November-December 2019
ISSN: 0193-4120 Page No. 2480 - 2485

Big Data Based Decision-making Support System Design for Efficient Analysis of the Performance of Software Education

Ji-Hoon Seo[1], Kil-Hong Joo[2]
[1]Department of Computer science, Incheon University, Korea, sserz@naver.com
[2]Department of Computer Education, Gyeongin National University of Education, Korea, khjoo@ginue.ac.kr

Article Info
Volume 81
Page Number: 2480 - 2485
Publication Issue:
November-December 2019

Article History
Article Received: *5 March 2019*
Revised: *18 May 2019*
Accepted: *24 September 2019*
Publication: *12 December 2019*

Abstract

This paper is aimed to analyze the performance of domestic software education that has been implemented thus far centering on the opinion mining visualization analysis technique based on big data related to domestic software education collected over the last five years such as data from portal site news, SNS service, Internet café, and other performance data and develop a big data based decision-making support system for Korean style software education in order to derive agendas that will be discussed hereafter. In addition, through the prediction of more advanced future-oriented education systems and analysis of performance factors based on the foregoing, solutions for big data based decision-making support prediction analysis can be prepared hereafter in the field of education and measures for cultivation of creative convergence talents and promotion of software education can be sought so that the directions of mid/long-term development of Korean style software education can be accurately set.

Keywords: *Big Data, Software Education, Decision-making support system*

1. INTRODUCTION

Recently, R&D project plans linked to the fourth industrial revolution on which the South Korean government focuses have been implemented. Those projects include national projects utilizing big data, which are applied to diverse fields such as ICT, BT, and Smart-City and are fully supported by the government including the establishment of new big data related organizations and laws. On the contrary, measures to utilize big data in the field of domestic education are still slow and are hardly supported at the government level [1][2]. However, thanks to the excellent domestic IT infrastructures, projects utilizing big data are now gradually progressing in the field of education and since it has become possible to make massive data accumulated over several years into big data, such pieces of big data information can now be linked with each other for analysis [3][4]. Meanwhile, since various departments have been implementing systematic programs intended to improve the constitution of education and cultivate talented persons such as software education and creativity, personality, and convergence education centered on learners' problem solving ability and computational thinking skills, diverse and huge data and pieces of information lie scattered [5][6]. While analyzing South Korean software education systems over the last five years, we have achieved eye-opening performance in reducing the sense of difference between education and the people [7]. However, there has been no method to concretely quantify and measure the performance and although the results of individual analyses exist, there is no appropriate tool that can organically link those pieces of information with each other for analysis [8]. In addition, in South Korea, still there is no system at all that can construct big data using unstructured texts to predict decision making for educational outcomes. Therefore, in this paper, based on big data analyses, a decision-making support system that can seek the performance and the direction of future oriented development of software education that has been reorganized into compulsory education at elementary/middle schools from this year will be designed, the education policies implemented by various government departments will be evaluated in general, the big data based on the unstructured data accumulated in online media will be analyzed centering on opinion mining among the three unstructured data analysis techniques, that is, text mining, data mining, and opinion mining, and based on the foregoing, a big data based decision-making system design model that can derive the prediction of changing future education systems and the direction of promotion of

polices as quantitative performance will be proposed.

2. RELATED WORKS

2.1 Core Strategies of South Korean Software Education

Focusing on the stable settlement of SW education in the South Korean school field, computing thinking ability was defined as 'the thinking ability to efficiently solve problems that may occur in students' daily lives based on the basic concept and principles of computing.' It has been pointed out that although the importance and necessity of SW education have been sufficiently recognized, for SW education to be effective, the goal and content of the curriculum should be concretized and the operation of the curriculum should be flexible[8][12]. Accordingly, although computing thinking centered studies have continuously achieved good outcomes, such studies show a shortcoming of being not sufficiently realistic [9][10].

2.2 Core Strategies of South Korean Software Education

The SW education implemented in the 2009 revised curriculum comprises one unit of the subject manual training for the 5th grade of elementary school and an elective subject (information) for middle school, which is selectively taught at some schools. Whereas the content of education of the subject information in the existing curriculum is focused on the utilization of information and communication technologies, the content of education in the operation guidelines for software education has been said to be focused on the reinforcement of information ethics and the improvement of problem solving ability utilizing algorithms and programming (Ministry of Education - KERIS, 2015). According to the 2015 revised curriculum, in the case of elementary school, the existing unit information centered on ICT utilization should be changed into a large unit centered on basic knowledge of SW to implement the education with a content, which is centered on program solving processes and experience in algorithms and programming and includes the fostering of information ethics consciousness [11][12].

2.3 Utilization of Education Big data

Cases where educational (big) data are utilized for services in South Korea include the Edu Data Service System (EDSS). The EDSS prepared a system to collect, link, and process the education related data accumulated in the Ministry of Education, Metropolitan and Provincial Offices of Education, and education related organizations and provides data requested by researchers for the purpose of academic research through a specified examination procedure. The EDSS provides data sets for a total of 11 fields comprising seven fields of elementary/middle/high school educational data (school information disclosure, elementary/secondary education statistics, college scholastic ability test data, national academic achievement evaluation data, special education statistics, NEIS data, Edufine data) and four fields of higher education statistics(university information disclosure, higher education statistics, employment statistics, lifelong education statistics) [13].

2.4 Education Data Mining

According to previous education big data utilization related studies, implications were derived through the analysis of the relationships between diverse variables such as the method of utilizing education big data for learning and the frequency of inquiries about education big data. However, such analytical models have limitations for analysis of performance factors in creativity education and seeking for mid/long-term development.

Thus far, education policy analysis methods enabled the relevant analyses only when the agenda has been presented in advance and accordingly, policies have been implemented through subjective keywords in the improvement of educational environments. In addition, although implications of education systems can be derived when the policy importance is judged based on the inquiry about the education big data, in the case of the databases, which are not inquired, reference data that will enable the application or comparison of policy demand become necessary [14].

2.5 Opinion Mining Analysis Technique

Opinion Mining is a classification of text mining and is also called reputation analysis. It determines the positive, negative, and neutral preferences of formal and informal texts in social media and is utilized for the forecast of the market sizes of specific services and products and analyses of consumers' responses and words-of-mouth, etc. Opinion mining extracts vocabulary information expressing positive and negative responses and recognizes object and sentences consisting of opinions about the object to measure positive and negative responses with the sum of patterns including opinions. As such, opinion mining can obtain more valuable information from informal data made up of opinions of many unspecified users. To interpret in another aspect, it is also called sentiment analysis and is interpreted as the broad meaning of natural language processing, computer linguistics analysis, and text mining [15].

3. PROPOSED METHOD

For the big data based decision-making support system proposed in this paper, a dictionary context based pre-treatment process was implemented through which all domestic software education related unstructured data distributed in online media were collected and inappropriate keywords were removed, and a process was configured for design of an efficient decision-making support system for

analysis of the performance of software education performance.

3.1 Development of Korean SW Education Data Collections, Storages, and Processing Technologies

In this paper, to reduce the dependency on real-time data collection and present correct directivity through the prediction of the future of domestic software education, the python-based crawling technique was used as a method to collect raw data. Atypical data such as online news articles, blogs, social networks(Twitter and Facebook) from 2012 when SW education became an issue to 2017 were collected and classification, preprocessing, definition for data standardization, and bidirectional sharing systems were grafted centering on big data storage and processing technologies for decision making support systems.

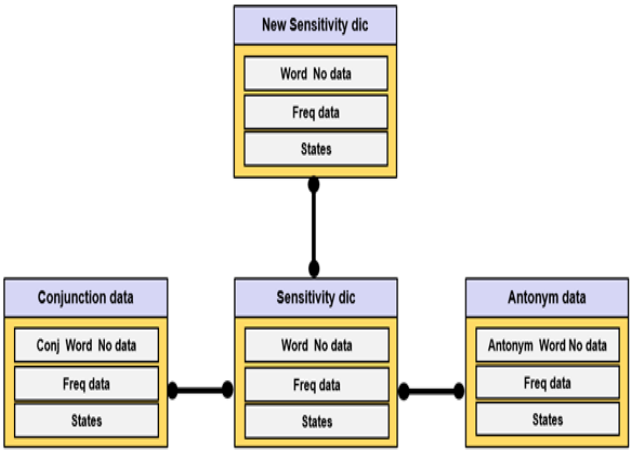

Figure 1: Korean Grammar based Sentiment Dictionary DB Join Design

3.2 Development of a Data Integration System for Prediction and Analysis of Atypical Data for Education

Since configuration systems that are stable in and suitable for physical environments such as DW(Data Warehouse) for decision making analysis, an environment where atypical data can be distributed for processing utilizing storage servers and master servers in constructing big data was constructed. An accuracy improving technique was designed to extract reliable affirmations and negations from text sentences in creative education contents using opinion sentiment dictionaries trained on individual atypical text documents. For ETL-based data extraction, transformation, and loading, a disk sharing mode system was constructed to store and refine past raw data, database modeling was constructed utilizing RDBMS, and Map Reduce-based and Stream-based big data environments were designed.

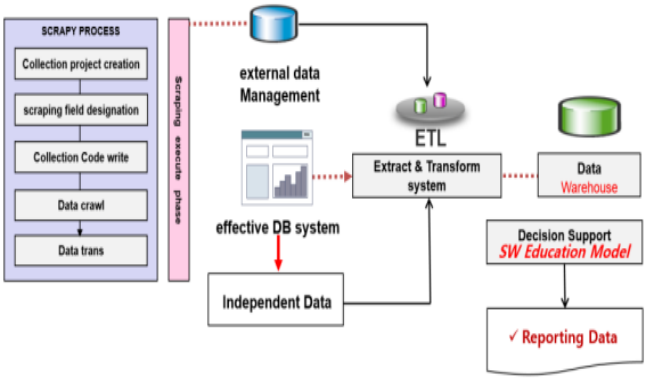

Figure 2: Big Data Decision Making Support and Storage Model of Software Education

3.3 Korean Grammar based Pre-Treatment Design

To design a decision-making support model for opinion mining analysis based on Korean grammar, appropriate word preprocessing design and grammatical rules should be applied. In this paper, a data preprocessing process was implemented to select candidate groups from the words in the sentiment dictionary and extract sample data for the decision-making support model from the candidate groups. In this paper, outliers, missing values, or wrong values are not shown in the data values because the preprocessing was performed based on the collected unstructured data. In particular, filtering was conducted to extract highly important words in sentences or important words that could be sentiment words. In the United States, in cases where sentiment words for English grammar are extracted, the words can be easily preprocessed utilizing the SWN(SentiWordNet). The SWN here is a vocabulary dictionary that contains sentiment words applied with English grammar and was constructed by giving sentiment values of meaningful affirmation and negation for sentiment classification and opinion mining analysis, which is utilized by at least 300 research groups throughout the world. A characteristic of the SWN is that it contains synonym sets based on 147,278 English words belonging to those parts of speech such as nouns, verbs, adjectives, and adverbs and that it is stipulated that the sum of the sentiment values of affirmation, negation, and neutrality in each set should be 1. That is, the SWN is specialized for extraction of sentiment words from sentences in English grammar because of the structure of its grammar algorithm and the structure makes it difficult to sentiment words from Korean sentences and becomes a factor to degrade prediction accuracy. Therefore, in this paper, the SWN was excluded and to efficiently derive sentiment words from sentences in Korean grammar, the rules as shown in <Table 1> were applied to perform word filtering.

Table 1: Word Stemming Filter Process

* **Filtering rules applied for the construction of a sentiment dictionary**
1. Remove special characters, English words, and disused vocabularies
2. Remove meaningless terms and one-letter texts
3. Separate the same words in the form of a conjunction to classify the natures of sentiment words
4. Distinguish homonyms and synonyms from each other
5. In the case of abbreviations and newly coined words, reflect only those that have been registered in WiKipedia or a Korean language dictionary

3.4 Design of Korean Grammar based Decision-Making Support System

For efficient opinion mining analysis in Korean grammar, those words that should be considered for decision-making support should be understood. In Korean grammar, there are words that sound identically but are different in meanings. Such words are called homonyms and a model is necessary for classification of words that happened to have the same sounds but are used with completely different meanings. The following are examples of such Korean words.

Table 2: Examples of Homonyms in Korean

Homonym	Meaning
Send	- It is a matter of strength. [힘이 부치는 일이다.] - I'll write you a letter. [편지를 부친다.] - We plant fields. [논밭을 부친다.] - Posting on Arbor Day [식목일에 부치는 글.] - An agenda for a meeting. [회의에 부치는 안건.] - Manuscript for printing. [인쇄에 부치는 원고]
Attach	- Stamp a stamp. [우표를 붙인다.] - Fire it. [불을 붙인다.] - Attach a watchdog. [감시원을 붙인다.] - Attach a condition. [조건을 붙인다.] - Nickname. [별명을 붙인다.]

In addition, there are words that sound differently but are the same in meanings and such words are called synonyms. The following are examples of synonyms.

Table 3: Examples of Synonyms in Korean

Synonym	Meaning
Smart Phone	- Smart Phone [휴대폰] - Cell Phone [휴대전화] - Hand Phone [핸드폰] - Mobile Phone [모바일폰]
Ship	- Marine [배]

Such a grammar consists of words that are used for analyzing opinion mining, and it is necessary to perform table classification on semantics and store the results in the database. That is, although "배" [boat] has the same meaning as "선박"[ship] that moves the water, there are "배" [pear] that means a fruit and "배" [stomach] that means a part of the body.

The following is the application of antonym rules. Typically, reputation analysis synthesizes the numbers of positive words and negative words in each sentence into means to analyze reputation with positivity and negativity in sentences. However, there are sentences where there are positive words followed by antonyms including concluding vocabularies that make the sentences into negative sentences. The analysis of these sentences reveals that the word rise in the overall context has a positive meaning and the word drop has a negative meaning but there are elements of antonyms that make the word rise show negative future forecast in the context of the flow of the entire sentence. In general, there are cases where in the process of deriving a sentiment dictionary, affirmative words are applied as they are without considering the irony of the following words. In this paper, positive words are considered as positive vocabularies in terms of the entire sentence; provided that, those words that are expected to be converted into negative vocabularies are filtered to construct a separate data dictionary. Finally, sets of associated words that correspond to the links of sentences are constructed. If the sentiments of opinions are judged in proportion to the appearance of positive or negative words in sentences, the accuracy will not be high enough. In this paper, under the assumption that there may be cases where even if a sentence has more positive words, the relevant sentence is a negative sentence when the meaning of the entire sentence is interpreted such as the case of an irony, word association rules are applied to reconstruct the sentiment dictionary.

3.5 Sentiment Word Tagging and Opinion Mining Analysis

Sentiment word taggings were classified into four types; positive, negative, neutral, and other taggings and the sentences in each document were compared to the data of word set groups in the sentiment dictionary to calculate the frequencies of positive and negative words. In this process, the values of the opinion data that had been divided by month and stored in each category were used to conduct time series analysis. The sentiment words were classified as follows.

November-December 2019
ISSN: 0193-4120 Page No. 2480 - 2485

NO	Word	Type
1	올바른	positive
2	개선...	positive
3	희망	positive
4	행복	positive
5	우수한	positive
6	효과적	positive
7	존중...	positive
8	좋다	positive
9	사랑...	positive
10	배려...	positive

Figure 3: Sensitive Word Classification

4. EXPERIMENTAL RESULTS

In this paper, to analyze the performance of Korean style software education, massive data were stored and accumulated and related preprocessing was performed. In addition, to improve the accuracy of the analysis, the decision-making support design technique for Korean grammar was used and the resultant antonym correspondence, homonym and synonym rule classification, and associated word decomposition were performed. Thereafter, the existing results of analysis of texts for software education and the results of reputation analysis conducted through the decision-making support system designed in this paper were compared for accuracy.

The environment of comparative analysis is as follows. First, in order to increase the accuracy of software education related text words, 41,000 words that appeared regularly at high frequencies were extracted by applying the related rules, and some of the nouns maintaining a close relation with the positive units among the extracted words were included in the important words. Second, all vocabularies recognized as one letter were excluded in the filtering process, and verbs and antonyms with high frequencies of association between words were extracted. Third, because of the characteristics of the online news articles that were written by a professional reporter's, which were not at the level of general comments, no slang or unspecified term appeared. However, numbers, special symbols, and abbreviations with the attributes of Internet language were removed for the accuracy of data. For instance, menbung in the sentence titled "Mengbung," is an abbreviation of mental collapse, and it is a newly coined word formally registered in the Wikipedia and the knowledge encyclopedia meaning severe shocks leading to mental collapse. However, this word was removed not because it was judged to be a irregular word but it was regarded as a word that cannot be identified with the use of the part-of-speech, that is, as a word consisting of noun(N) plus noun(N).

words	t_count	d_count	status
멘붕	2	1	BAD

Figure 4: Example of Inappropriate Newly Coined Words

With regard to the performance analysis, the results of reputation analysis using the existing sentiment words were compared with the results of reputation analysis using the design technique presented in this paper. According to the results, the accuracy of the existing technique was shown to be 78% in the case of positive word and 82% in the case of negative words. The accuracy of the negative word was 78% Accuracy was 82% accuracy. The accuracy of reputation analysis not lower than 80% corresponds to stable values. The accuracy of reputation analysis that introduced the design technique presented in this paper was shown to be 82% in the case of positive words with an improvement by 4% compared to the existing model and 83% in the case of negative words with an improvement by 1% compared to the existing model.

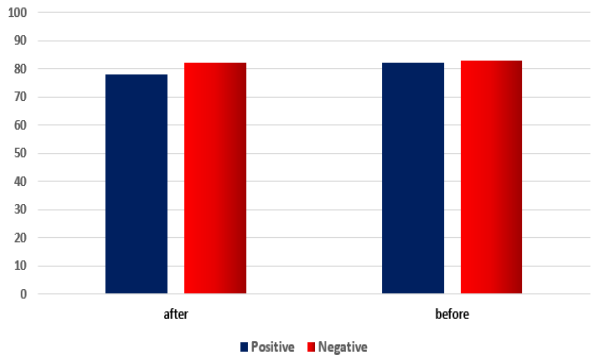

Figure 5: Performance Analysis

5. CONCLUSION

In this paper, a decision making support platform was designed for analysis of the performance of Korean style software education [16-17]. If the proposed big data analysis platform is used, change factors for diverse education environments can be analyzed and future-oriented agendas for creative talent cultivation and coding education can be discovered. Learner education systems can be improved so that excellent creativity with good problem solving ability to solutions based on new and unique perspectives can be discovered and present ways to suggest diverse solutions for similar problems and if the data in the field of education that have been already accumulated, ways to implement education suitable for students can be presented. In addition, to compare with previous analysis methods, whereas methods such as brainstorming, Delphi, and expert panels were used in the past, this paper enabled the presentation of data based analysis results through big data decision making support platform using research literature, media articles, and related

November-December 2019
ISSN: 0193-4120 Page No. 2480 - 2485

reports. Therefore, the results of this paper are expected to be widely utilizable in the field of convergence in the domain of humanities too.

REFERENCES

(Periodical style)

1. Chen, M., Mao, S., Liu, Y.: **Big data: a survey**. Mob. Netw. Appl. 19, 2014, pp. 171–209.
2. Xindong Wu, Xingquan Zhu, Gong-Qing Wu and Wei Ding, **Data Mining with Big Data**, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING , Vol.26, No.1, 2014, pp. 97-107.
3. Tang, C. and C. Liu.2008. **Method of Chinese grammar rules automatically access based on association rules**, In: Proceedings of the. Computer Science and Computational Technology volume, 1 pp. 265-268 ISCSCT, Shanghai, Dec. 20-22, 2008.
4. Irfan Ajmal Khan, Jin Tak Choi, **An Application of Educational Data Mining (EDM) Technique for Scholarship Prediction**, International Journal of Software Engineering and Its Applications, Vol. 8, No. 12, 2014, pp. 31-42
5. Xu, Yue, Li, Yuefeng, & Shaw, Gavin, **Reliable representations for association rules,** Data & Knowledge Engineering, Volume 70 Issue 6, 2011, pp. 555-575.
6. Bo pang, Lillian Lee and Shivakumar Vaithyanathan, **Thumbs up?: sentiment classification using machine learning techniques**, Proceedings of the ACL-02 Conference on Empirical methods in Natural Language Processing, Vol.10, 2002, pp.79-86.
7. Berman, J.J., **Principles of Big Data Preparing, Sharing, and Analyzing Complex Information**, Elsevier, Waltham, 2013.
8. P. Allen, S. Higgins, P. McRaie, H. Schlaman, **Service orientation: winning strategy and best practices**, Cambridge University Press, New York, NY, 2006.
9. Klein, D., Tran-Gia, P., Hartmann, M., **Big data**, Informatik-Spektrum 36, 2013, pp. 319–323.
10. Grünwald, M., Taubner, D., **Business intelligence**, Informatik-Spektrum 32, 2009, pp. 398–403.
11. Rupali Bhardwaj and Sonia Vatta, **Implementation of ID3 Algorithm**, International Journal of Advanced Research in Computer Science and Software Engineering(IJARCSSE), Vol.3, Issue.6, pp.845-851, June, 2013.
12. Jihoon Seo, Eunmi Cho and Kilhong Joo, **Analysis of agenda prediction according to big data based creative education performance factors**, Lecture Notes in Electrical Engineering, vol. 474, pp. 1876-1100, 2018.
13. E. Courses and T, Surveys, **Using Sentiment SentiWordNet for multilingual sentiment analysis**, IEEE 24th International Conference on Data Engineering Workshop, Cancun, Mexico, 2008, pp.507-512.
14. Stonebraker, M., **SQL Databases v. NoSQL Databases**, Communications of the ACM, Vol.53, No.4, 2010, pp.10-11.
15. Daas, D., Hurkmans, T., Overbeek, S., Bouwman, H., **Developing a decision support system for business model design. Electron**, Mark. 23, 2012, pp.251–265.
16. Ślusarczyk, B., Haseeb, M., & Hussain, H. I. (2019). **Fourth industrial revolution: a way forward to attain better performance in the textile industry**. Engineering Management in Production and Services, 11(2), 52-69.
17. Prakash, G., Darbandi, M., Gafar, N., Jabarullah, N.H., & Jalali, M.R. (2019) **A New Design of 2-Bit Universal Shift Register Using Rotated Majority Gate Based on Quantum-Dot Cellular Automata Technology**, International Journal of Theoretical Physics, https://doi.org/10.1007/s10773-019-04181-w.