ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645
Indian Journal of Science and Technology, *Vol 8(S8), 58-63, April 2015*
DOI: 10.17485/ijst/2015/v8iS8/64714

Creating Values from a Noisy Accumulated Contents based on Data Analysis

ChulSu Lim , Wongoo Lee, YoonYoung Joon and Kangryul Shon*

NTIS Center, Korea Institute of Science and Technology Information, Daejeon, 305-806, South Korea;
krshon@kisti.re.krr

Abstract

The contents of the information system play a major role to user services. The quality of contents not only depends on the accuracy and availability but also depend on the depth of the information. As the size and the quality of the information item increases, the need to create meaningful analyzed data also increases. But it is not easy to extract valuable information from the unfiltered noisy data. Using these accumulated data, we want to add valuable information based on data analysis. With a preliminary validation of data items in a preparation step, we found that about 70% of data items could be used as a source of getting statistics. After applying time series analysis, correlation analysis between data items and regression analysis we found some informative relations between the data items. These value added information could be added to the original data set as a source of another analysis.

Keywords: Analysis, Regression, Relation, Time-series

1. Introduction

As information is growing in size with high quality, the importance of the utilization of the information are increasing. The contents of the information system play a major role to user services. The quality of contents not only depends on the accuracy and availability but also depend on the depth of the information. As the size and the quality of the information item increases, the need to create meaningful analyzed data also increases. But it is not easy to extract valuable information from the unfiltered noisy data. Using these accumulated data, we want to add valuable information based on data analysis. NTIS (National Science and Technology Information Service) provides overall Korean national R&D information[1–2,5]. This service provides more than 107 million information items related to national program information gathered from 17 ministries and offices. The information gathered from these institutions is based on national R&D standard information. The information includes government invested programs, projects in the program, participants for the project, equipment used for the project, and outcome from the project. For the quality of the information, there are data constraints to validate the soundness of the data. The quality reached the highest level in public sector in Korea[2]. To enhance data quality, integrity constraints are managed and incorporated into a database using triggers[3,4].

2. Related Works

While detail 330 information items on NTIS service pages play their roles, but there could be more sophisticated values based on the information items. From 2000, the information items was investigated and analyzed to produce statistics to help understand the information1. The analyzed statistics and charts are used as a basis for the policy decisions related to national R&D activities like budget investment or budget distribution. The items for this analysis are in 21 categories. For example title of the project and period of the project belong to 'basic project information' category. Within these categories, there

*Author for correspondence

are 114 information items. The results of the analysis are grouped in 3 step hierarchy for the easy access of the information. Users will find charts and corresponding statistics on their screen as shown in Figure 1[5].

role in this data set. And all other attributes are linked to the 'Project' attribute with an identifier for a project. Many entities like 'Paper', 'Patent', 'Research Paper' belongs to 'Outcomes' entity which are also linked to the 'Project' entity. In general, this data set is considered

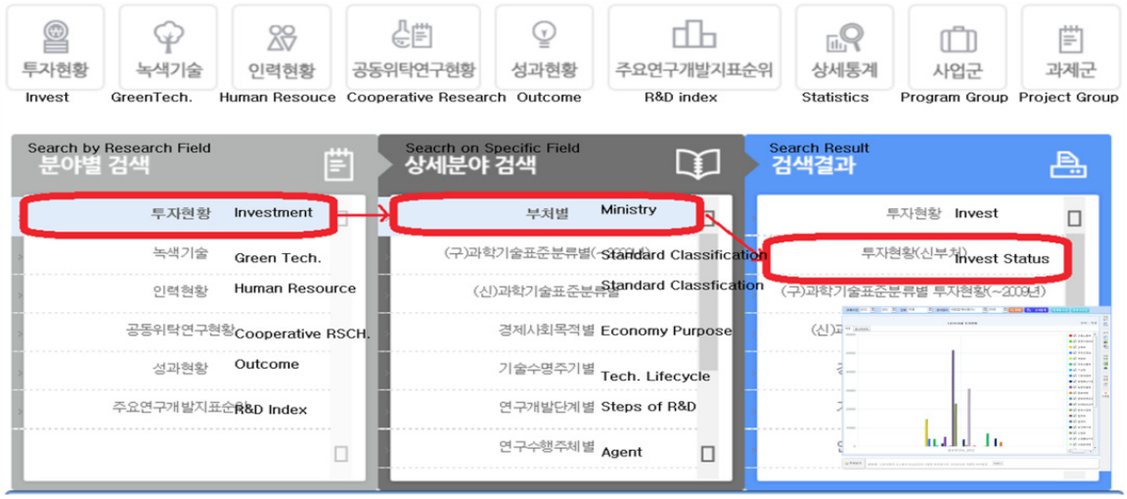

Figure 1. Services of an Analysis Result.

3. Creating Values using Data Analysis

3.1 Learning Target Data

Figure 2 shows a data schema for national research and development information management.

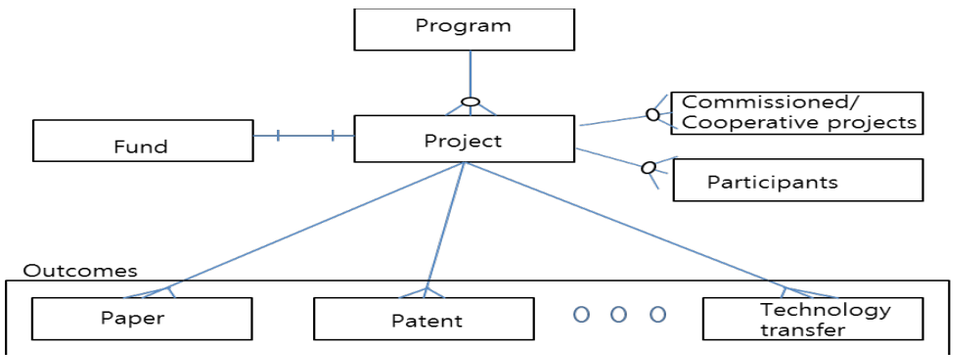

Figure 2. A Data schema for R&D information

As shown in Figure 2, project information is the key attribute to connect other information items like program, participant, fund and outcomes. Among these items, selected 114 information items are our target data to analyze. So the attributes in 'Project' entity plays major role as a structured data because the structure of the entity and attributes in the entity is fully described. But some of the attributes like 'abstract', 'expected effect' in the data-set belongs to unstructured data, in which there are no restriction in the text, in order words these attributes have no inherent structure.

3.2 Data Preparation

Figure 3 shows steps for the analysis of data. These steps are comparable with the data analytics lifecycle which contains discovery step, data preparation step, data modeling step, model building step, communicate result

step, and operationalize step. In the lifecycle the steps are reversible whenever required. 'Survey' step is a kind of discovery phase in which we investigate and learn specific business domain knowledge. We have to check previous attempts, analytic methods, applied techniques and outcomes in that domain area. The understanding of the domain area is critical factor that affects entire process. Deeper understanding of domain knowledge helps determine methods and models we have to apply to the specific problem we want to solve.

After the survey on the data sets, the data could be prepared for the analysis. In this stage, error correction or refinement of data could be conducted. The prepared data could be used as an input to a preliminary test. The validated data sets are used as the input to the analysis stage. The analysis models are listed for model planning. Among many models considered, promising models could be used as a basic analysis. After evaluating the model to the date, we could recheck the data set and models for further analysis. Basically, these steps are prerequisite to many big data analysis platform based on a hadoop file system[9–11].

2012. About 80% of them are based on budgets of the projects. And remaining 20% of the graphs are focus on the participants of the projects.

The main topics of the analysis are the subject of the researches, kind of researches, and location of the researchers. For the subject of the researches, the graph shows type of organization of the researcher, gender and age of the researcher and degree of the researcher. For the kind of researches, the analyzed graphs shows number of projects with predefined classifications such as 'standard classification of national science and technology', ' 6T (for example, IT (Information Technology), ET (Environmental Technology), BT (Bio Technology))', 'NTRM (National Technology Road Map)'.

Among 44 graphs, 18(41%) graphs adopt time series analysis. Most of them are for 3 years or 5 years depending on the availability of the data attributes.

To expand analysis to other data items, we should understand logical meaning of the data shown in Figure 3, Figure 4 shows a part of preliminary analysis result for the national R&D information.

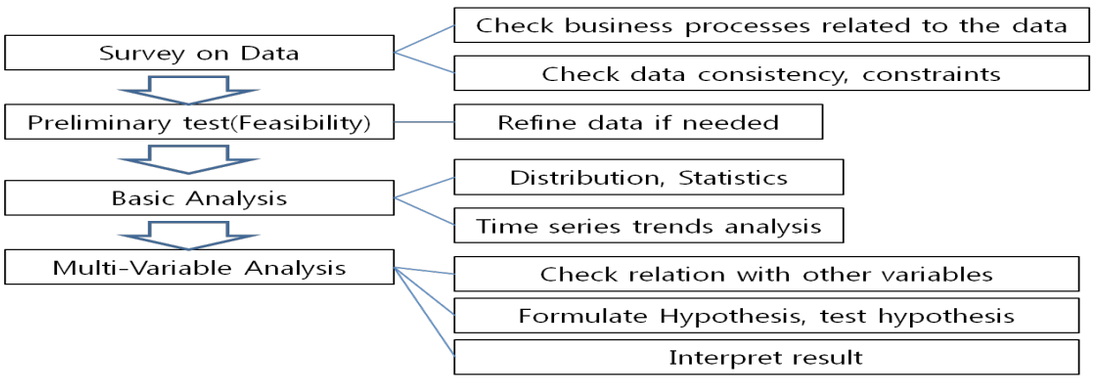

Figure 3. Steps for the analysis of data.

3.3 Time Series Data Analysis

These statistics includes investment by ministries, participant by degrees. These analyzed results which focus on just one year could be expanded to multi-year time series analysis. For example, this result could be expanded 3 year trends over time. For short term trends we use 3 year data and for long term trends, we use 5 year data.

3.4 Preliminary Data Analysis

Among 330 items, only 114 items are used for the analysis for now. And 37 items are reserved for another use. In the current analysis report, there are 44 major graphs on

We excluded data items which consist of free text. Those are not appropriate for statistical processing. These items could be used as information source for semantic knowledge acquisition later.

From 56 items which are in formats as code data, numbers, fixed names, dates, 39 items (69.6%) proved to be applicable to further analysis. The validity for the analysis was marked in 'availability' column 'Y' if the full scanning of the contents indicates sufficient occurrences with tolerable error rates.

3.5 Data Analysis

category	item	availability	Data scan result
program	Executed fund	Y	Valid from 2010 to 2013
	Investment priority	N	NULL 70%
	Contract fund	Y	Valid from 2010 to 2013
project	Start date	Y	
	End date	Y	
	Managing institute	Y	
	performing institute	Y	
participant	Start date	Y	
	End date	Y	
	Ratio of participation	Y	Valid from 2008 to 2013
Fund	Cash amount of enterprise fund	Y	
	spot amount of enterprise fund	Y	
	Fund of participating country	Y	
	Funds except government	Y	
	Governmental funds	Y	
Support for training	Number of papers	N	Valid through 2012-2013
	Number of dr. degrees	N	Valid through 2012-2013
	Number of ms. degrees	N	Valid through 2012-2013
	Number of bs. degrees	N	Valid through 2012-2013

Figure 4. Preliminary analysis of items.

3.5 Data Analysis

With 39 feasible information items which proved useful from preliminary analysis, we analyzed the data to find out implication of the analysis. Figure 5 shows an example analysis result with project start year and month information item.

Initial hypothesis was projects starts in January and ends December. But as a matter of fact, May is the most preferable month for the start of a project and likewise instead of December, May or August of next year are the most frequent project ending month. That means expected month for research report is May instead of December.

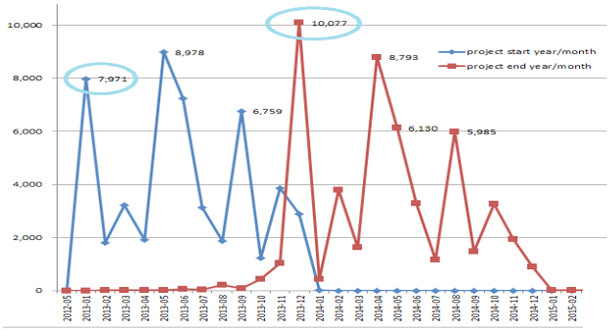

Figure 5. An analysis result with project start and end item.

3.6 Time Series Trend Analysis

After analyzing basic statistics and distribution of an information item, we could compare the information with previous results. With time series analysis we could understand current information. Also, on the basis of the trend information from the analysis we could predict future results.

3.7 Correlation Analysis

A variable could be dependent or related to another variable. The correlation value between the variables will tell whether the variables are related. The value of the correlation is between -1 to 1. Closer to 1 (-1) means positively (negatively) related and closer to 0 means no relation.

Table 1. Relation between two variables

Year	# of projects in basic researches	# of papers from the project
2002	128	209
2003	299	569
2004	762	1,444
2005	2,333	4,965
2006	4,912	19,565
2007	5,560	18,869
2008	6,169	32,606
2009	5,527	33,381
2010	6,979	37,324
2011	8,875	43,179
2012	9,876	47,377
2013	8,480	36,456

In this example in Table 1, the correlation value of the two fields is 0.9715, which means these two variables have close relation. That means as number of basic research project increase, the number of papers from the projects also increases. With this analysis result, we could make a decision to invest more projects in basic researches to get more papers from the projects.

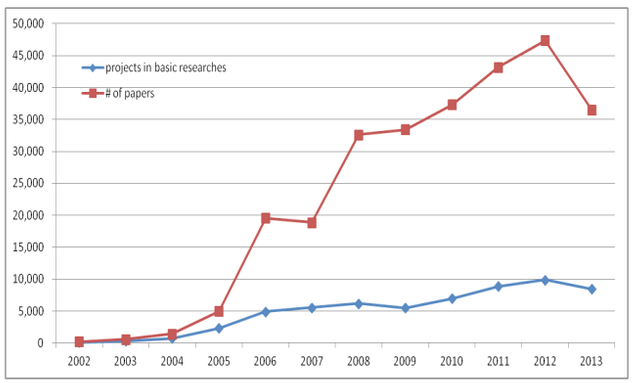

Figure 6. A graph of the number of papers versus number of projects.

Figure 6 shows a graph with the two variables shown in Table 1. With the graph, we can guess a kind of linear relation between the two variables, which could be confirmed by a regression analysis.

A regression analysis shows R square value is 0.94, which is a very good fit. This values means 94% of the number of the papers could be explained by the number of the projects in basic research areas. The closer to 1, the better the regression line fits the data. And the significance F is 1.41061E-07 which is low enough to rely on the statistics. The value should be below 0.05 to check if the results are reliable. If the value is greater than 0.05, then we should stop using this variable to predict another variable. In a situation where we want to identify meaningful dependencies between the variables, we could delete these independent variables with these higher p-values.

The regression line is 'Y = -2134.938 + 5.0344 X'. In other worlds, for each increase in number of the projects, the number of the paper increases by 5.0344 units. This line could be used for a forecast. The residuals as in Figure 7 show you how far away the actual data points are from the predicted data points using the equation. We could review the difference between the predicted values and actual values.

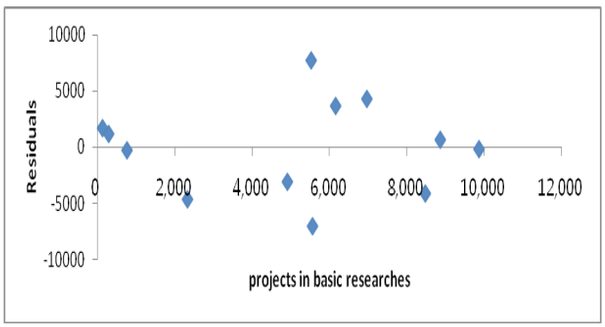

Figure 7. The residuals of the prediction and actual data points.

We apply these correlation and regression analysis to the variables in national research and development data sets. The value for the correlation between the number of paper/patents and the number of the commercialization is 0.00291, with which we could say there is no relation between them. The list is the founding from the analysis.

- The amount of fund by the government vs. the number of participants in the projects : weak relation (0.27)
- The amount of fund by the government vs. the amount of fund of commissioned project : weak relation (0.25)
- The amount of fund by the government vs. the amount of fund for cooperative project : no relation (0.17)
- The amount of fund by the government vs. The age of the director of the project : no relation (0.08)

With the analysis of the gathered national R&D information, we could extract values from basic distribution statistics to time series trends information and relation information between the variables.

4. Conclusion

This paper discusses about analysis of constructed national R&D information items. The increased size of the information can be a good source for quality information. But not only the data itself but also statistics and induced analysis of data could be a valuable information source. To extract valuable information from the unfiltered noisy data, we filtered the information for preliminary analysis in data preparation stage. Preliminary analysis identified many candidate items to create values with national R&D data meaningful information could be derived using those items. With the preliminary validation of data items in a preparation step, we found that about 70% of data items could be used as a source of getting statistics. After applying time series analysis, correlation analysis between data items and regression analysis we found some informative relations between the data items. These value added information could be added to the original data set as a source of another analysis. The relations with these items should be studied in future researches.

5. Acknowledgement

This research was supported by the Sharing and Diffusion of National R&D Outcome funded by the Korea Institute of Science and Technology Information.

6. References

1. LLim CS, Kim JM, Yoon YJ, Shon KR, Kim JS. Data content constraint management for national r&d data quality improvement. International Conference on Convergence Contents; 2011. p.17–18.
2. Lim CS, Shon Kr, Kim TH, Han SW, Lee WG, Kim JM. Improvement of R&D information management to support convergence research. International Conference on Convergence Technology; 2013.
3. Ceri S, Widom J. Deriving Production Rules for Constraint Maintenance. Proceedings of 16th International Conference on VLDB; 1990 Aug 13-16; Brisbane, Australia. p 566–7.
4. Ceri S, Cochrane R, Widom J. Practical applications of triggers and constraints: success and lingering issues. Proceedings of 26th International Conference on VLDB; 2000 Sep 10-14; Cairo, Egypt. p. 254–62.
5. Available from: http://www.ntis.go.kr
6. Mosteller, Frederick, Tukey JW. Data analysis and regression: a second course in statistics. Addison-Wesley Series in Behavioral Science: Quantitative Methods. 1977.
7. Hosmer, David W, Lemeshow S, Sturdivant RX. Introduction to the logistic regression model. John Wiley & Sons, Inc; 2000.
8. Phillips, Peter CB, Perron P. Testing for a unit root in time series regression. Biometrika. 1988; 75(2):335–46.
9. Zikopoulos P, Eaton C. Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. 1st ed. McGraw-Hill Osborne Media; 2011.
10. Kwak JH, Yoon J, Jung YH, Hahm J, Park D. Large-scale data analysis based on hadoop for Astroinformatics. Journal of KIISE. 2011;17(11):587–91.
11. Borthakur D. The hadoop distributed file system: Architecture and design. Apache Software Foundation; 2007.