Business Data Extraction from Social Networking

Asif Uddin Khan, Bikram Kesari Ratha
Dept of Computer Science
Utkal University, Vani Vihar
Bhubaneswar, India
asifkhan.iiit@gmail.com, b_ratha@hotmail.com

***Abstract*—Social networking sites like Facebook, twitter and Google++, contains valuable business information. We can extract information about the business like address, phone number, catalogs, special offer and business hour from these sites. In this paper we have proposed a method to extract valuable business information of a business organization from Facebook using crawler and IE techniques. We build a search engine using apache solr to extract useful information from Facebook**

Keywords—information extraction(IE); crawler; social networking; business data; jason;precission; recall;f1-mesure; solr.

I. INTRODUCTION

Web data extraction [1,2] is an important web application where useful and valuable information can be extracted from the web.Usefull information is attached in html web page or as a text itself in the page which can be extracted and further processed into a structured layout that can be used for various purposes. There are various applications of web data extraction system such as web text analysis, competitive intelligence, crawling, and bio-informatics and so on. Large amount of data is present in the web and also growing every day. We can collect these data by using web data extraction system efficiently. Again analysis of this huge amount of data for extracting scientific, social, economic and business information manually is a hard task. Business organizations can analyze their customer behavior and do market analysis and activity of their competitors [3, 4] by using web data extraction system.

Widely used Social networking sites such as Facebook, Google+, YouTube, LinkedIn, Instagram, Pinterest, Vine, Tumblr, and Twitter contain valuable information about business and shops and also incorporate new information and communication tools such as mobile connectivity, photo/video/sharing blogging [5] and valuable business data. We can extract information about the business like address, phone number, catalogs and business hour from these sites. In this paper we collected web pages using crawling and then applied information extraction techniques to extract business information from Facebook. One reason is that in popular social networking sites, many businesses will have their own fan page, at the top of the news published in their own stores, such as temporary rest of the class of goods or discount store, and if through efficient IE and IR words on Facebook, the user can bring considerable convenience. In addition to web content itself outside, json is also an important source of information provided by Facebook, the number provided above "Like" above the current, very useful when the page after statistical activities, can be used as basis for the recommended time. In this paper we have used an open source tool Apache SOLR [6] for designing our search engine.

The rest of this paper is organized as follows: Section II presents the problem statement, section-III describes related work, in section-IV we describe the proposed method, and section-V describes crawling strategy and method section-VI describes evaluation methods performance of crawling, in section- VII we describe the method of data extraction using our search engine, section-VIII describes evaluation of searching based on ranking, finally in section-IX we conclude the paper.

II. PROBLEM STATEMENT

Data extraction from the web is an important problem that has been studied widely using different scientific tools and applications. Many approaches have been proposed and designed to solve specific problems and operate in ad-hoc domains. Now a day's business organizations provide useful information in social sites like Facebook, YouTube or Flickr.Small organizations which do not have personal websites they rely on social networks to give their service information to the customers. We aim to extract business data like address, phone number, catalog, business hour and activities like special offer posted by the organization from Facebook id collected from internet. In this paper we extract phone number and business hour, address of the organization, longitude, latitude rating using user likes and we estimate the score of each searched information of the organizations by using efficient information extraction technique.

III. RELATED WORK

Social media Data Extraction is covered by a number of papers. Zhao et al. [7] presented a comprehensive knowledge extraction approach for social networks to guide latent dimensions analysis. Valkanas et al. [8] discuss the particularities of geocoding in online social networks and present a simple, lightweight, efficient approach for location extraction. In [9] Wang et al introduce a methodology to collect and analyze those personal data and by this for extracting social networks from the data. In [10] Kopka et al. constructs the social network from the process log in the given context finds communities in this network

978-1-4799-8579-1/16/$31.00 ©2016 IEEE

and communities were analyzed using knowledge of the business process and the environment in which the process operates. In [11] Borg et al. investigate the impact of using social network data extracted from an E-mail corpus to improve spam detection

IV. PROPOSED METHOD

We introduce the process and architecture of our system as in figure-1. We use Google search engine for querying keyword and add some search syntax to get the search results. We search all pages in Facebook. We use regular expression to fetch Facebook id and then insert Facebook id and other data to our MYSQL Database. Now our crawler program know facebook id so it can get Web Facebook page, json Strut graph and Facebook data by http request. In this work we choose the json Struct graph, Facebook data for extraction object as in figure-4.We developed efficient information extraction technique to extract phone number and Business hour, address, category, description and other information as in figure-2 and figure-3.

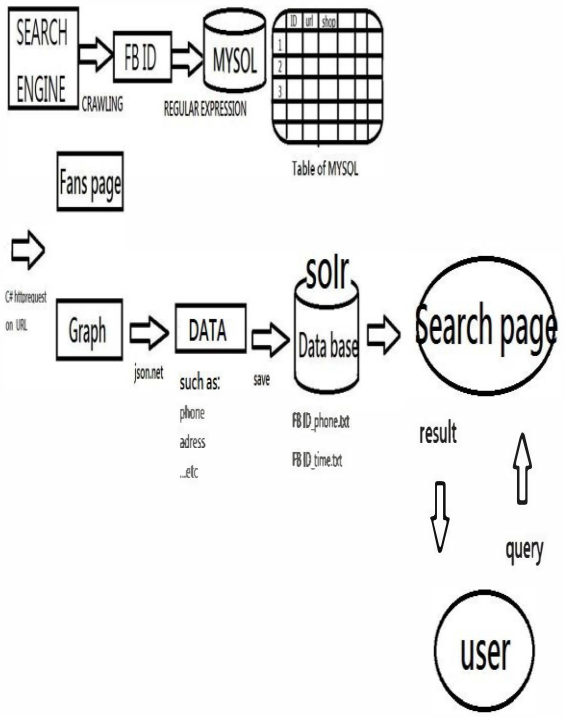

Figure-1: System Architecture

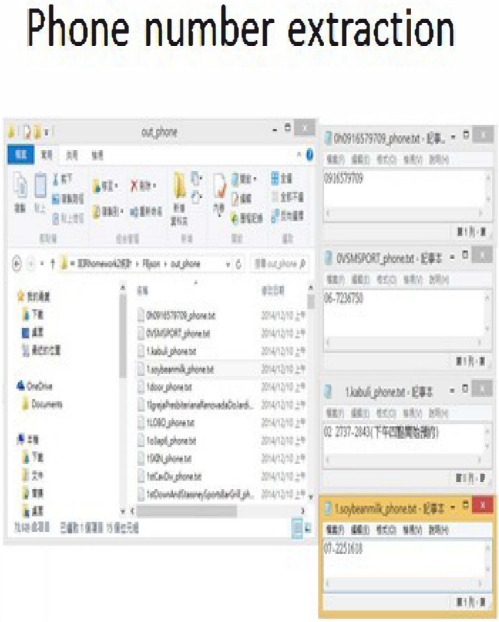

Figure-2: Phone number

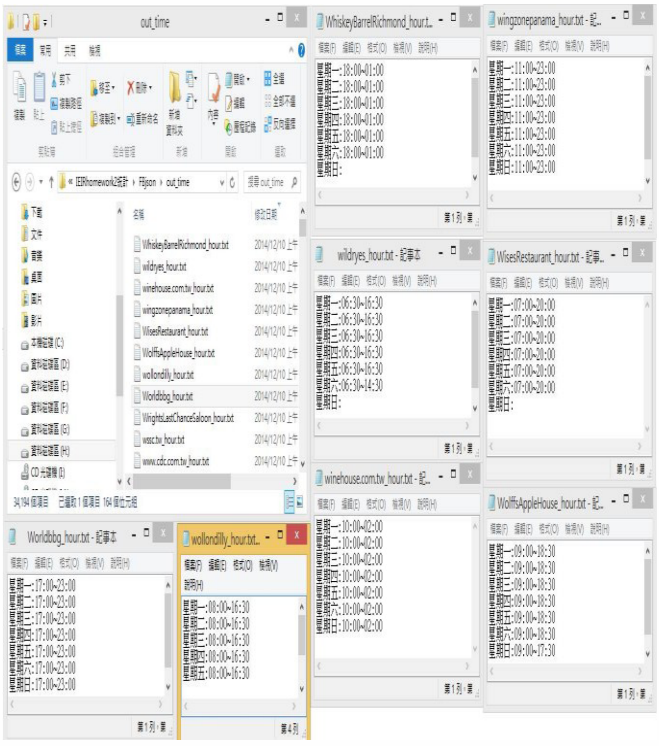

Figure-3: Business hour

Graph FB data

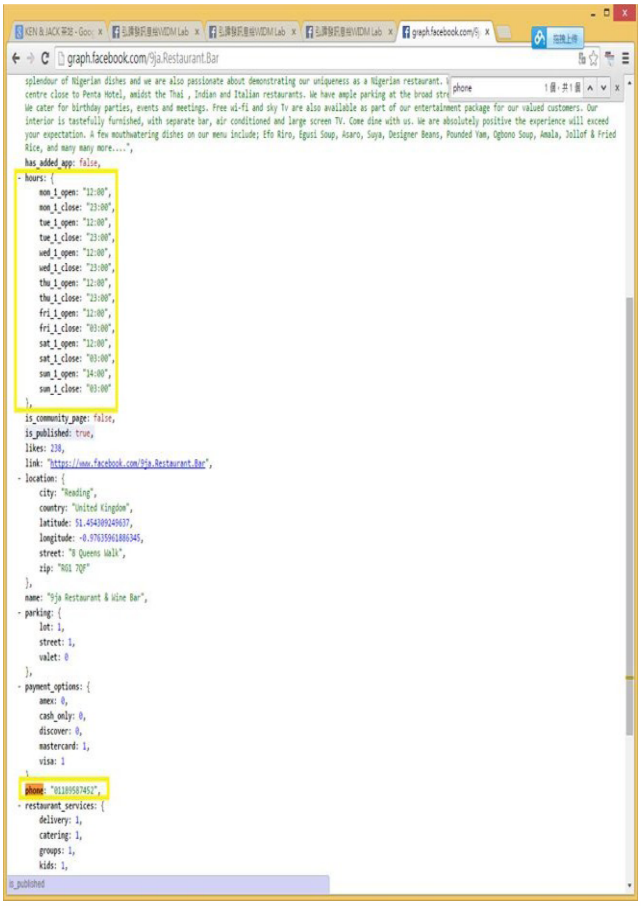

Figure-4: graph data

V. CRAWLING STRATEGY AND METHOD

We use crawler for collecting Facebook ids as in figure-2. Our crawler strategy designed for two blocks, one for obtaining business Facebook ID; and another one for web content and json.On obtaining FB Id we use keyword: "store name" site: facebook.com, in url:, in the Google search engine, query ID (the store name from the Solr) "rf =?".Id deposited is collected back into MYSQL database. The data collected are id, FB Id, Title, and other important information for future use. Where Id is the mean number MYSQL deposit and, after the use of mod it gives workstation id, conduct web and json collection as in figure-5. After the workstation gets Id use http client, the data sequentially and down, the data is stored only way is to retain a unified web information into the folder, it overwrite the old data with future updates. Json with each update date folder, Facebook Id folder for the file name is placed under, after the completion of the entire index back to the server through the establishment of xml ftp.

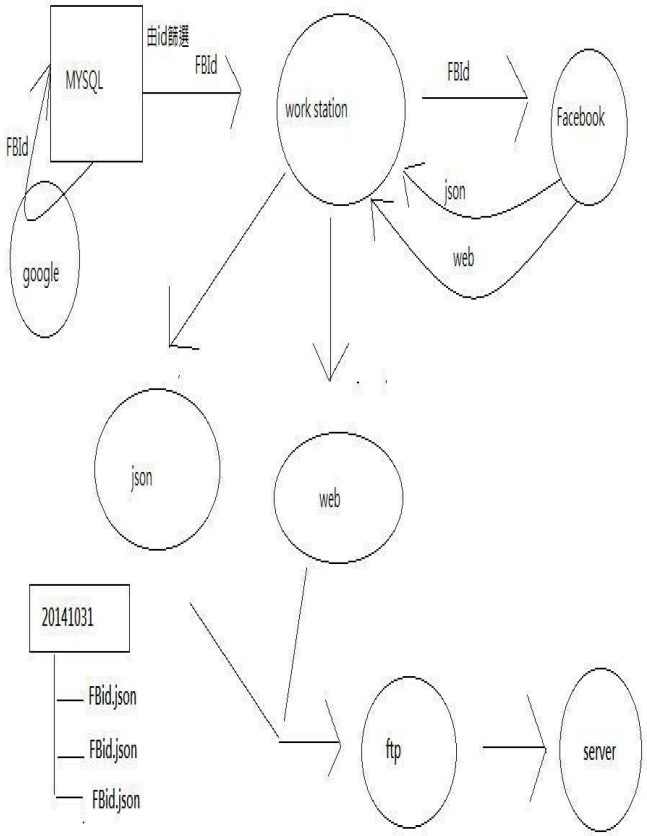

Figure-5: Crawling

VI. EVALUATION AND RESULTS OF CRAWLLING

We have used C# programming for crawling web pages and collecting data from the pages. After crawling we collected 147254 number of total Facebook ids out of which 133737 are valid ids.

- Total phone number extracted = 73839
- Total business hour information extracted = 34194

We evaluate the result by using ROI [12], precision, recall and f1-mesure [13]

A. *Evaluation of ROI*

ROI can be calculated as following

ROI = Number of entities obtained/number of ids crawled

- ROI of valid id = 133737/147254 = 0.9082
- ROI of phone number = 73839/147254 = 0.5014
- ROI of business hour information = 34194/147254 = 0.2322

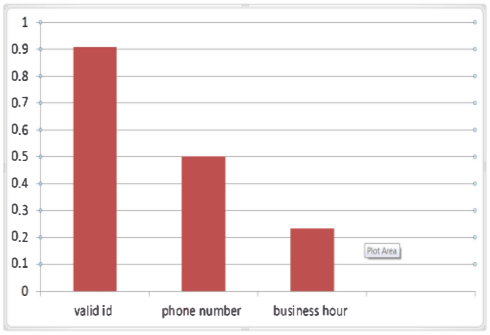

Figure-6: ROI

B. Evaluation of precission,recall and f1-mesure

Precision (P) of phone number and business hour

- P(phone number)= 73839/133737=0.5521
- P(business hour) = 34194/133737=0.2556

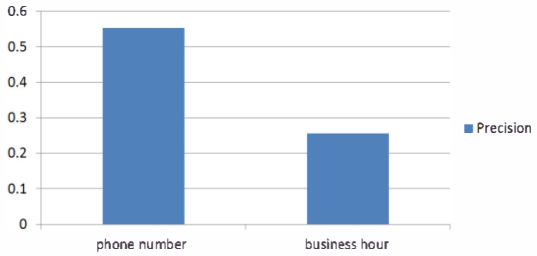

Figure-7: precision of phone no and business hour

Recall(r) of phone number and business hour

- r(phone number)= 73839/73839=1
- r(business hour) = 34194/73839=1

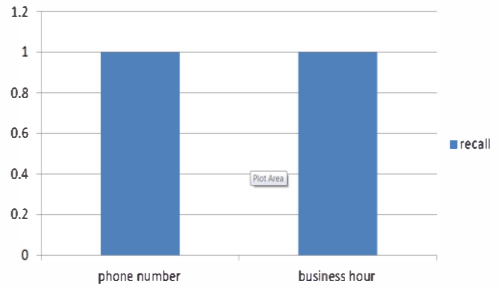

Figure-8: recall of phone no and business hour

F1-mesure (phone number) = (2 ×p(phone) ×r(phone))/ precision+ recall=(2 × 0.5521 ×1)/1.5521=0.7114
F1-mesure (business hour)=(2 ×p(business hour) ×r(business hour))/ precision+ recall=(2 × 0.2556 ×1)/1.2556=0.4071

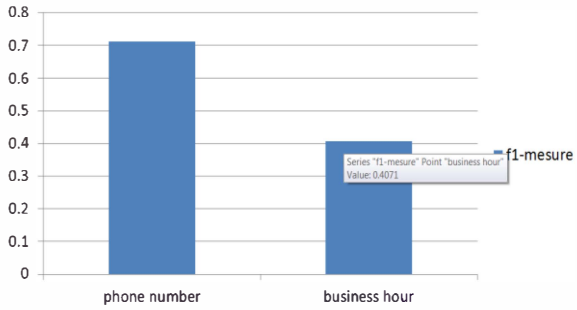

Figure-9:f1-mesure of phone no and business hour

VII. DATA EXTRACTION USING OUR SEARCH ENGINE

After collecting Facebook Jason page our program collects important information of the organization and stores in the SOLR database. For extracting querying and showing data to the user, we have designed a search engine in java as shown in figure-10 where user can type query and get the searched results on the interface.

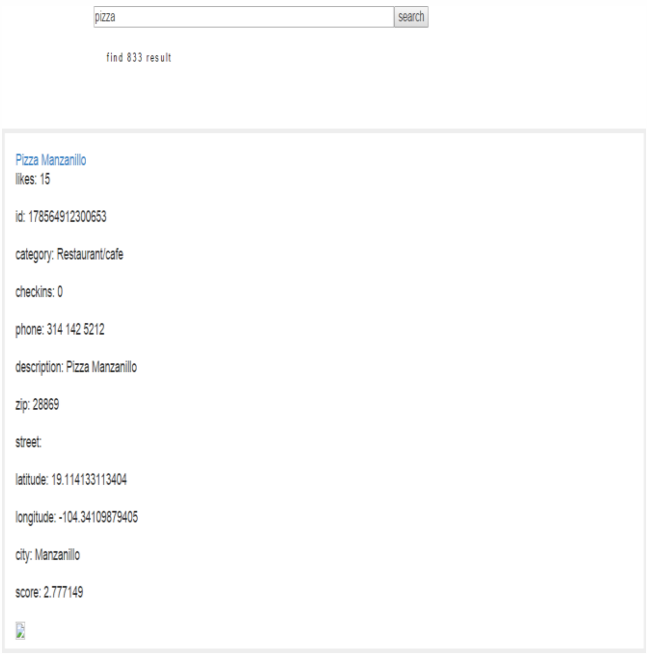

Figure-10: search interface

VIII. EVALUATION OF SEARCHING

For evaluating the search results we have used Rank-Based Measures such as Precision@K (P@K), Mean Average Precision (MAP) on 400000 Facebook id posted into solr database. We have evaluated the search result by using different keywords and used ranking for k=10, 20, 30…50 as shown in figure-11, figure-12, figure-13, figure-14.

978-1-4799-8579-1/16/$31.00 ©2016 IEEE

Precision (P@K):

P@K=rel/K

Where rel=no of relevant document

K=Rank threshold

keyword="Restaurant" "Taipei"

P@10=9/10=0.9

P@20=17/20=0.85

P@30=26/30=0.86

P@40=28/40=0.7

P@50=30/50=0.6

keyword=Restaurant in Kaohsiung

P@10=7/10=0.7

P@20=8/20=0.4

P@30=9/30=0.3

P@40=9/40=0.225

p@50=10/50=0.2

keyword=Restaurant in Mumbai

P@10=0.2

P@20=2/20=0.1

P@30=2/30=0.067

P@40=2/40=0.05

p@50=2/50=0.04

keyword=Restaurant with wifi

P@10=0.1

P@20=1/20=0.05

P@30=2/30=0.067

P@40=2/40=0.05

p@50=2/50=0.04

Mean Average Precision (MAP)

MAP=(P@K1+ P@K2+..P@Kn)/n

MAP("Restaurant" "Taipei")=

(P@10+P@20+...+P@50)/5=0.786

MAP("Restaurant in Kaohsiung")=0.365

MAP("Restaurant in Mumbai")=0.0914

MAP("Restaurant with wifi")=0.0614

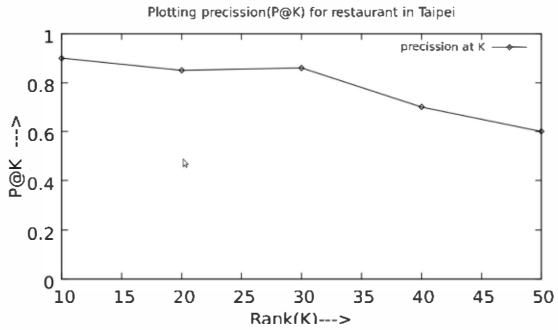

Figure-11: precision for restaurant in Taipei

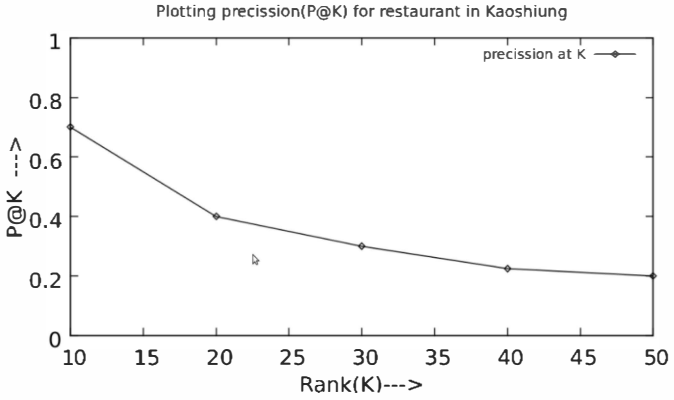

Figure-12: precision for restaurant in Kaohsiung

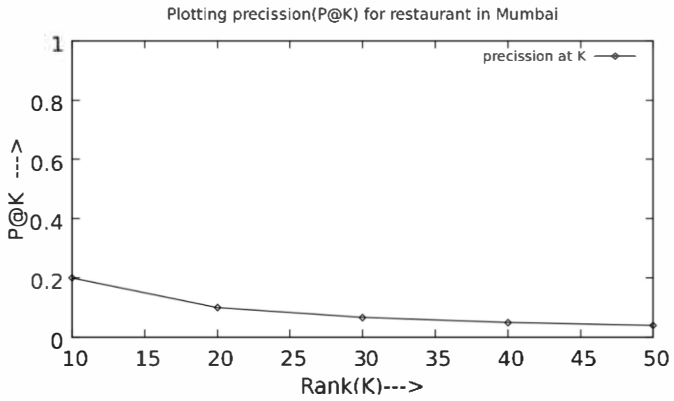

Figure-13: precision for restaurant in Mumbai

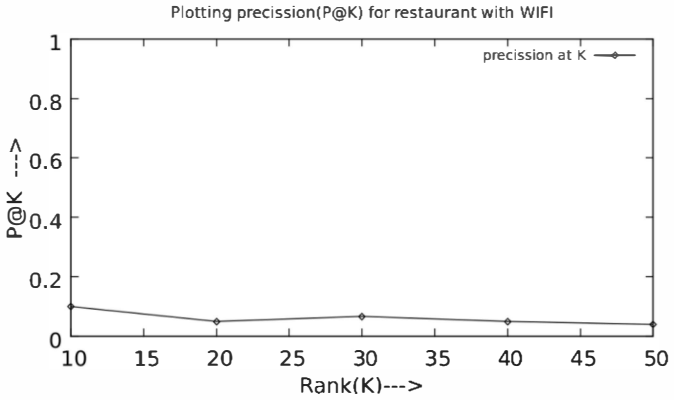

Figure-14: precision for restaurant with WIFI

In the graph above we can see that precision is maximum at k=10 and then the precision decreases, this is because at k=10 the page has high rank with high score and when the score and rank decreases the precision decreases accordingly. In figure-13 and figure-14 the precision of search for the keyword restaurant in Mumbai and restaurant with WIFI is less in comparison with restaurant in Taipei and Kaohsiung because we have fewer ids of WIFI and Mumbai restaurant in our database.

978-1-4799-8579-1/16/$31.00 ©2016 IEEE

IX. CONCLUSION

We crawled 147254 numbers of pages and extracted phone number and business hour and other important information of the organization. We calculated ROI and recall of business hour and phone number and compared them. Because the input of phone and business hour is keyed by human, so there might be some problem if it is not correct. Then we stored the collected data in SOLR database and designed our search engine for searching important information from Facebook. In future work we aim to extract business activity like special offer and other important information from fans page. We also aim to crawl more number of pages and also extend to other popular social networking sites like twitter, Google plus.

REFERENCES

[1] Hua Wang; Yang Zhang, "Web Data Extraction Based on Simple Tree Matching," in *Information Engineering (ICIE), 2010 WASE International Conference on* , vol.2, no., pp.15-18, 14-15 Aug. 2010.

[2] Siwu Fan; Xinjun Wang; Yongquan Dong, "Web Data Extraction Based on Visual Information and Partial Tree Alignment," in *Web Information System and Application Conference (WISA), 2014 11th* , vol., no., pp.18-23, 12-14 Sept. 2014.

[3] Li Kong; Yuchen Fu; Xiaoke Zhou; Quan Liu; Zhiming Cui, "Study on Competitive Intelligence System based on Web," in *Intelligent Information Technology Application, Workshop on* , vol., no., pp.339-342, 2-3 Dec. 2007.

[4] Chu, S., "Competitive intelligence on the World Wide Web," in *Professional Communication Conference, 1999. IPCC 99. Communication Jazz: Improvising the New International Communication Culture. Proceedings. 1999 IEEE International* , vol., no., pp.237-243, 1999.

[5] Journal of Computer-Mediated Communication Volume 13, Issue 1, pages 210–230, October 2007 .

[6] http://lucene.apache.org/solr/4_10_2/tutorial.html

[7] Yun Wei Zhao; van den Heuvel, W.-J.; Xiaojun Ye, "Exploring big data in small forms: A multi-layered knowledge extraction of social networks," Big Data, 2013 IEEE International Conference on , vol., no., pp.60,67, 6-9 Oct. 2013

[8] Valkanas, G.; Gunopulos, D., "Location Extraction from Social Networks with Commodity Software and Online Data," Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on , vol., no., pp.827,834, 10-10 Dec. 2012

[9] Kai-Yu Wang; I-Hsien Ting; Hui-Ju Wu; Pei-Shan Chang, "A Dynamic and Task-Oriented Social Network Extraction System Based on Analyzing Personal Social Data," Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference on , vol., no., pp.464,469, 9-11 Aug. 2010

[10] Kopka, M.; Kudelka, M.; Stolfa, J.; Kobersky, O.; Snasel, V., "Extraction and analysis social networks from process data," Computational Aspects of Social Networks (CASoN), 2013 FifthInternational Conference on , vol., no., pp.38,43, 12-14 Aug. 2013

[11] Borg, A.; Lavesson, N., "E-mail Classification Using Social Network Information," Availability, Reliability and Security (ARES), 2012 Seventh International Conference on
, vol., no., pp.168,173, 20-24 Aug. 2012

[12] http://en.wikipedia.org/wiki/Region_of_interest

[13] http://en.wikipedia.org/wiki/Precision_and_recall

978-1-4799-8579-1/16/$31.00 ©2016 IEEE