A Keyword-Based Big Data Analysis for Individualized Health Activity: Focusing on Methodological Approach

김한별*, 배근표**, 허준호***
* Unomic
**부산가톨릭대학교 응용과학대학 소프트웨어학과
***부산가톨릭대학교 응용과학대학 소프트웨어학과 조교수
**교신저자 e-mail : 72networks@cup.ac.kr

A Keyword-Based Big Data Analysis for Individualized Health Activity: Focusing on Methodological Approach

Han-Byul Kim*, Geun-Pyo Bae**, Jun-Ho Huh***
*Unomic, Busan, Republic of Korea
**Dept. of Software, Catholic University of Pusan
***Assistant Professor of Dept. of Software, Catholic University of Pusan

Abstract

It will be possible to solve some of the major issues in our society and economy with the emerging Big Data used across 21st century global digital economy. One of the main areas where big data can be quite useful is the medical and health area. IT technology is being used extensively in this area and expected to expand its application field further. However, there is still room for improvement in the usage of Big Data as it is difficult to search unstructured data contained in Big Data and collect statistics for them. This limits wider application of Big Data. Depending on data collection and analysis method, the results from a Big Data can be varied. Some of them could be positive or negative so that it is essential that Big Data should be handled adequately and appropriately adapting to a purpose. Therefore, a Big Data has been constructed in this study to applying Crawling technique for data mining and analyzed with R. Also, the data were visualized for easier recognition and this was effective in developing an individualized health plan from different angles.

1. Introduction

In the Republic of Korea (ROK), the number of obese people is rapidly increasing due to the changes in their lifestyle and eating habits [1-2]. According to the Ministry of Health and Welfare (MOHW), the prevalence of obesity for the population over 19 years old has increased from 26.0% (1998) to 29.2% (2001) and then to 31.7% in 2007 but remains at 31~32% for the last 7 years. During the same period, the rate increased from 25.1% (1998) to 36.2% (2007) showing an increase of 11pp for the last 9 years and remains at 35~38% afterwards whereas females are keeping the level of approximately 25% from 1998 to 2014 (2014 National Health Statistics I by MOHW, 2015).

Obesity can cause a variety of complications so that it has been considered as one of the major health and social problems around the world. Nevertheless, most of obese patients spend their lives without knowing a proper treatment for themselves. Although there are various obesity management services on and offline but each method is not quite helpful most of the time so that people are not interested in using the services.

2. National Health Data

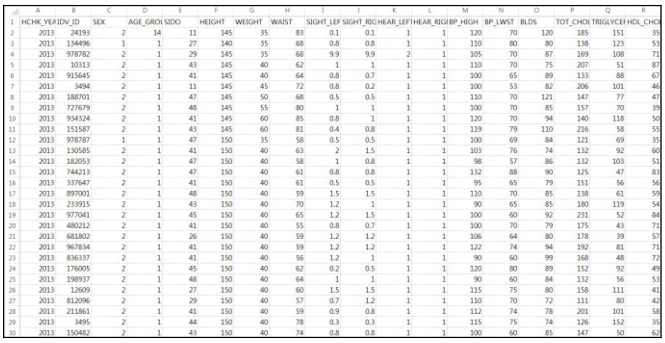
The National Health Data is a highly demanded public data opened to the citizens and private organizations in accordance with the Government 3.0 Policy. This data includes the information pertaining to the medical history, prescriptions and health checks of national health insurance subscribers accumulated by the National Health Insurance Corporation but excludes or masks personal or sensitive information for the safety. Currently, the data allowed to be disclosed are the ones that documented from 2002 to 2013 and the government is planning to increase the volume continuously. [Table. 1] shows the selection criteria for the National Health Data.

[Table. 1] The Selection Criteria for the National Health Data.

Classification	Content
Sample Extraction	Randomly extract one million patients who have received a health examination or a treatment in each year.
Limitation in Data Combination	Provide data by differentiating the personal serial number in each DB from the information request serial number.
Removal of Personal Identifier	Resident registration No. → Individual serial number (8-digit).
Categorization	Grouping the age levels: ages → age groups (between every 5 years of age), Categorize the people over 85 as ‘Above 85'.
Data Masking	Roughly classify sensitive Injury/Illness codes (D, O, P, X, Y :5 Classes, 114 symptoms).
Provision of Area Codes	Considering the perception of sample population residing in a small region, provide only the City/Province codes (17 Cities and provinces).

3. Reference of Big Data and Analysis Method

The health examination information is a dataset for the National Health Information and it consists of basic information and health examination results of 100 subscribers randomly selected from the subscribers who have received examinations for the selected items (Diseases) in the relevant year. The information dataset is constructed every year and includes 1million data and 34 attributes. (Fig. 1) shows the 2013 health examination dataset provided by the National Health Insurance Corp.

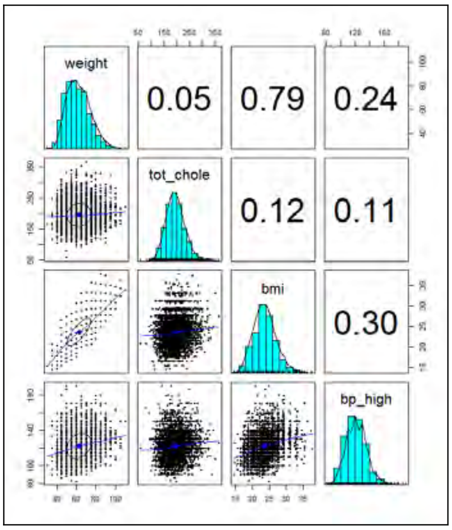


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	HCHK_YEA	IDV_ID	SEX	AGE_GROI	SIDO	HEIGHT	WEIGHT	WAIST	SIGHT_LEF	SIGHT_RIG	HEAR_LEFT	HEAR_RIGI	BP_HIGH	BP_LWST	BLDS	TOT_CHOI	TRIGLYCEI	HDL_CHO
2	2013	24193	2	14	11	145	35	83	0.1	0.1	1	1	120	70	120	185	151	35
3	2013	134496	1	1	27	140	35	68	0.8	0.8	1	1	110	80	80	138	123	53
4	2013	978782	2	1	29	145	35	68	9.9	9.9	2	1	105	70	87	169	108	71
5	2013	10313	2	1	43	145	40	62	1	1	1	1	110	70	75	207	51	87
6	2013	915645	2	1	41	145	40	64	0.8	0.7	1	1	100	65	89	133	88	67
7	2013	3494	2	1	11	145	45	72	0.8	0.2	1	1	100	53	82	206	101	46
8	2013	188701	2	1	47	145	50	68	0.5	0.5	1	1	110	70	121	147	77	47
9	2013	727679	2	1	48	145	55	80	1	1	1	1	100	70	85	157	70	39
10	2013	934324	2	1	41	145	60	85	0.8	1	1	1	120	70	94	140	118	50
11	2013	151587	2	1	43	145	60	81	0.4	0.8	1	1	119	79	110	216	58	55
12	2013	978787	1	1	47	150	35	58	0.5	0.5	1	1	100	69	84	121	69	35
13	2013	130585	2	1	41	150	40	63	2	1.5	1	1	103	76	74	132	92	60
14	2013	182053	2	1	47	150	40	58	1	0.8	1	1	98	57	86	132	103	51
15	2013	744213	2	1	47	150	40	61	0.8	0.8	1	1	132	88	90	125	47	83
16	2013	337647	2	1	41	150	40	61	0.5	0.5	1	1	95	65	79	151	56	56
17	2013	897001	2	1	48	150	40	59	1.5	1.5	1	1	110	70	85	138	61	59
18	2013	233915	2	1	43	150	40	70	1.2	1	1	1	90	65	85	180	119	54
19	2013	977041	2	1	45	150	40	65	1.2	1.5	1	1	100	60	92	231	52	84
20	2013	480212	2	1	41	150	40	55	0.8	0.7	1	1	100	70	79	175	43	71
21	2013	681802	2	1	26	150	40	59	1.2	1.2	1	1	106	64	80	178	39	57
22	2013	967834	2	1	41	150	40	59	1.2	1.2	1	1	122	74	94	192	81	71
23	2013	836337	2	1	41	150	40	56	1.2	1	1	1	90	60	99	168	48	72
24	2013	176005	2	1	45	150	40	62	0.2	0.5	1	1	120	80	89	152	92	49
25	2013	198937	2	1	48	150	40	64	1	1	1	1	90	60	84	132	56	53
26	2013	12609	2	1	27	150	40	60	1.5	1.5	1	1	115	75	80	158	111	41
27	2013	812096	2	1	29	150	40	57	0.7	1.2	1	1	110	70	72	111	80	42
28	2013	211861	2	1	41	150	40	59	0.9	0.8	1	1	112	74	78	201	101	58
29	2013	3495	2	1	44	150	40	78	0.3	0.3	1	1	115	75	74	126	152	35
30	2013	150482	2	1	43	150	40	74	0.8	0.8	1	1	100	60	85	147	50	62

(Fig. 1) The 2013 Health Examination Dataset Provided by the National Health Insurance Corp.

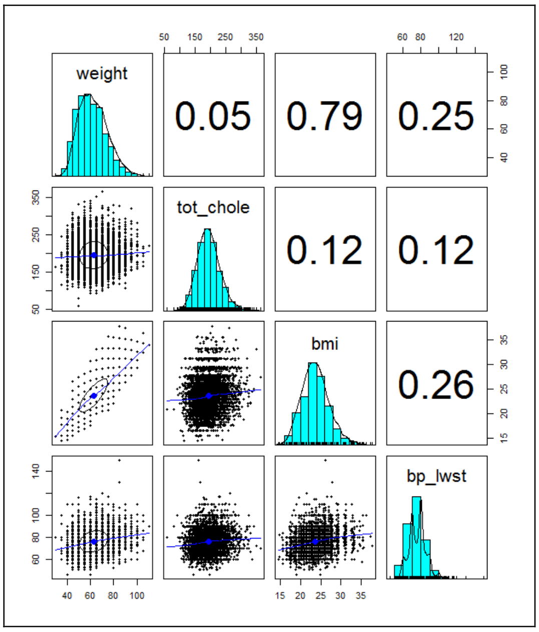
3.1. Analysis Method

Machine learning is not included in the basic setting of R. The RWeka, Class and Stats packages should be installed by using install.packages() function to use the machine learning algorithm implemented with R. In order to apply machine learning to the data, use library() function when necessary and then load the package(s). The first work is to create the Scatter Plot Matrixes for the variables such as 'Weight', 'Total Cholesterol', 'BMI (Bidy Mass Index)' 'Systolic Blood Pressure (maximal blood pressure)' and ' Diastolic Blood Pressure (minimal blood pressure)' to visualize the correlations between major attributes. (Fig. 2) shows the Scatter Plot Diagram for Systolic Blood Pressure and Variables.



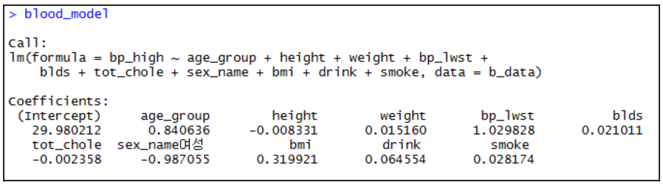
(Fig. 2) The Scatter Plot Diagram for Systolic Blood Pressure and Variables.

(Fig. 3) shows the scatter plot diagram for diastolic blood pressure and variables.



(Fig. 3) The Scatter Plot Diagram for Diastolic Blood Pressure and variables.

Here, the ellipse on the diagram is a correlation ellipse that shows the strength between variable. If the diagram stretches out and forms an ellipse as in the case of 'Weight and 'BMI', the correlation is considered to be strong but if the form becomes closer to a circle (e.g., 'Weight and Total Cholesterol), the correlation gets weaker. Next step is to create a binary variable 'Smoke' and 'Drink' by using ifelse() function to compare the blood pressures of those who Smoke (Drink) or do not Smoke (Drink). For the data model training, lm() function was used.



```
> blood_model

Call:
lm(formula = bp_high ~ age_group + height + weight + bp_lwst +
    blds + tot_chole + sex_name + bmi + drink + smoke, data = b_data)

Coefficients:
 (Intercept)     age_group        height        weight       bp_lwst          blds
   29.980212      0.840636     -0.008331      0.015160      1.029828      0.021011
   tot_chole   sex_name여성           bmi         drink         smoke
   -0.002358     -0.987055      0.319921      0.064554      0.028174
```

(Fig. 4) The Regression Coefficient of 'Blood_model' Model Object.

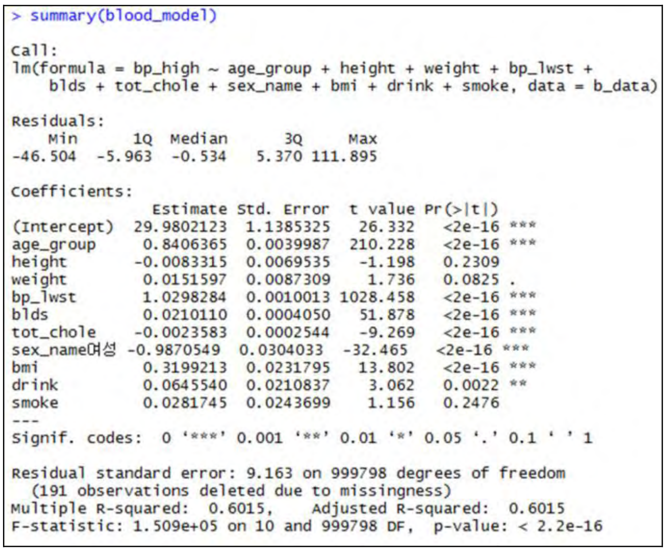
As shown in (Fig. 4), the estimated regression coefficient indicates how much bp_high (maximal blood pressure) has increased against the increase in each attribute while other attributes maintain a stable value. If the value of the age code age_group has increased by 1 while others maintained respective stable values, the bp_high increases by 0.83. As individual regression coefficients for 'Height', 'Weight', 'Blds (fasting serum glucose) and tot_chole (total cholesterol) were quite small, these values are considered insufficient to explain resulting blood pressure. The minimal blood pressure bp_l west has increased similarly to the maximal blood pressure and females had lesser blood pressure levels (0.98, in average) while 'BMI', 'Drink' and 'Smoke' had a closer relationship with blood pressure than other variables. Next, the model's performance was evaluated with summary().

3.2. Performance Analysis

(Fig. 5) shows the performance analysis of 'blood_model'.

The section 'Residuals' provides the summarized statistics for the errors and the maximum error of 111.895 means that the model would have that much difference in its predictive value in at least 1 exercise. The value from the multiple R-squared value indicates how well the model explains the dependent variables. Similar to the correlation coefficient, the model explains the data more perfectly as this value becomes closer to 1.0. For example, when the R-squared value is 0.6015, it means that the model (blood_model) can explain the dependent variables at the level of about 60%. A model with more attributes will be able to provide a higher value. Although the size of error is questionable, the regression model blood_model has a value of 0.6015 so that it can be considered that this model is actually working well.

As the National Health Insurance Corporation is currently providing the citizens' health examination data accumulated till 2013 only, analyses of the same data for other years are not easy. Thus, the next part of this research aims to provide an individualized/customized health information by studying various issues concerning national health problems and public's areas of concern by analyzing a variety of keywords in information disclosed through media.



```
> summary(blood_model)

Call:
lm(formula = bp_high ~ age_group + height + weight + bp_lwst +
    blds + tot_chole + sex_name + bmi + drink + smoke, data = b_data)

Residuals:
    Min      1Q  Median      3Q     Max
-46.504  -5.963  -0.534   5.370 111.895

Coefficients:
              Estimate Std. Error  t value Pr(>|t|)
(Intercept) 29.9802123  1.1385325   26.332   <2e-16 ***
age_group    0.8406365  0.0039987  210.228   <2e-16 ***
height      -0.0083315  0.0069535   -1.198   0.2309
weight       0.0151597  0.0087309    1.736   0.0825 .
bp_lwst      1.0298284  0.0010013 1028.458   <2e-16 ***
blds         0.0210110  0.0004050   51.878   <2e-16 ***
tot_chole   -0.0023583  0.0002544   -9.269   <2e-16 ***
sex_name여성 -0.9870549  0.0304033  -32.465   <2e-16 ***
bmi          0.3199213  0.0231795   13.802   <2e-16 ***
drink        0.0645540  0.0210837    3.062   0.0022 **
smoke        0.0281745  0.0243699    1.156   0.2476
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.163 on 999798 degrees of freedom
  (191 observations deleted due to missingness)
Multiple R-squared:  0.6015,    Adjusted R-squared:  0.6015
F-statistic: 1.509e+05 on 10 and 999798 DF,  p-value: < 2.2e-16
```

(Fig. 5) The Performance Analysis of 'blood_model'.

4. Method of Keyword Analysis

First, in order to analyze public's areas of concern, the study borrows the news published in online Naver News. Crawling technique was used to collect the articles related to obesity for the past year and analyze the frequency of the relevant words used. The other additional method is using a big data-based service where the trend in keyword search has been schematized for real time display, similar to the method used by Never Trend or Google Trend. The same process used for the earlier dataset analysis is also used for the collection of internet news. That is, using R for web crawling, and this time, the 'Text Mining' technique will be used to find some significant information implicit in the words used.

4.1. Text Mining

Text mining is a technique that extracts and processes high-level information such as pattern, trend and distribution obtained by analyzing the unstructured text [3]. Recently, due to widely spread use of Big Data, there is a growing interest in analysis technology for the high-volume texts, and as a result, the importance of the text mining technology is being emphasized. Text mining basically displays unstructured/ structured data into a simplified model. 'Word Cloud', which is widely used for visualization of data, is an image or graph composed of words used in a particular text or subject, in which the size of each word indicates its frequency or importance. Such a graph can provide analysis results economically and effectively. 'Wordle' is the most popular word cloud currently as anyone can use it on the open webs. A more detailed expression can be created by using R.

4.2. Word Cloud

Word cloud is a technique that visualizes the keyword(s) in a text to grasp its meaning or concept [4-5]. For example, the size of the frequently mentioned word will be enhanced depending on the frequency of use. This technique is employed to deduce the characteristics of data when analyzing the big data where a huge volume of data is contained. The packages such as ‘KoNLP’, ‘wordcloud’, ‘XML’, ‘stringr’, ‘httr‘, ‘rvest‘, ‘dplyr‘ are being offered by the big data analysis tool R for crawling, text mining and word cloud so that we have collected data through crawling after installing necessary packages.

4.3. Web Crawling

Web crawling is a computer program-based processing technique which explores the WWW systematically and automatically to collect a specific type of information [3-6]. This can be performed with the packages like library(httr), library(rvest), library(dplyr) offered by R. Analysis is usually conducted in the order of news search -> collection of relevant URLs -> collection of URLs of news articles -> perform crawling for the words within the article -> extract and store target texts.

First, the Naver News has been searched to identify public's areas of interest in obesity in each season. From January 1st to March 1st of 2016, the total number of webpages was 655 and the total number of articles was 6545. Crawling was performed for 500 articles in 50 pages. The same process has been taken for other seasonal periods and the number of target pages and articles are shown below. [Table. 2] shows the number of webpages and articles searched on the Naver News for a word 'Obesity'.

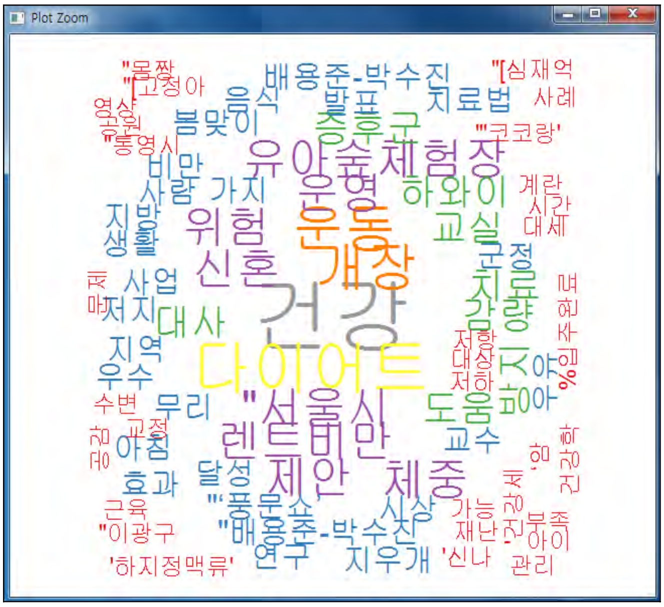
[Table. 2] The number of webpages and articles searched on the Naver News for a word 'Obesity'.

Period (2016)	No. of pages	No. of articles	Targeted for Crawling
Jan. 1st~ Mar. 1st	655	6545	500
Mar. 2nd-Jun 1st.	915	9145	450
Jun. 2nd-Sep. 1st	772	7716	535
Sep. 2nd-Dec. 1st	887	8863	525

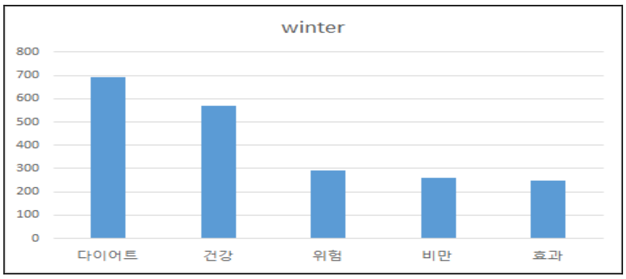
The collected text files might contain not only the words significantly related to obesity but also unnecessary words like special characters and numbers so that gsup() function which filters out these meaningless words was used when performing the keyword-based extractions. This function changes the user-designated characters or signs into desired characters or spaces.

And then, filtered 'winter.txt' was represented with word cloud to check the frequency of each word. Next, the top five most used words in the period were selected and visualized with a graph to interpret their significant correlations.

(Fig. 6) shows search for the obesity-related words in winter season using word cloud. Also, (Fig. 7) shows the graph of frequency of 'winter.txt' words.



(Fig. 6) Search for The Obesity-related Words in Winter Season Using Word Cloud.



(Fig. 7) The Graph of Frequency of 'winter.txt' Words.

The most frequently used words during the winter season were Diet, Health, Danger, Obesity and Effectiveness. From the analysis, we were able to find that the obese people were interested in the words such as diet and health, as well as the possible dangers due to obesity during this season. Next, we perform a crawling-based analysis for other seasons.

5. Conclusion

The targets for data mining were the news articles included in the Big Data-based services such as Google Trend or Naver News. Crawling technique was applied to data mine the dataset of approximately one million health-related information provided by National Health Insurance Corporation. Machine learning was applied for data analysis with R. The primary goal of this is to identify the distribution of obesity-related data including target patients' heights, weights and genders. The focus was on finding out the degree of obesity depending on each patient's weight. The result from the analysis emphasized that the patients over 85kg need to be careful about their diet as the degree was distributed extensively over that boundary. Considering that there could be many other variables contributing to obesity in today's hyper-nutritive world, the dataset should be analyzed from many other angles as well. Although Crawling and Word Cloud are being considered as a useful data schematiztion tool to clearly and elliptically distinguish data, the users of these tools should be aware that they need to use them carefully and correctly as the results may vary from their application methods.

Acknowledgments

The 4D Health Care Project Group of Catholic University of Pusan aims to cultivate the creative talent who have capabilities in developing 4D contents required for rehabilitation and health care of modern people. Both Department of Physical Therapy and Department of Software of this university are participating and operating the group jointly to perform the task.

The 4D Health Care refers to an advance health care technology which is used for the operation in a 4D-based mixed reality where human senses, cognition and experiences (1D) have been converged with both real and virtual information (3D) and the project group runs various curricular and extracurricular programs to train every participating student to acquire a 4D technology-based health care contents development skills.

This has been written with the support of the 4D Health Care Project Group and the author wishes to express his gratitude to the Ministry of Education, National Research Foundation of Korea, as well as the CK Project Group. And this paper is the product of a team project performed in “Advanced Java Programing” Course at Dept. of Software, Catholic University of Pusan for the undergraduates.

Also, my gratitude extends to Catholic University of Pusan and the Lord who has provided me with his wisdom and grace.

References

[1] Jun-Ho Huh, Han-Byul Kim, Kyungryong Seo, “Preliminary Analysis Model of Big Data for Prevention of Bioaccumulation of Heavy Metal-Based Pollutants: Focusing on the Atmospheric Data Analyses,” ASTL, SERSC, 129 (2016), 159-164.

[2] D. Battré, M. Hovestadt, B. Lohrmann, A. Stanik, D. Warneke, “Detecting bottlenecks in parallel dag-based data flow programs,” In: 2010 3rd Workshop on Many-Task Computing on Grids and Supercomputers, (2010).

[3] A. Behm, V.R. Borkar, M.J. Carey et al., “Asterix: towards a scalable, semistructured data platform for evolving-world models,” Distrib. Parallel Databases, 29(3), 185-216. 2011.

[4] K.S. Beyer et al., “Jaql: a scripting language for large scale semistructured data analysis,” Proceedings of the VLDB Endowment, 4(12), 1272-1283, 2011.

[5] C. Boden et al., “Large-scale social media analytics on stratosphere,” Proceedings of the 22nd International Conference on World Wide Web - WWW '13 Companion, 2013.

[6] V.R. Borkar et al., "Hyracks: a flexible and extensible foundation for data-intensive computing," 2011 IEEE 27th International Conference on Data Engineering, (2011), 1151-1162.