International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 01 | Jan-2018 www.irjet.net p-ISSN: 2395-0072

Semantically-Interlinked Based on Rich Site Summary Bank for Sites of Indonesia Online News

Andrea Stevens Karnyoto[1], Parea Russan Rangan[2], Abedneigo Carter Rambulangi[3], Indo Intan[4]

[1]Teknik Informatika of Universitas Kristen Indonesia. Tana Toraja, 91817 South Sulawesi, Indonesia,
[2]Teknik Sipil of Universitas Kristen Indonesia. Tana Toraja, 91817 South Sulawesi, Indonesia,
[3]Manajemen of Universitas Kristen Indonesia. Tana Toraja, 91817 South Sulawesi, Indonesia,
[4]Teknik Informatika of STMIK Dipanegara. Makassar, 90000 South Sulawesi, Indonesia,

---

Abstract - *Currently, Indonesia is against hoaxes. To develop an information system that can reduce the hoax spread is needed. We are improving technological aspects to create the information system that can calculate percentages of connectivity of semantically-interlinked by using keyword similarity method in several online media. As we know, the technical elements are needed to support people make more accessible to categorize the news and to get a relation. The Indonesia Government has done several ways, such as doing socialization due to social media, blocking several hoaxes at social media, and eliminated hoaxes sites. Based on this research, by using this keyword similarity method, it was found the relation between an online news to another with 5 keywords have the connectivity of semantically-interlinked shows the lowest is 0.823%, and highest is 1.253%. Lack of linkages between the media every online media happens because they were using the different grammar even though same topic.*

Key Words: Google Trends, Hoax, Indonesia Media Online, Keywords Generator Algorithm, News, Rich Site Summary, Semantically-Interlinked, Words Connectivity.

1. INTRODUCTION

Hoaxes are a threat in the internet, Indonesia is also against it. We have conducted a preliminary observation; we found that hoaxes were targeting essential objects such as a person, incident, and even particular groups or organization. Hoax created by individuals and the groups of people. Hoax makers and hoax spreader use information technology in their action. To minimize the hoaxes activity, we need to develop an information system that can reduce it. The government has to exercise two aspects to decrease the hoaxes, that is the social aspect and technology aspects. Social aspect required to realization humanely and give understanding to people with persuasive way. Technological are needed to assist people to make more accessible to categorize the news and the relation between one and another news. The Indonesia Government has done several ways to socialize due to social media, blocking several creators, and eliminate hoaxes sites.

Alexa.com released a ranking of Indonesia trusted online news sites that have an excellent rating for regional Indonesia. Tribunnews.com is the 4th ranked, detik.com at the 5th, liputan6.com at the 6th, kompas.com at the 10th, and okezone.com at the 14th. It is becoming a reference that Indonesia's people have the penchant for reading news.

When internet becomes an integral part of modern daily life, indication of internet keyword search has become one important indicator for people to understand social development [2]. Even Google Trends improve effectively to predict unemployment rate [5]. Similarly to Twitter, every one hour Google releases a list of the top-20 trending search keywords (i.e., the so-called hot trends or hot queries), which we refer to as web trends [3]. Taking the popular word or phrase on Google Trend would assist vastly to develop the categorization system that will support to combat hoax. Google Trends provides links to popular daily keywords so software developers can easily access and retrieve keyword information. Google released at least 20 hot keywords each day. Alexa.com release a rating that shows Google is the most search engine widely used by Indonesian People.

Since its inception in mid-90s, the World Wide Web has become a major platform for presenting and distributing data in various domains. Much instrumentation and measurement data also are automatically processed and posted on the Web to reach more people of interest [4]. The internet becomes a complex thing so people create simple method to sharing news by collected news summary in a link or file, this method easier than read the entire contents of website. Disadvantage of this method is text limit of summary. Rich Site Summary (RSS) is a way to get briefed items on a website as the web site gets updated, and these are called feeds [1]. A feed is often a series of headlines and brief summaries of all the articles published on a web page. Instead of having to visit numerous web sites to get weather, sports, latest gossip, or latest political debate [1]. Many internet research using RSS because it is simple and full feature.

Our research is to develop a website that will connect the news from online media sites Indonesia, so people can quickly to get news interlinked and make conclusions from the news that has circulated. News from trusted online media can be a reference as material to know the story is

happening. The data should not only get from one online media site but should compare with the websites of another online news. This research is just developing technological aspect, and this paper is an explanation of news semantically-interlinked process and result. This research using RSS (Rich Site Summary) from several online media as data and keywords from Google Trends as interlink keys. The advantage RSS data is the convenience to get news that legal to be duplicated and processed by our site. Furthermore, RSS ability is to share and feed to many websites. Besides, every trusted online media in Indonesia have facilities such as RSS link or Atom link. RSS entities associated with this research are title, description, and summary. Moreover, the popular word from Google Trends as keys to getting connectivity of news from several online media is needed. First, Google is the most significant search engine and regularly update. Next, Google has facilities to give the developer accessibility to using data from Google trends.

2. RESEARCH METHOD

2.1 Main Process

Main process consist three processes. First, download, parsing and save RSS data to the database. Second, Download and analyze Google Trends keywords, and last, Semantically-Interlinked process of RSS. The result of all manner which will display to the web.

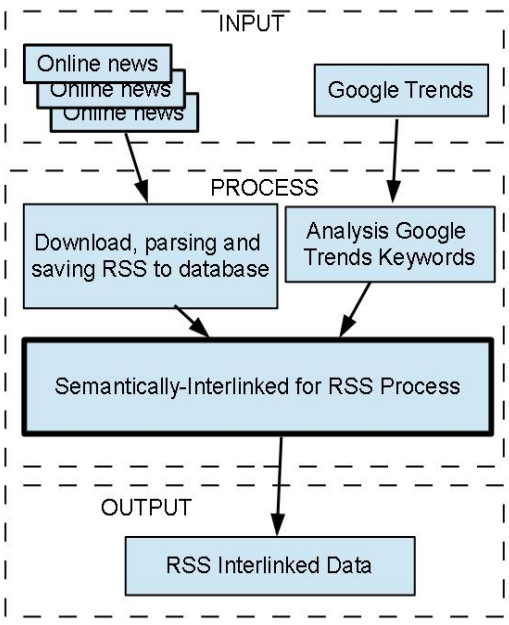

Fig-1 Main Process

This paper only explains the interlinked process in sub-process "Semantically-Interlinked for RSS Process" as shown fig 1. The process in this section is to make interlink between downloaded of RSS items which have resulting from section "Download, parsing and saving RSS to database" and sequence of keywords resulting from section "Analyze Google Trends keywords." The final result of this section is interlinked data RSS that using keywords from Google Trends as connectivity key and RSS yang has been downloaded several media online. This format makes information structured to facilitate people searching relevant information.

2.2 RSS Semantically-Interlinked

XML component in RSS has base entities such as Title, Description, guId, and Summary. This base entity will be used to make interlinked by connecting through popular keywords in the news.

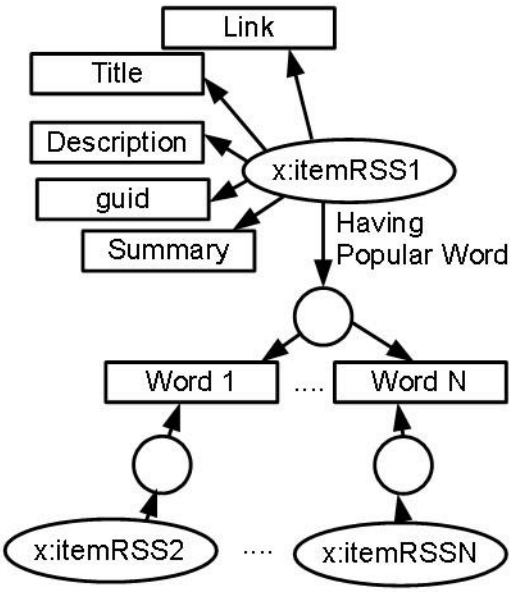

Fig-2 RSS Semantically-Interlinked Scheme

Fig 2 shown that linkage between one RSS items with another RSS items uses keywords. The system will connect the keyword to entities of RSS, i.e., title, description, and summary. With this design of the relationship as seen in fig. 2, it can be concluded that a keyword can be in many RSS items and an RSS can have many keywords. To search connectivity is by accessing the table that stores the relationship between keywords and RSS. The degree of interlinked will get by calculating the number of similarities.

2.3 Database Relationship

The data RSS that has downloaded will store in the table and processing of interlinked using SQL commands. The reasons we are choosing to use MySQL database engine because of the speed of access, easy to apply on many platforms and the ability to perform tasks on a vast scale data.

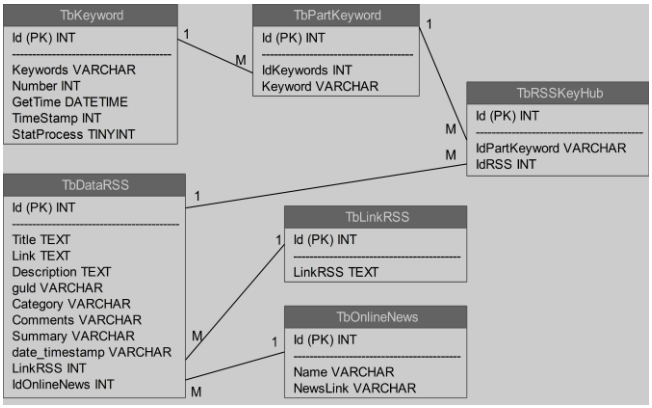

Fig-3 Database Relational

Fig 3 shows the tables relations for this system. With this relations is allow for each RSS record to have many popular keywords and popular keywords can be found in many RSS records. Each of the links stored on the TbLinkRSS generates RSS items; one RSS link contains many RSS items. In general, online media categorize each link to make it easier for developers to access their RSS link. Keywords that have more than one word will be split into words and stored in the table TbPartKeyword. Database relationships allow searching popular keywords in the TbDataRSS table by only using simple SQL commands. The degree of interlinked calculation results will facilitate count the percentage of RSS and search process by users.

2.4 Keywords Generator Algorithm

The popular keyword derived from Google trends are re-combined as the sequence of keywords, so the system has data to calculate semantically-interlinked in RSS table. Re-combined of sequence keywords using keywords from all topic that have entered into the TbPartKeyword table. This method to increase the possibility of keywords semantically-interlinked.

```
function GetKeywordCombination($arkwrd) {
 $rkwrd = array();
 $tkwrd = count($arkwrd);
 for ($i = 0; $i <= $tkwrd - 1; $i++) {
  for ($m = $i; $m <= $tkwrd - 1; $m++) {
     for ($j = $i; $j <= $tkwrd - 1; $j++) {
         $gkwrd = array();
         $gkwrd[] = $arkwrd[$i];
         for ($k = $m + 1; $k <= $j; $k++) {
            $gkwrd[] = $arkwrd[$k];
         }
         if (!in_array($gkwrd, $rkwrd)) {
            $rkwrd[] = $gkwrd;
         }
       }
     }
  }
  return $rkwrd;
}
```

The code shows an algorithm that generates a sequence of keyword combinations. The number of keywords determines the amount of keywords combinations; more keywords will resulting more amount of keywords combinations. The algorithm does not allow giving a repeated keyword in a sequence of keywords and ignores keyword position. Each sequence combinations of keywords entered in an array so the system can perform query process on the TbDataRSS table.

Table -1: Keyword and Combination.

Number of Keyword	Number of Combination
1	1
2	3
3	7
4	14
5	25
6	41
7	63
8	92
9	129
10	179

Table 1 shows the output of the algorithm to get the sequence number of combinations by referring to the amount of the keyword. It proves to 3 of the number of keywords will generate the sequence number of combinations is 7. For example, first, we take three keywords, i.e. ('jokowi', 'proyek', 'tol laut'). Finally, the result of a sequence number of combination that keywords is seven i.e. (('Jokowi'), ('Jokowi', Proyek'), ('Jokowi', 'Proyek', 'Tol Laut'), ('Jokowi', 'Tol Laut'), ('Proyek'), ('Proyek', 'Tol Laut'), ('Tol Laut')).

2.5. The Business Process Management Notation

The following diagram to describe how the system works. We use BPMN diagram to provide the overview.

a. Target RSS URLs

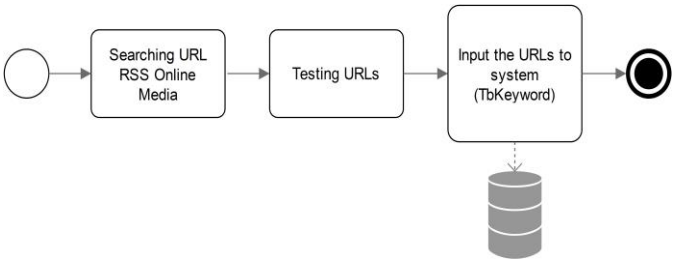

Fig-4 Input Target RSS URLs

In this process, first, the researchers get a list of URLs from each online media by searching the website of online media manually. Second, before the link stored in the database, we tested it whether could be found or not. Sometimes an RSS URL link cannot be found for some reason, i.e., 1) when the programmer of the online media edits the web page, they do not write the RSS Link URL. 2) The link is changed, for example the programmer write link using HTTP but in fact it is using HTTPS. Third, the list of the links inputted into the database by the system. Last, the transfer of RSS media online data into database system will execute the links when sub-process "download, parse and

save RSS data to the database" run. Each RSS item will occupy a record in the TbDataRSS table.

b. Create The Relation

The system will connect the keywords from Google Trends in TbPartKeyword table to RSS in TbDataRSS that containing those keywords; the easiest way to connect two tables is using a connecting table.

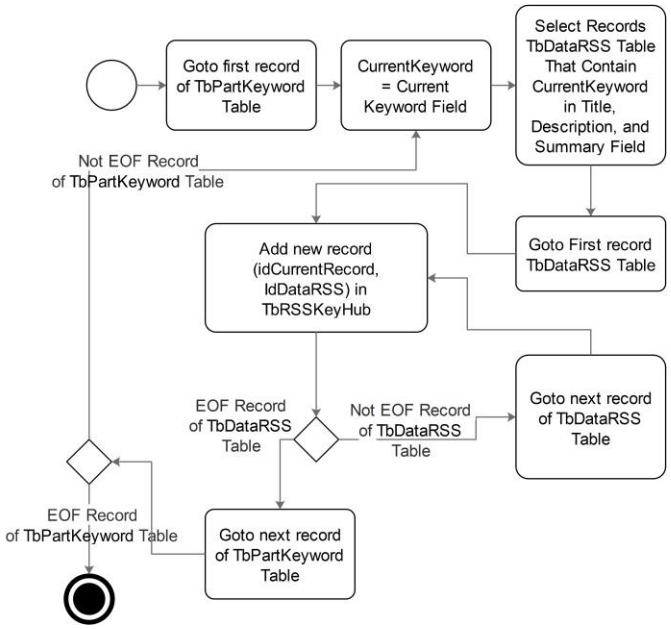

Fig-5 Create the Relation

Fig 5 shows that to connecting between TbPartKeyword table with TbDataRSS table is using Table TbRSSKeyHub. TbRSSKeyHub table will facilitate the search and interlink calculations because to calculate the interlinked percentage we only access one table. So in an RSS record there are two or more keywords it will be recorded on different records in TBRSSKeyHub Table.

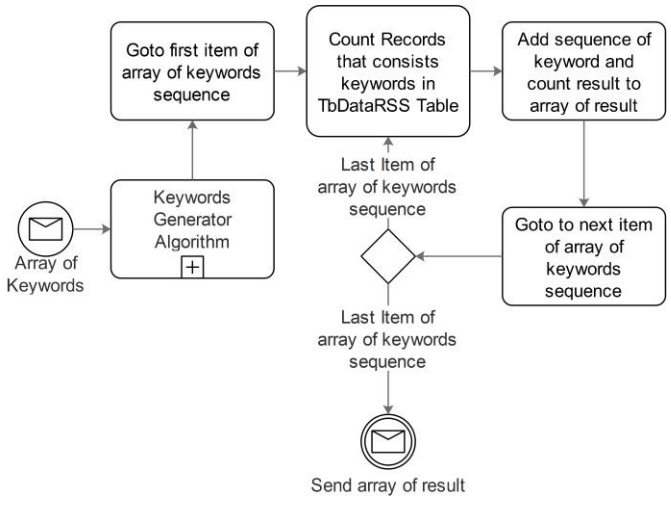

Fig-6 Search RSS That Consist Keyword

To facilitate the search with various sequences of keywords then the researcher makes a class with a groove as shown in fig 6. A sequence of keywords can consist 1 to infinity of Keyword. To avoid repeated keyword in a sequence of keywords use Keyword Generator Algorithm. The result of this class is the barrage of keywords and the number of records in TbDataRSS.

c. Semantically Percentages of Online Media

To compare between 2 online media used five same keywords. The following BPMN as shown in fig 7 is a process to obtain percentages of similarities.

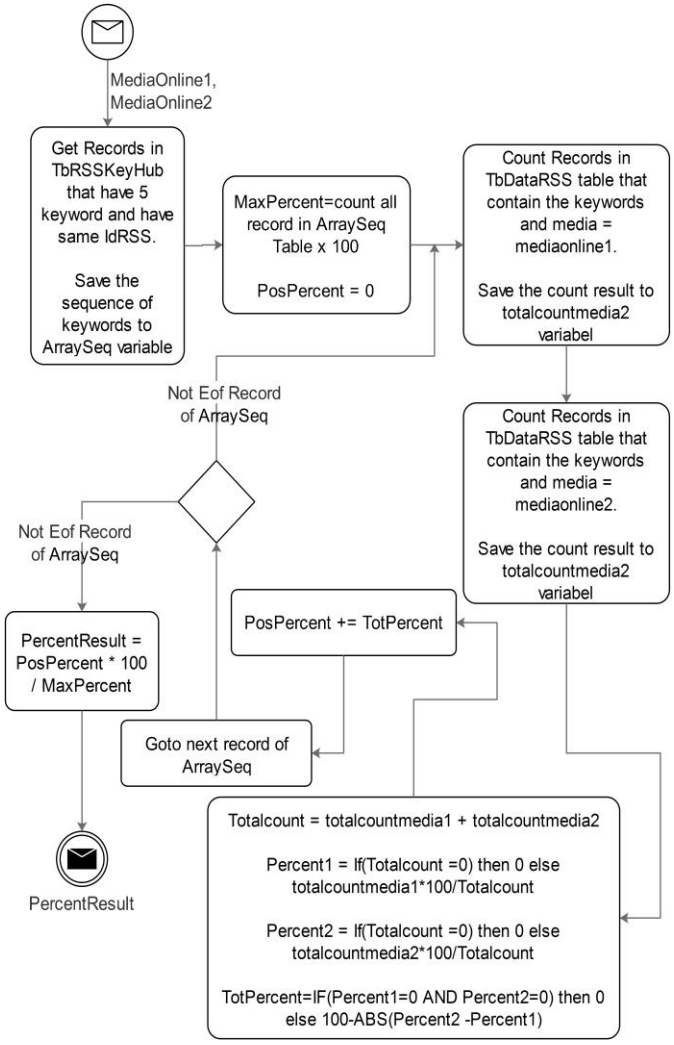

Fig-7 Process to Obtain Percentages of Similarities

Fig. 7 shows the process to get the percentage of interlinked from two online media. The input is the id of two online media. The next process is to get all RSS items from TbDataRSS that have 5 or more keywords. These keywords are objects for calculating the similarities of both online media. The result is the percentage of similarity of both online media.

3. RESULT

In a test of the system, we define the target RSS URL, each online media has different RSS versions, for some media use atoms.

In this case, Analyze Google Trends keywords result the number of keywords is 795.

3.1 Target RSS URLs

The popular online news is our research target; there is a few RSS link in a site. It was grouping to particular categories such as science, sports, technologies, business, automotive, finance, international, regional, property, etc. The process used to get the RSS link shown in figure 4.

Table-2: Total Links anda Total RSS News

No	Online News		
	Name	Total Link	Total RSS News
1	http://www.bbc.co.uk	6	469
2	http://www.detik.com	9	456
3	http://www.vivanews.com	16	412
4	http://www.antaranews.com	9	502
5	http://www.republika.co.id	33	565
6	http://www.okezone.com	13	370
7	http://www.suaramerdeka.com	10	400
	Total	127	4042

Seen in Table 2, we only took the material processing of RSS feeds from 7 online media, there are 127 RSS links, and the system successfully extracts 4042 of RSS items. The amount has already obtained from of various categories.

3.2 News and Relations

By using the process of search with various sequences of keywords as shown in Figure 6 and using the Keyword Generator Algorithm in section 2.4 to get a sequence of keywords. We take one record from TbDataRSS as a sample to know keywords and related news link. The examples are the current favorite news, i.e., "Kebut Program Tol Laut, Jokowi Resmikan 5 Pelabuhan di Maluku". From that news, we found five keywords in TbRSSKeyHub, so the system generates 25 sequences of keywords. There is not seen repeated keywords in the sequence of keywords because the system is designed to ignore the order of keywords of the news. Table 3 shows the result.

Table-3: Relationship Between News, Keyword And Number Of Relate News

No	Title News : Kebut Program Tol Laut, Jokowi Resmikan 5 Pelabuhan di Maluku	
	Kewords	Related News Links
1	Jokowi	74
2	Jokowi, Maluku	15
3	Jokowi, maluku, proyek	4
4	Jokowi, maluku, proyek, tol laut	3
5	Jokowi, maluku, proyek, tol laut, pelabuhan	2
6	Jokowi, proyek	45
7	Jokowi, proyek, tol laut	8
8	Jokowi, proyek, tol laut, pelabuhan	3
9	Jokowi, tol laut	15
10	Jokowi, tol laut, pelabuhan	9
11	Jokowi, pelabuhan	13
12	Maluku	6
13	Maluku, proyek	4
14	Maluku, proyek, tol laut	2
15	Maluku, proyek, tol laut, pelabuhan	2
16	Maluku, tol laut	2
17	Maluku, tol laut, pelabuhan	2
18	Maluku, pelabuhan	2
19	Proyek	6
20	Proyek, tol laut	15
21	Proyek, tol laut, pelabuhan	13
22	Proyek, pelabuhan	16
23	Tol laut	34
24	Tol laut, pelabuhan	14
25	Pelabuhan	23

3.3 Percentage of Semantically

The calculations to compare the similarities of the two online media uses the process as shown in Figure 7. Although Keyword Generator Algorithm on the system generates some 795 keywords but we use just five keywords as a source for determining the degree of similarity of news. The number of online news already displayed in table 2.

Table-4: Semantically Online News

No. Media	1	2	3	4	5	6	7
1	100%	1.041 %	1.109%	0.953%	0.972%	1.201%	1.212%
2		100%	1.095%	0.823%	0.993%	1.113%	1.001%
3			100%	1.253%	1.021%	0.989%	0.988%
4				100%	1.133%	1.011%	1.198%
5					100%	1.124%	1.107%
6						100%	1.095%
7							100%

Table 4 shows that semantically using five keywords has the highest level between (3) http://www.vivanews.com and (4) http://www.antaranews.com that is 1.253%, and percentage of the smallest similarity is between (2)http://www.detik.com and (4)http://www.antaranews.com that is 0.823%.

Next, we use five popular keywords to get the amount of news that exist in each online media. The sequence of five keywords that most frequently appear on our data shown in Table 5.

	Keywords	Online News							Total
		1	2	3	4	5	6	7	
1	Piala, Presiden, Jokowi, Sepakbola, Final	6	2	4	7	8	3	4	34
2	Aksi, 313, Pilkada, Damai, Jakarta	5	1	3	6	8	3	5	31
3	Organ, Tubuh, Anak, Penculikan, Jakarta	5	2	4	5	6	4	2	28
4	Arsenal, Sepakbola, manchester city, vs, liga inggris	4	3	2	1	4	3	2	19

Table 5 shows that sequence keywords are most numerous in the news is (Piala, Presiden, Jokowi, Sepakbola, Final) and the amount is 34 news.

4. CONCLUSION

Based on this research, that by using this keyword similarity method obtained a relation of one news to another online news. Connectivity of semantically-interlinked shows the lowest is 0.823% and highest is 1.253%. Lack of linkages between the news due to the use of different grammar and different words in every online media even though they are raising the same topic. Adding the number of search keywords will improve the quality of the news interlinked. Further research is needed to be semantically-interlinked no only use RSS because it has limited text length. Text length limitations brought the number of keywords is decreasing.

ACKNOWLEDGEMENT

This paper is the one of output of "Research Lecturer Beginners" funded by "Directorate of research and Community Service. Ministry of Research, Technology, and Higher Education. Indonesia".

REFERENCES

[1] D Veeraiah, Y V Ramanjaneyulu, D Yakobu, T Sahithi. (2016). "A Novel Approach for Extraction and Representation of Main Data from Web Pages to Android Application", IEEE International Conference On Recent Trends In Electronics Information Communication Technology. PP 1126-1130.

[2] Fan, M.H., Chen M.Y., Liao, E.C. (2014) A TAIEX Forecasting Model Based on Changes of Keyword Search Volume on Google Trends. 978-1-4799-5375-2/14 ©2014 IEEE.

[3] Giummolè, F., Orlando, S., Tolomei, G., Trending Topics on Twitter Improve the Prediction of Google Hot Queries., SocialCom/PASSAT/BigData/EconCom/BioMedCom 2013. pp 39-44.

[4] I Chen, S Hu, D Shih. (2009). "A Platform for Syndicating and Manipulating Instrumentation and Measurement Data on the Web". VECIMS 2009 - International Conference on Virtual Environments, Human-Computer Interfaces and Measurements Systems.

[5] Karamé, F., & Fondeur, Y. (2012). Can Google Data Help Predict French Youth Unemployment? Centre d'Études des Politiques Économiques (EPEE), Université d'Evry Val d'Essonne. pp 12-03.