes in the structure of the content and do a browse or fetch automatically follow the pattern of the occurrences of the news. The results of the tests, with the threshold 0.3 for MDR and similarity threshold score 0.65 for STM, using recall and precision values produce f-measure average 92.6%. While the results of the tests of the exponential estimation smoothing using α = 0.5 produces MAE 18.2 datarecord duplicate. It slowed down to 3.6 datarecord from 21.8 datarecord results schedule download/fetch fix in an average time of occurrence news.

Keywords: Web Scrapping, Exponential Smoothing, MDR, STM

1. INTRODUCTION

The process for collecting any information about competitors from public spaces on the internet with actionable form short or long term strategy is the field of Competitive Intelligent [1]. Companies protect themselves from slander and erroneous rumor of news online. On the other hand the news sites, forums, social media, blogs, etc. can serve as a trigger for customer to customer communication or the customer to the company. And consequently enlarge the area of the market share[2]. Information retrieval techniques from this public site is an early stage that must exist in the system search engine as well as other web mining[3].

Any public sites are autonomous, so that changes to the structure of the HTML tags on the server can occur at any time without being noticed by users. So a method, which may classify datarecord [4] without duplicating by visual senses such as eyes but adaptive to changes in the structure of HTML is required. Method i-robot crawlers [5] picking data record and perform construction of sitemap source site, then do the traverse to the news pages [6]. However, i-robot does not detect more than two of the same news posted on different time. So it needed an additional method for identifying ID post.

From the view point of the server's news sources, if the server detects the intensity of the fetch of the client on the page the same source simultaneously, then it can be categorized under DDoS attack by the system IDS [7]. From the view point of two-way client-server and the problem of duplication of record news on adaptive crawling, the author proposes a method for predicting new posts a news source. so that the intensity of the download page is not often. Then it is not detected as the machine/robot crawlers by the server.

There are several methods of crawling the news with predicting a new news post. First, the novel architecture and algorithm for web page change detection [9], time complexity is small but does not detect changes in the structure of the tags so it is not adaptive and prone to duplication of records.

Second, method Carbon Dating The Web: Estimating the Age of Web Resources, a method where a change of news gathered from several approaches through the header response, RSS XML, backlinks and google's index, then the delta t is calculated from the change in its content [8]. Unfortunately on the server side scripting, date create is dynamic.

Third, A Novel Combine Forecasting Method for Predicting News Update Time, This method uses a combination of Exponential Smoothing and Naive Bayes. The root of the Naive Bayes is exponential smoothing level. The first leaf is a type of news category. Then the second leaf is the number of occurrences of a data set of training [10]. But this method only gets data from RSS.

Fourth, the approach of Mining Data Record [4], a method based on similarity of news segment visually and unsupervised adaptive to change the structure of the HTML source of the news. Although this is not a method to estimate the changes of new content. But the method chosen by the author by combining the exponential smoothing methods. And helpful to know changes some new news in one page. So the datatime series can be taken specifically from each group different news categories in a single page.

Testing on the efficiency of the system are done with the Mean Absolute Error. Testing data on the accuracy of news who successfully learned are done using Recall and Precision to assess the suitability of the record that is being drawn with the original records on the source website[11].

2. THE PROPOSED METHOD

Schedule the fetch/download, which uses exponential smoothing (ES)[12], used to slow the intensity download/fetch followed the appearance of new news site estimate source of its purpose. In Figure 1 below, the system is divided into 4 sub methods namely preparations, MDR, temporary classification, archiving data records, and estimator using ES method.

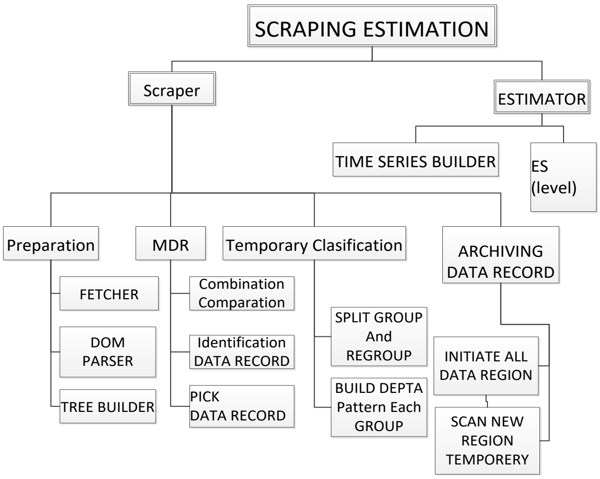

Figure 1. Taxonomy Method Scraping with Estimated New News

2.1 Preparing

Agent response, generated by the download/fetch in the HTML form, in the format parsed into the Document Object Model (DOM) [14]. MDR method uses the DOM data structure as a model for the purposes of comparisons between similar subtree using the edit distance levenstein[16]. To speed up the process of comparison/combination on the MDR method, all nodes except the type element must be removed from the DOM-tree, and each node is modified to contain an array of the results of the combination/comparison for each of the k nodes. In the method of MDR, k is the number of combinations tagtree in a generalize node. In this research the maximum number of combinations is 10 sub tag tree.

2.2 Mining Data Record

Every edit distance results from combinations and comparisons begin n = 1 to k are stored into the data array in each tagnode. At the stage of identification a dataregion, every tagnode checked if the value generated from a combination of levenshtein comparisons under the threshold.

On stage pick datarecord, every child from dataregion sliced into array object datarecord. Each datarecord is formed from one or several combined tagnode also has sub tagtree. At the time of the formation of datarecord also note the download time is saved in each object datarecord.

2.3 Temporary Classification

The entire datarecord in every new dataregion, at this stage, will be divided if between datarecord has similarities Simple Tree Matching (STM) is less than the value of the threshold score[18]. ReGroup algorithm can be seen in Figure 3. The value of the similarity score obtained from a similar number of nodes of the STM results on two tagtree (every tagtree comes from datarecord to compare). To normalize the amount of score values, tagnode in two tagtree calculated then total tagnode divided in half, resulting in an average number of tagnode. A normalized score is the number of matching results STM tagnode divided average number of tagnode. In this research, we uses the value score 0.65.

Although in a dataregion, if there are two datarecord level similarities under the score threshold then it should be split into two different group. And each group has a pattern in the form of tagtree. Tagtree pattern is obtained by aligning [13] the entire tagtree of every member from datarecord group.

2.4 Temporary Classification

After a group has members all datarecord tagtree similar to each other. So any new datarecord in each group is stored into a dataregion which has existed in the form of local files. Before it is stored into a local file datarecord then must be checked against all the href attribute of each anchor tag <a>in datarecord. Researchers use the mongo database to store the array into a single document hrefs. Every document in the database save file location and array contains attribute datarecord hrefs from datarecord. If the combination of hrefs datarecord recently matched with one of the document database, then the datarecord need not be saved. But note download time remains stored in the database for the purposes of the evaluation of the new datarecord post.

Then a combination of hrefs, in each of the daterecord, will be an ID (identity) for every datarecord. Thus, the duplication will not occur when looping download/fetch on the same attribute datarecord hrefs at different time.

The pattern comes from the dataregion which has existed in the local data and the new pattern of a similar group will be carried out using the method of partial tree aligning depta. So the new patterns, aligning results, in the form of tagtree, stored into the pattern in each dataregion.

If there are two equal record posted on different time, all hrefs in datarecord looping will occur. If the entire hrefs on two record match then it can be considered as duplicate records, so it does not need to be saved as archive files. If the entire hrefs don't match the database queries, then save the new datarecord into local files, note also the url, number dataregion, and download time.

2.5 Estimator

This estimation intends to calculate how many new news datarecord that appear during the span of delta T. Thus, emergence of new news time can be estimated. With more, the intensity of the process of repeating browsing on the same content can be slowed.

Forecasting method used is the method of exponential smooting (ES) with parameter α [12]. Input comes from data download time per record. First, calculate how long the average time it takes the appearance of a new datarecord from certain dataregion from one site news source. P is the length of the dataregion. P is calculated from the sum of all datarecord every dataregion which can be displayed on a webpage. Delta T for dataseri derived from the average time the emergence of all datarecord as many as P. for every interval series, count how many datarecord in each interval of time. The format of the input data in the form of a series of new record how many downloaded on any delta t. Then dataseries modelled in the exponential smooting. The SES model used is a single/level [15] :

$$S_t = \alpha X_t + (1-\alpha)S_{t-1} \quad (1)$$

$$\hat{X}_t(m) = S_t \quad (2)$$

3. ESTIMATION OF DATA EXTRACTION

Referring to figure 1, this method is generally broken down into two major parts, namely web data capture and record appearance estimator news. Web data retrieval starting from fetcher, MDR, a temporary classification until datarecord archiving. While the estimated new news starting from the initiation of length P datarecord and average download time (ΔT) every dataregion, then formation of dataseries, then estimate the amount of n records that appear in the next interval. If ΔT is the average download time is full of P one data region. Then, the difference of time Δti is posting the datarecord i subtracted to the time posting datarecord i-1. A provision on the interval to the i as follows:

$$\Delta ti = \frac{T}{n_i} \qquad (3)$$

$$next\ fetch = lastfetch + \Delta ti * (P - 2) \qquad (4)$$

nextfetch is the estimated time of next fetching, which type of data is datetime. While lastfetch is the latest time fetch/download. i is interval, which is the time t exist in there. P is the length (maximum amount) of datarecord which is able to display at web page at certain dataregion.

3.1 Data Used

Sampling data web is the latest news from datarecord dataregion which has the pattern combination of generalized-node 1n [4]. Estimator is the sample data dataseries formed from records stored per dataregion from each site. With the initiation of shown in table 1 below.

Table 1. Characteristics of latest news datarecord at 27 Juni 2018 12:48 until 28 Juni 2018 04:57 UTC

URL	NoRegion/ Length P/ amount Dataregion	average ΔT (minute)
http://www.cnnindonesia.com/	1/6/6	52.6
http://www.tribunnews.com/	6/49/11	84.2
http://www.metrotvnews.com/?	14/26/21	92.3
http://www.merdeka.com	3/40/6	170.6
http://www.tempo.co	5/7/6	23.7
http://www.kompasiana.com	4/95/9	568
http://www.harianjogja.com	5/20/5	152.2
http://www.dream.co.id	6/7/7	565.5
http://www.antaranews.com	4/33/11	98.98
http://www.detik.com	5/38/6	62.7
http://www.rmol.co	4/39/8	244.2
http://www.inilah.com	8/28/14	185.4

3.2 Data Flow System Diagram

Plot this data describes the workings of the system, starting from setting URL parameter, T, scoreT, and α. URL parameters used as target fetch web scraping system. T is the levenshtein edit distance threshold on MDR. While scoreT is a normalized similarity score results of STM[18]. The parameter scoreT is used on a temporary classification and archiving, for comparisons of datarecord in the form of tagtree.

In Figure 4 below, the parameter α is the parameter for the level (St) in the exponential smoothing. Estimation of the next Download is shown in equation (3) and (4). Whereas temporary classification and archiving datarecord implemented in dataRecord archiving algorithms, SaveCheckDuplicate, and Regroup in Figure 2, 4 and 5. As shown below:

```
SaveCheckDuplicate(NoRg, DRCi, DRGLs)

FileLocation=queryDB(DRCi.hrefs)
if FileLocation is not exist
 NewFileLocation=SaveScrap(DRCi,
 DRCi.hrefs)
 SaveDB(NoRg, NewFileLocation, DRGLs)
End If
```

Figure. 2 Algorithm : Duplication check before hrefs are saved.

In Figure 2, NoRg is index number from the dataregion. Each dataregion which have been at the local file is always given the ID number in the region. QueryDB is query on mongo database to check whether the combination of hrefs in datarecord has existed with the return value in the form of file storage location datarecord. DRCi datarecord is no. i have just extracted. DRGLs is all the dataregion which has been in the local. SimpanDB is a function/method stores the address of the file locations on the database, on the DRCi dataregion number to NoRg from DataRegion DRGLs.

```
ReGroup(DRGi)

DRG2s=init collection dataregion
while(DRGi ≠ 0)
 DRC=pop up datarecord from DRGi
 for(i=1; i <= DRG2s.size; i++)
  if STM(DRG2s[i].pattern,DRC > =
  scoreT) then exit loop for;
  end if
 end for
 $T_s$=DRC
 if i <= DRG2s.size then
  $T_s$=PartialTreeAlign(DRG2s[i].pola, DRC)
  If DRC unable aligned then
   i= DRG2s.size+1
  End if
 End if
 DRG2s[i].pattern=$T_s$
 DRG2s[i].add(DRC)
End while
return DRG2s
```

Figure. 3 ReGroup Algorithm for temporary classification .

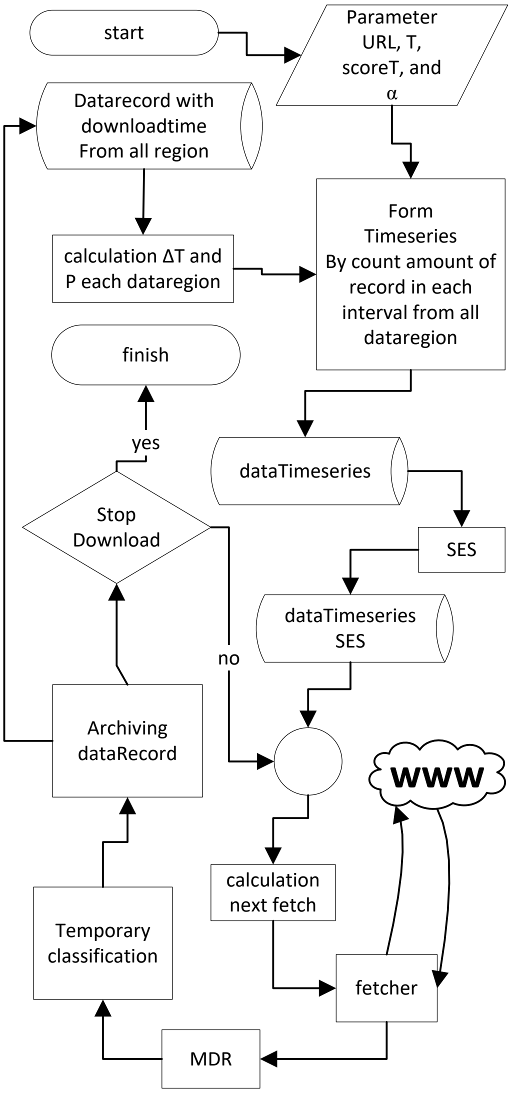
start
Parameter URL, T, scoreT, and α
Datarecord with downloadtime From all region
calculation ΔT and P each dataregion
Form Timeseries By count amount of record in each interval from all dataregion
finish
yes
Stop Download
dataTimeseries
SES
dataTimeseries SES
no
Archiving dataRecord
www
calculation next fetch
Temporary classification
fetcher
MDR

Figure. 4 Diagram of a system of Scraping with the fetch being estimated.

```
ArchivingDataRecord(DRGs,DRGLs)

1.  while(DRGs ≠ 0)
2.     DRGi=pop up dataregion from DRGs
3.     DRG2s=ReGroup(DRGi)
4.     while(DRG2s ≠ 0)
5.       DRG2i=popup dataregion from DRG2s
6.       polaDRGi=take from DRG2i.patern
7.       for(i=1;i<=DRGLs.size;i++)
8.          if STM(polaDRGi,DRGLs[i].patern) > =
       scoreT
9.               then    NoRg=i
10.      End For
11.      if NoRg not yet set then NoRg= DRGLs.size+1
12.        while(DRG2i  ≠ 0)
13.          DRCi=popup datarecord from DRG2i
14.          SaveCheckDuplicate(NoRg, DRCi, DRLGs)
15.        End while
16.    End while
17. End while
```

Fig. 5 Algorithm of Data Record Archiving.

In Figure 3, DRGi is a temporary Dataregion with the number i. DRGi generated by MDR after fetch/download just happened. On line 1 Figure 3, DRG2s is an empty dataregion. DRC on line 3 is temporary datarecord originating from dataregion DRGi. Line 4 will be skipped if DRG2s still empty. On line 5, compare the pattern of the dataregion tagtree belongs to each DRG2s with datarecord tagtree DRC.

In figure 5, polaDRGi at line 6 is tagtree, tagtree pattern which is belong to each the dataregion i's DRG2. tagtree Patterns on everytime dataregion used for pattern matching to determine the suitability of the dataregion on certain dataregion. STM on line 8 is the SimpleTreeMatching algorithm[18].

3.3 Accuracy Experiment

This research uses 12 URL test. Testing at this stage of this web data capture must be done before the stage of forecasting. This is to tell how valid the data used in web data retrieval system, including this estimation. Testing is done using web data recall and precision[11]. In Figure 4, there is a section of the MDR, temporary classification, archiving datarecord which respectively using threshold editdistance not more than 0.3 for levenshtein and threshold score similarity ternormalisasi not less than 0.65 for STM. While the selected dataregion is a dataregion which latest news is shown.

Table 2. Testing : TP(True Positive), Recall, precision, dan f-measure for 12 news sites at 2 August 2017 12:10 until 7 Agustus 2017 05:16 UTC

Sites	TP	Recall (%)	f-measure
www.inilah.com	28	87.5	93.3%
www.cnnindonesia.com	18	85.7	92.3%
www.tribunnews.com	48	92.3	96%
www.metrotvnews.com	26	96.3	98.1%
www.merdeka.com	29	96.7	98.3%
www.dream.co.id	13	76,5	86.7%
www.tempo.co	31	100	100%
www.kompasiana.com	100	100	100%
www.antaranews.com	50	98	99%
www.harianjogja.com	17	95.2	97.6%
www.detik.com	36	94.7	97.3%
www.rmol.co	29	87.9	93.5%
Average		92.6	98.8

The precision column is not shown in table 2 due to all urls have value 100%. 12 sites in table 2 have a value of 100% precision so there are no data received incorrect/false. The average recall values was 92.6%, meaning that there is a still unread datarecord in small amounts, usually the last datarecord in the dataregion. But the data still can be used for datatime series because no data was wrong. In addition, the latest data on the web view will shift. So in the end, will still be readable datarecord by scraping system.

The next testing phase is to test the efficiency. The second test is done using MAE for measuring efficiency of download traffic. Efficiency is done by comparing the value of MAE download fix and value estimation-based download MAE. Download the fix, with time, done in the time span ΔT – (P-2) * Δti, Δti value obtained from equation (3). While the range of download time that ter-estimation is calculated based on the difference in time nextfetch subtracted by lastftech in units of minutes. The estimation is done with a value of α 0.5 can be seen in the following table.

Table 3. The average test results of MAE on schedule scraping, miss and duplication datarecord refers to the fix average history.

Sites	MAE Duplication
www.inilah.com	11.96
www.cnnindonesia.com	5.88
www.tribunnews.com	37.1
www.metrotvnews.com	12.7
www.merdeka.com	14.4
www.dream.co.id	6.9
www.tempo.co	6.8
www.kompasiana.com	79.2
www.antaranews.com	16.5
www.harianjogja.com	9.9
www.detik.com	37.1
www.rmol.co	23.8
Average	21.8

At the second trial was performed on the parameter α 0.5 which means balanced between the influence of the actual data and the average history results estimation. Testing with parameters α = 0.5 on 12 URL address is performed on every dataregion, since each URL address has more than one data region. The results of testing with α = 0.5 can be seen in the following table:

Table 4. The average test results of MAE on schedule the scraping miss and duplication datarecord refers to ES estimation.

Sites	MAE Duplication
www.inilah.com	9.5
www.cnnindonesia.com	5.99
www.tribunnews.com	12.1
www.metrotvnews.com	7.8
www.merdeka.com	7.3
www.dream.co.id	6.96
www.tempo.co	6.5
www.kompasiana.com	89.1
www.antaranews.com	10.95
www.harianjogja.com	8.3
www.detik.com	37.3
www.rmol.co	16.5
Average	18.2

Test estimation based on fix average time dataregion and test-based estimation of exponential smooting with parameters α = 0.5 can lower MAE of 3.6.

4. CONCLUSION

In this paper, an approach web scraping with estimated download time is used to avoid the intensity of downloads are too often. Test data shows that there is no negative data that is read by the system. Test of scraping has done using levenshtein edit distance threshold parameters of 0.3 and similarity threshold parameter STM of 0.65, which shows the f-measure 98.8% with 100% precision. So none datarecord which will not be stored in the database. The next test on the efficiency of the system with the SES 0.5 alpha parameters required. Test results indicate MAE decrease download time of 3.6. It showed a reduced number of duplicate datarecord who found a result too often download/fetch. In other words schedule fetch became slower, so the vacuum of download time can reduce the network traffic load. And reduce the impact of the system IDS from the server site news.

5. FUTURE WORK

Minus 2 on formula 4 is used to match the records of the database. then the next job is how to get rid of minus 2. So we need a method to recognize the paging url of the datarecord in the dataregion.

6. REFERENCES

[1] W. Y. C. Thompson S.H. Teo, "Accessing The Impact Of Using The Internet For Competitive Intelligence," Information And Management, 2001.

[2] D. J. F. W. Glynn Mangold, "Social media: The new hybrid element of the promotion mix," Business Horizons, pp. 257-365, 2009.

[3] S. P. Yugandhara Patil, "Review of Web Crawlers with Specification and Working," IJARCCE, pp. 220-223, 2016.

[4] R. G. Y. Z. Bing Liu, "Mining Data Records in Web Pages," SIGKDD, 2003.

[5] J.-M. Y. W. L. Y. W. L. Z. Rui Cai, "iRobot: An Intelligent Crawler for Web Forums," WWW 2008, 2008.

[6] B. S. P. P. Namrata H.S Bamrah, "Web Forum Crawling Techniques," International Journal of Computer Applications, vol. 8, pp. 36-41, 2014.

[7] A. A. N. V. Dusan Stevanovic, "Feature evaluation for web crawler detection with data mining techniques," Expert Systems with Applications, 2012.

[8] N. Salah Eldeen, "Carbon Dating The Web: Estimating the Age of Web Resources," International World Wide Web Conference Committee (IW3C2), pp. 1075-1082, 2013.

[9] D. K. S. Deepak Kumar Ganeshiya, "A novel architecture and algorithm for web page change detection," International Advance Computing Conference (IACC), 2013.

[10] X. Z. W. Z. Y. W. Memeng Wang, "A Novel Combine Forecasting Method for Predicting News Update Time," Fourth International Symposium on Information Science and Engineering, pp. 227-231 , 2012.

[11] D. M. W. Powers, "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness &

Correlation," Journal of Machine Learning Technologies, p. 37–63, 2011.

[12] T. M. J. A. Cooray, Applied Time Series: Analysis and Forecasting, Oxford: Alpha Science International Limited, 2008.

[13] Y. Zhai and B. Liu, "Web data extraction based on partial tree alignment," Chiba, Japan, 2005.

[14] Philippe Le Hégaret; Ray Whitmer; Lauren Wood, "Document Object Model (DOM)," 19 January 2005. [Online]. Available: https://www.w3.org/DOM.

[15] Gardner and E. S., "Exponential smoothing: The state of the art—Part II," ELSEVIER, International Journal of Forecasting, p. 637–666, 2006.

[16] V. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," Soviet Physics Doklady, p. 707–710, 1966.

[17] A. Papana, "Short-Term Time Series Prediction For A logistics Outsourcing Company," in Outsourcing Management for Supply Chain Operations and Logistics Service, Thessaloniki, Bussines Science Reference, 2013, pp. 150-160.

[18] W. Yang, "Identifying syntactic differences between two programs," Software—Practice & Experience, pp. 739 - 755, 1991.