「지방정부연구」 제19권 제1호(2015 봄): 75-94
The Korean Journal of Local Government Studies, Vol.19 No.1 (2015 Spring)

Search Data Analysis and Its Application to Social Science Research and Implications
: A Case Study on Living Conditions of Single Youth Households*

Kim Do Hee

Abstract

This study attempts to analyze living conditions of single youth households by using the Internet search data which is one of typical examples of big data of today. Also, this study provides a practical methodology for using the Internet search data as a new data source for analyzing a social phenomenon. More specifically, we determine research variables relevant to the study of single youth households through correlation analysis among the Internet search trends. By doing so, we demonstrate the effectiveness of the proposed methodology in terms of promptness and cost required for analyzing a social phenomenon, compared to the conventional survey methodologies. The importance of this research lies in the fact that a methodology for collection and interpretation of big data such as the Internet search data to analyze an actual social problem has been demonstrated, whereas most of the previous research focused on illustrating the potential of big data in the domain of social science. Accordingly, the proposed methodology can be utilized for more demand-oriented policy making, customized policies according to age groups, and timely analysis of effectiveness of a policy after enforcement thereof.

Key Words: search data analysis, single-youth household, living conditions

Ⅰ. Introduction

Obama's success in the 2012 presidential campaign is often referred to as one of the most famous moments in the application of big data, which has employed micro-targeting based on a huge amount of voter profiles. Scott Vandenplas, then the DevOps director at Obama camp, described the campaign on the Twitter: *4Gb/s, 10k req/s, 2k nodes, 3 datacenters, 180TB and 8.5*

* 이 논문은 2014년도 울산대학교 교내연구비 지원에 의하여 연구되었음.

** 울산대학교 정책대학원

billion req. Design, deploy, dismantle in 583 days to elect the President[1]. Big data usually refers to data sets so large and complex that they become difficult to work with traditional statistical software or data processing applications (Snijders & Matzat & Reips, 2012). The rise of big data is attributed to such things as advancement of information technology, higher computing performance at a lower price, domination of social networks, and widespread use of smart devices (Lim, Yong-Jae *et al*., 2013). Many disparate domains are expected to benefit from big data analysis and decisions made therefrom, examples of which include health care, public sector administration, retail, manufacturing, and personal location data (McKinsey, 2011).

Recently, in Korea, many researchers have paid great attention to applications of big data. Lee has conducted significant research on the various aspects of big data as a means to implement Government 3.0 (Lee, 2013), pointing out the role of big data in the process of developing government policies. Lee and Yoon described how keyword spotting can be used to extract people's interest with respect to a particular sightseeing place and proposed how advertisement or web-hosting strategy can benefit from the search result (Lee & Yoon, 2014). Kim investigated how utilization of public data could be enhanced, proposing three tasks to be performed: development of information services based on big data, building a master data management system, and improving reliability on e-government (Kim, Hyung-Sung, 2012).

However, as indicated in the work of Moon and Cho (Moon & Cho, 2012), big data is still an ambiguous concept in a sense that it is difficult to determine which technique is suitable for which step in a particular problem. In fact, most of previous works only describe potential applications and capability of big data, along with proposals for a strategic use of big data and policy making principles (Kim & Trimi & Chung, 2014; O'Malley, Martin, 2014). Even though a few of the research works articulate actual analysis methods used and steps of how their problems are solved based on the use of big data, there are in fact very little efforts to provide techniques or methods that can be actually referenced. This fact then indicates the very initial stage of the use of big data to social studies; while, at the same time, it indicates that the success of big data application highly depends on a specific goal and problem to be solved.

Utilization of big data for local administration is in the much the same situation as described above. In fact, there are only a few research works published (Noh, 2014; Kim, Jongphyo, 2013; Lee & Park & Jeong, 2014), though many local governments are trying to develop effective usage of big data to their needs.

As summarized in the work of Noh (Noh, 2014), where utilization examples of big data for local administration are reported, it can be regarded that utilization of big data for local administration or strategic policy development is at its early stage. Based on the literature survey,

[1] Morales, A. Weber. (2013). DevOps teams are on the move. SD Times, March 29th.

Noh categorized four main utilization areas of big data for local administration: monitoring, prediction, policy making, and development of customized administrative services. The research provides directions of strategic utilization of big data only at an abstract level; thus, practical insights about how big data can be obtained and utilized still remain obscure.

In a more specific example of big data application to local administration, Kim studied a crime prevention system in Busan area (Kim, Jongpyo, 2013) proposing three steps of implementation: community mapping of crime, crime map building based on public data, and big data based crime prevention system. Lee et al. has studied accessibility changes in the metropolitan Seoul subway system based on the analysis of traffic card transaction databases (Lee & Park & Jeong, 2014). In the application of big data to tourist preference analysis, communication cell-based transient population information, credit card usage databases, and public data of Jeju have been utilized (Lee, Eunjoo, 2015). As can be seen from these works, access to commercial and/or public databases is of paramount importance for problem solving based on big data[2]. Depending on situations, such databases may or may not be available for individual social science studies. Also, there is a crucial issue about selection of particular databases; whether a particular set of databases (in fact, research variables) is sufficient to describe the corresponding social phenomenon is still unanswered.

With these findings in mind, this research tries to set up a data collection method to construct a research framework. More specifically, this research relies on the author's previous research work (Kim, Dohee, 2013) for a theoretical framework but attempts to construct a methodology to find meaningful research variables from the search volume provided by commercial search services such as NAVER and Google trends. Particular search terms are eventually used as research variables if the search volume of the corresponding term exhibits non-trivial correlation with the increase of single youth households.

In the previous work, the author analyzed the living conditions of single youth households by using traditional data coming from government organizations. The data are traditional in a sense that it takes significant resources to finally obtain the data, comprising face-to-face interviews with people, survey through telephone, and basic literature survey. Moreover, some of the data, for example, the population consensus conducted every five years, show a considerable temporal lag to be useful as an estimate of the current mind shaping of the general public and as an up-to-date reference for policy making.

To summarize, the conventional methods employed for determining research variables can be characterized by their expensive costs, non-real time analysis, and inflexibility to change sample

[2] Another case study of utilizing big data for local administration in Seoul can be found at Local Information Magazine, 91(3/4):14-19, 2015. In this case, personal communication database, population database, income data, digital techno-graph (DTG) data of about 70,000 taxis, traffic data, and so on have been actively utilized to provide disparate administrative services.

sets (for example, the size, geographic variety, or demographic variety of the sample set cannot be easily changed). In this sense, search data from the Internet can be a great resource to compensate the weakness of the conventional methods in terms of inexpensiveness, promptness, and flexibility to add or modify sample sets. Therefore, in this research, we examine if it is possible to derive research variables for a proposed research framework directly or indirectly from search data analysis. As indicated above, it is important to set up a research goal to systematically search for data suitable and to proceed with subsequent steps leading to decisions with respect to the current problem (Lee, 2013).

It should be noted that this research deals with the initial step of the whole data value chain as proposed in the work of Miller (Miller & Mork, 2013). Although this research focuses on developing a methodology for data collection to construct a research framework, it is our belief that the data collection step is crucial for subsequent analysis steps of a current social problem. This is so partly because big data analytics usually employs traditional tools coming from statistics, data mining, pattern recognition, and so on, all of which have their own theoretical foundation for data representation and interpretation. On the other hand, collection of big data sets and extraction of meaningful variables from big data is relatively new, which requires further research and evidence for wider application to real-world problems.

The paper is organized as follows. In Section II, we show that big data can be used to derive analysis variables for a proposed research framework. We employ our previous work on single youth households and describe how the state-of-the-art search engines can be utilized to this purpose. Section III describes an in-depth analysis of living conditions of single youth households with respect to the search volume of corresponding terms discovered from the search data analysis. Finally, we discuss how the proposed method can be used for social studies and policy making in a more general setting.

Ⅱ. Theoretical Backgrounds

1. Usage of the Internet Search Data: A Review of Query Validation and Modeling Methods

A big data framework employing the Internet search inquiries is usually constructed via two steps: search query validation and system modeling.

The first step of search query validation is needed to filter out irrelevant search data and to increase prediction performance of a found search inquiry as an indicator. Meaningful search queries (or indicators to a current problem) can be found from the validation step if the problem can be described by a limited number of specific term(s), as compared to a social problem comprising a plurality of economical and/or social issues.

In that sense, Google flu trends can be regarded as an excellent example showing how search keywords can be chosen (Ginsberg et al., 2009). In the case of Google flu trends, a relationship between an influenza-like illness (ILI) physician visit and and ILI-related search query is modeled first. Then 50 million candidate queries are separately tested against the model to see goodness-of-fit based on the respective queries. For each search query, linear regression with 4-fold cross validation was performed to fit the model and correlation coefficient was calculated. Finally, top 45 queries in the order of goodness-of-fit were selected. It should be noted that the top 45 queries consist of various keyword expressions associated with the influenza (for example, flu remedy, symptoms, complication, and so on).

Similarly, in the case of tuberculosis (TB) surveillance based on Google trends (Zhou, Ye, & Feng, 2011), 19 TB-related terms were chosen manually by the authors based on their domain knowledge. Further investigation of performance of the respective terms was not performed. Instead, a set of linear equations was built to predict the future behavior of TB cases, where a classic iterative process of the Kalman filter was employed for the prediction.

Also, in the case of private consumption forecast (Vosen & Schmidt, 2011), instead of using individual search queries, authors chose 56 consumption-relevant categories from a total of 605 categories provided by the Google trends. They used the search volume generated from each chosen category as a data source for developing a forecast model. The forecast model was built around an autoregressive model augmented with a set of macroeconomic variables determined in an arbitrary manner based on the authors' expert knowledge.

As can be noticed, methods for selecting search queries and modeling data behavior vary according to the characteristics of a current problem. Moreover, search volumes are often distorted by the so-called media effect. However, since ideas or thoughts formed within the general public or a target group can be captured promptly through a proper procedure of using search data, search data analysis can be useful to instantly determine the validity of chosen research variables, which may not be easily done through traditional methods employed for social studies. Also, it should be noted that compared with a large number of big data applications in the engineering fields (HU *et al*., 2014; Jardak & Mhnen & RiihiJrvi, 2014; Ferreira *et al*., 2013; Srinivasan & Arunasalam, 2013), the adoption of big data into social studies is relatively new and slow. It is partly so because social studies usually have to deal with a variety of social behaviors that are difficult to model into mathematical forms, different from engineering problems dominated by more structured and formal representation of data.

2. A Proposed Method for Validation of Search Queries

This research aims to describe our approach to analyze social phenomena and to demonstrate how policy design can benefit from analysis of big data such as the Internet search volume. Social studies often depend on surveys published at large intervals. The population consensus is one typical example of this kind, which is published every five years. Therefore, it is often the case that expert insights and domain knowledge about a particular subject are required to interpolate the gaps between survey intervals and eventually develop an appropriate research framework. For example, to build a research framework on which to analyze living conditions of single youth households, unemployment rate, late marriage, residential moving due to education, Internet utilization rate, and others were chosen to reflect the living conditions of single youth households (Kim, Dohee, 2013). An extensive literature survey was conducted to get the initial understanding of the social phenomenon, including census data provided by National Statistics Office of Korea.

Contrary to the conventional approach, more recent search tools such as social METRICS, NAVER trends, and Google trends can be used to determine indicators of a particular social phenomenon, which is the increase of the number of single youth households in this research.

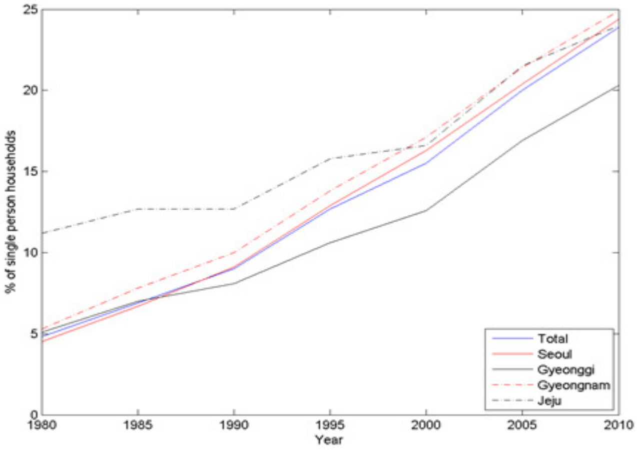
1) Keyword selection

A traditional approach to extract opinions of people in their twenties and thirties would be an interview or personalized survey. Another approach is a literature survey to extract common keywords, which is commonly found in the previous research works. Those surveys are still a useful means for social studies since an initial set of meaningful keywords can be found from the survey and can be used to narrow down the scope of search queries to be evaluated.

In this study, we first examine the demographic data to see the distribution of single person households in Korea. From the demographic data, we determined a majority age group and obtained a population change of the corresponding age group as a statistical pattern to be compared with that of search volume of a particular query.

Figure 1 shows variations of the percentage of single person households, recompiled from the population consensus data published by the National Statistics Office of Korea. From the figure, one can easily notice the gradual increase of the number of single person households. Table 1 further details the ratio of single person households according to age groups, which shows that young people in their 20s and 30s form the majority of the single person households.

〈Fig. 1〉 Variations of percentage of single person households[a].



[a] Recompiled from the population census data published by the National Statistics Office of Korea.

〈Table 1〉 The ratio of single person households according to age groups in metropolitan cities of Korea (re-organized from the population consensus published by the National Statistics Office of Korea for the year of 2000, 2005, and 2010).

City	10s	20-30s	40s	50s	60s or more
Seoul	1.20	54.03	14.87	10.43	19.40
Busan	1.27	32.77	16.60	16.20	33.17
Daegu	1.67	39.57	15.73	13.43	29.60
Incheon	1.13	44.67	17.90	12.30	23.97
Gwangju	2.60	47.00	13.13	11.00	26.27
Daejeon	3.20	51.70	14.40	11.13	19.60
Ulsan	0.70	43.93	18.67	14.00	22.77

Then, we conducted the keyword selection by combining the social METRICS and Google trends. In this research, a group of single youth household-related terms were chosen from a previous research work (Kim, Dohee, 2013). The social METRICS search data was then used to examine the initial, general distribution of interests of people embedded in the search keywords against the terms listed in Table 2.

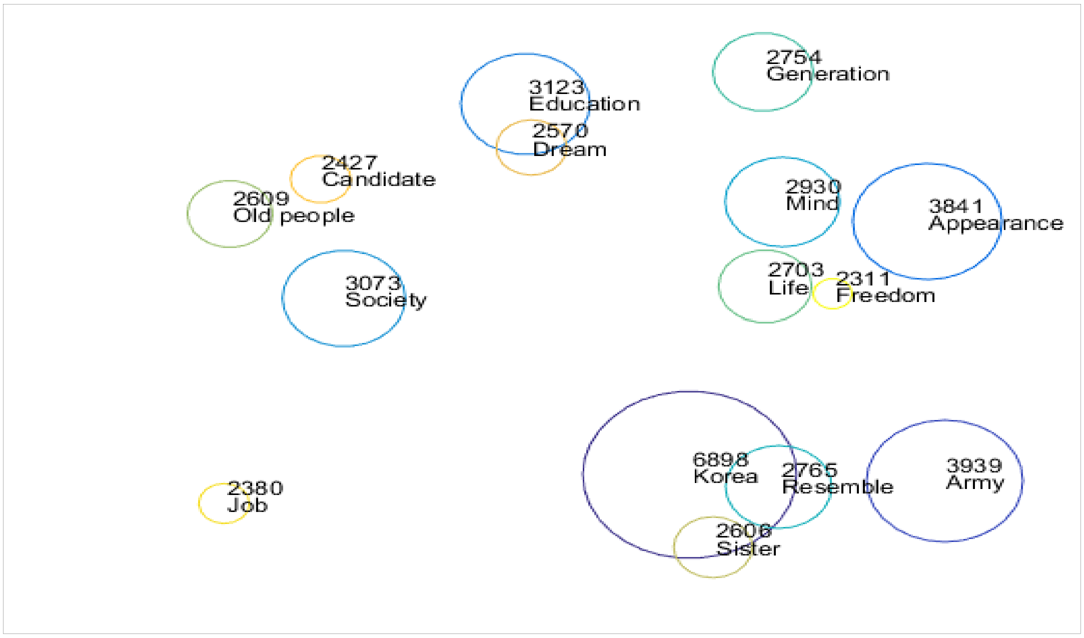
〈Table 2〉 The initial set of terms used for evaluating relevance of search volume

Classification	Search Terms
Economical aspect	Employment, unemployment, job, career
Socio-cultural aspect	Marriage, late-marriage, unmarriage, singleness
Educational aspect	University, education, higher education

To determine the factors which cause the increase of the number of single youth households, one may resort to search keywords that are closely linked to the lifestyle or living conditions of young people. As done in the work of Zhou, Ye, and Feng (Zhou, Ye & Feng, 2011), if the subject in question is a technical matter (which is tuberculosis in that case), search keywords may be determined relatively easily. Since living conditions of young people for a particular time period are related to the value system, spirit of the times, and so on, we employed the social METRICS which returns a list of keywords with a non-trivial search volume related to a given inquiry. In this research, we chose *youth* as the search inquiry. The social METRICS utilizes search inquiries entered only during the last one month. Thus, the order of resulting keywords may be distorted by recent, sensational events which in fact are not closely related to the living conditions.

Figure 2 visualizes search volume for each associated keyword with respect to the search inquiry of youth. Each number for a circle represents the number of occurrences of the corresponding keyword found in blogs and twits. If the search period can be extended to several years from the default period of one month, accuracy of estimating the opinion of young people can be improved significantly. Since it is not the case at this time, we need to further examine relevancy of the respective keywords obtained. Moreover, it is often the case that a significant part of the search result may contain irrelevant topic, showing only the relevance in terms of literal similarity or matching.

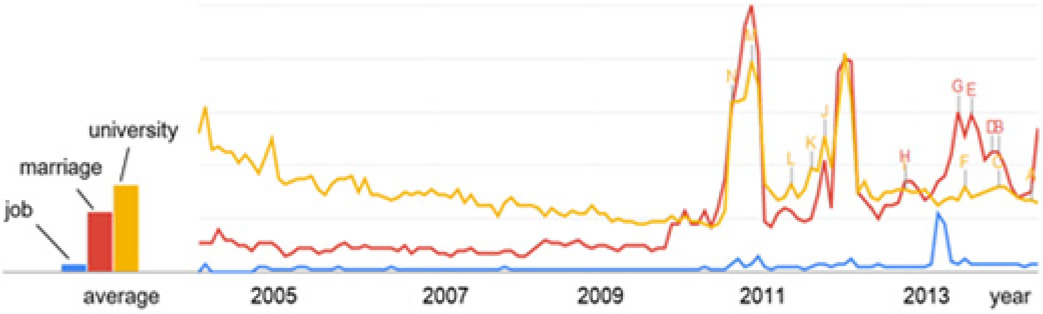
〈Fig. 2〉 Visualization of search volume for each associated keyword with respect to the inquiry of youth[b].



[b] Each number for a circle represents the number of occurrences of the corresponding keyword found in blogs and twits (re-organized from the search result provided by Social METRICS).

In this regard, Google trends is more relevant to the purpose of keyword spotting except that the result from Google trends contains a search result comprising the whole age groups. Google trends is based on the search volume with respect to news articles since 2004; therefore, one can regard that Google trends provides time series data spanning ten years, thus supporting more reliable analysis of people's opinion. Figure 3 shows search volume of three particular inquires chosen from the result of social METRICS, including job, marriage, and university among other things. By examining the search pattern, one can conclude that there is a significant resemblance between the Google trends and the growth pattern of the population census data of Fig. 2. It should be noted that the search trends of Fig. 3 is a result obtained from a search conducted in Korea.

<Fig. 3> Search trends for inquiries of job, marriage, and university provided by Google trends.



2) Implications of keyword selection using social monitoring tools

As described above with reference to Figs. 1 to 3, the social monitoring tools provided freely from Internet search service providers are valuable in a sense that they provide a quick overview of current trends or opinions of people without invoking additional expenses or resource investment. For example, social METRICS gives an instant monitoring result with respect to a target keyword: associated keywords, sentiment analysis result, and variation along timeline. The Google trends gives a similar outlook on a particular subject. Therefore, if they are used as tools for examining social trends at the initial stage of research or as tools for confirming traditional research findings, they can play a significant role as a new means to facilitate the social study. What is more, the social monitoring tools can be used as a probe to immediately find out public opinion, social trends, and so on, which is useful for decision making for business or government policy design.

In what follows, based on the initial findings from the social monitoring tools, we proceed with analyzing significance of individual search inquiries on the living conditions of single youth households.

Ⅲ. Analysis of living conditions of single youth households using search data

Ginsberg and others (Ginsberg et al., 2009) demonstrated that there is a remarkably similar pattern between U.S. Centers for Disease Control and Prevention (CDC) and a search pattern obtained from a Google search for flu or influenza. This kind of findings obtained from analysis of search data reveals that there is an association between what occupy people's minds and observed social phenomena captured through keyword spotting through Google search, for example. Based on this insight, this research further demonstrates that key research variables can be properly selected and confirmed from analysis of search data. In this study, we further employed a search trends service provided from NAVER since the corresponding search service is deemed to be more adapted to Korea as can be evidenced from its high market share in Korea[3].

1. Living conditions of single youth households from an economic aspect

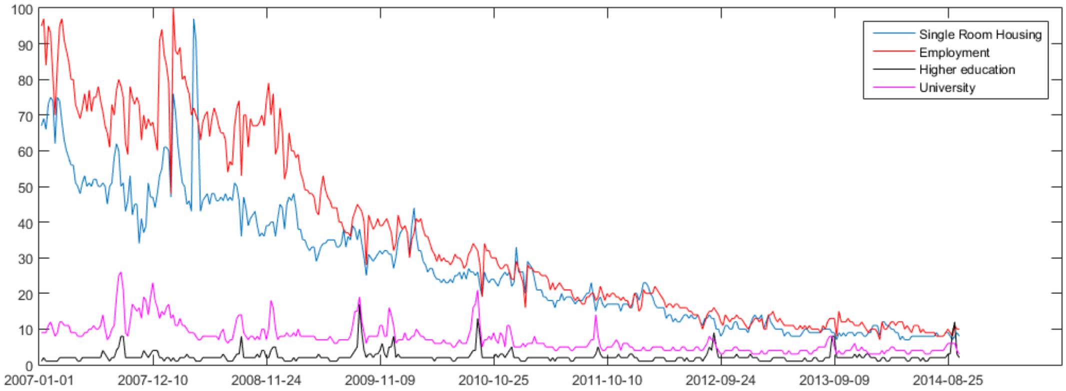
Based on the initial finding from the social METRICS (see Fig. 2), we calculated correlation between the type of single youth household and individual search keywords. As indicated in Table 2, households consisting of young people in their 20s and 30s are of primary concern in this study; therefore, we examined the percentage of individual housing types popular for the young people. From a recent survey of housing types, it was found that the single room housing is the most popular household type for the young people[4].

Next, we examined the NAVER search trends with a search keyword of single room housing (which is called *one room* in Korea). Differently from the Google trends, the NAVER trends service provides search keyword trends from the year of 2007. The search terms of Table 2 were used to generate the trends data of Fig. 4. As the figure shows, one can notice there is a good similarity between the search trend of single room housing and that of employment. We calculated the Pearson correlation coefficient and obtained 0.9521, which is a quite high value and indicates that the rise or fall of the number of single youth households is closely related to the employment status of young people. However, on the contrary to our expectation, the correlation between the single room housing and higher education (0.0085), university (0.6750), marriage (not shown), and other variables was found to be very small, from which we may conclude that the social behavior due to the single youth households is not significantly related to education and marriage.

[3] As of Aug. 25, 2014, the Korea market share of the Internet search service by NAVER is 76.69% while that of Google amounts only to 2.06% (news.einfomax.co.kr/news/articleView.html?idxno=118339).

[4] As of Jan. 7, 2014, the single room housing occupies 43%, followed by apartment (20%), studio (16%), multiplex housing (11%), and others (www.newswire.co.kr/newsRead.php?no=731302).

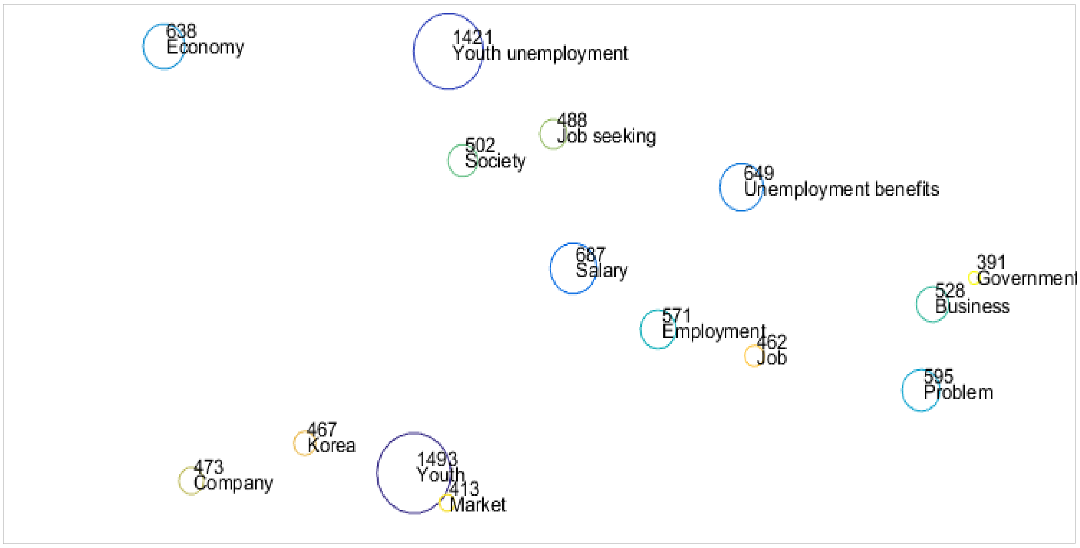
〈Fig. 4〉 A comparison of NAVER trend patterns[c].



[c] The NAVER trend patterns are related to search queries of employment, higher education, and university against the single room housing. The vertical axis shows the number of hits from the keyword search, which is normalized to 100 by the largest search volume. The horizontal axis represents the search period.

Next, we examined public sentiment due to unemployment to check whether young people in their 20s and 30s are actually influenced by the issues of employment. If so, we can determine with more confidence that employment is in fact one of the indicators closely related to and describing the pattern of single youth households. To this purpose, we employed the social METRICS which is more into the examination of popular sentiment and the social monitoring.

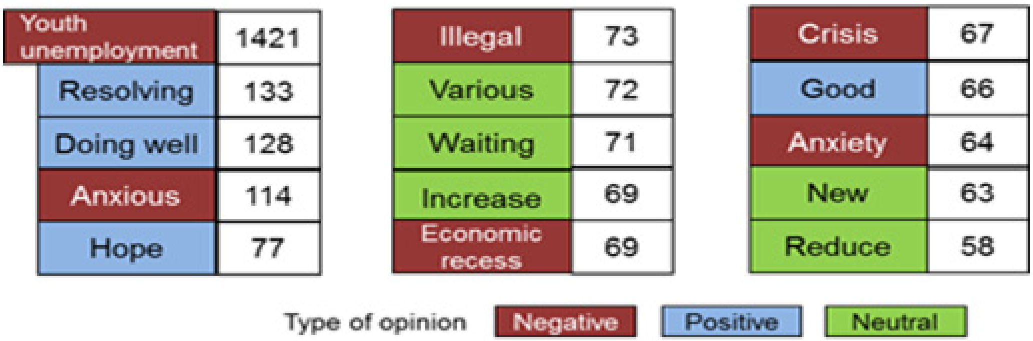
〈Fig. 5〉 Keyword map obtained from the search with respect to the query of unemployment[d].



[d] This is a re-organized visualization of the original social METRICS output. The circle radius represents the corresponding search volume, where the number of occurrences of the corresponding keyword is shown together.

The social METRICS now reinforces our findings with respect to a relationship between unemployment rate and people's social activity. According to the analysis result of social media comprising blogs and twitters with respect to unemployment, youth is ranked in the first place, followed by youth unemployment, salary, unemployment allowances, economy, employment, and others (see Fig. 5). Moreover, the sentiment analysis result provided by the social METRICS reveals that employment issue gives rise to unhappy feelings among the population; negative responses consisting of youth unemployment, anxiety, illegal, economic recession, crisis, and others are found to form the minds of the public, particularly for the youth (see Fig. 6).

〈Fig. 6〉 Sentiment analysis with respect to the keyword map of Fig. 5[e)]



🟥 Youth unemployment	1421
🟦 Resolving	133
🟦 Doing well	128
🟥 Anxious	114
🟦 Hope	77

🟥 Illegal	73
🟩 Various	72
🟩 Waiting	71
🟩 Increase	69
🟥 Economic recess	69

🟥 Crisis	67
🟦 Good	66
🟥 Anxiety	64
🟩 New	63
🟩 Reduce	58

Type of opinion 🟥 Negative 🟦 Positive 🟩 Neutral

[e)] The number shown in the neighboring right cell of each keyword represents the corresponding search volume. As the figure shows, people's opinion is formed in a negative way.

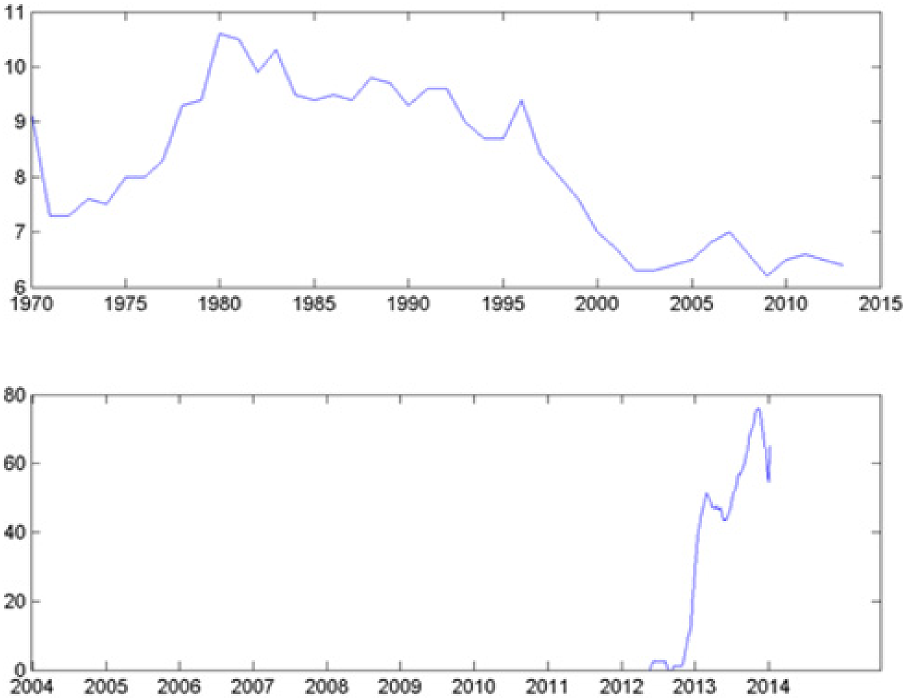
2. Living conditions of single youth households from a socio-cultural aspect

Though in the previous search data analysis, marriage-related queries revealed weak relevance to the behavior of single youth households pattern, we further examined the relationship between the two because young people in their 20s and 30s are at their best age for marriage and there might be still a relationship between the two in some manner.

According to the author's previous work, the median age at first marriage is getting older, making single youth households spread more firmly. Based on a recent data provided by the National Statistics Office of Korea, the crude marriage rate shows an obvious tendency of people's attitude toward marriage, namely, the tendency of getting married at an older age (see Fig. 7). Though decrease of the crude marriage rate is a trend commonly found among most of OECD nations, the crude marriage rate in Korea exhibits the highest rate at 1980 and decreases gradually down to 6.4 at 2013. The crude marriage rate shows a considerable fall compared with the pre-2000s. Interestingly, this tendency can be confirmed through social activity monitoring based on Google trends. As can be seen in the bottom figure of Fig. 7, there is a virtually zero search result before the year of 2012. For the last few years, the keyword of single life returns a significant search result, coinciding with the social trends of increase of late marriages as measured by the National Statistics Office of Korea. Therefore, one may conclude that this aspect also reinforces the significance of social query data through search engines. Also, one may further conclude that the weak relationship between the search volume via the keyword of marriage and the pattern of single youth households mainly results from the virtually zero interest in the single life before 2010, providing a very small correlation coefficient (0.0425).

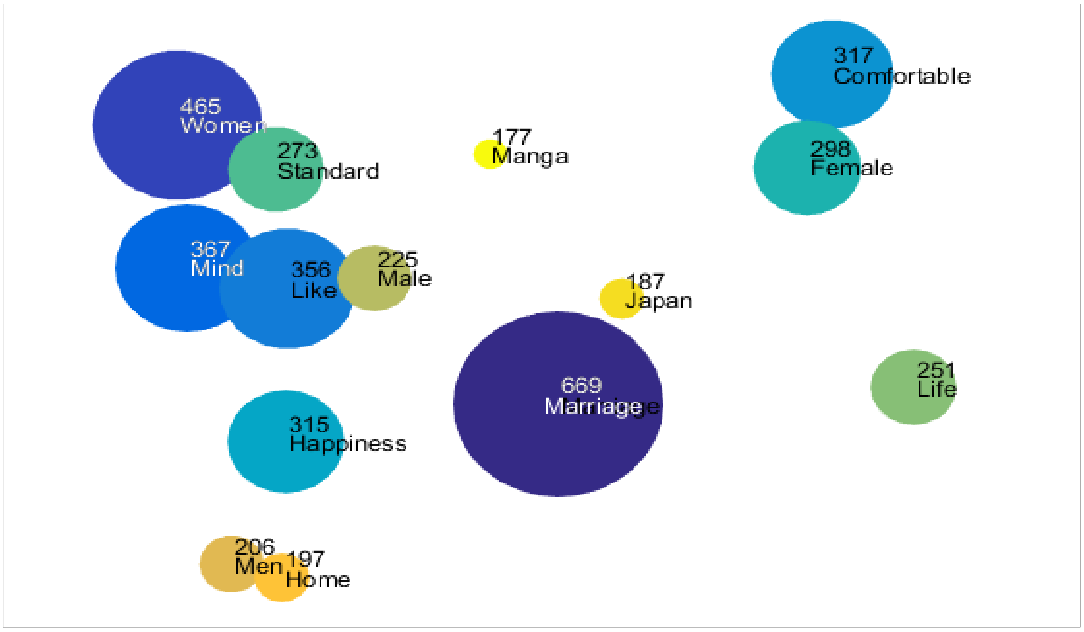
Again, similar to the analysis from the economic aspect, Fig. 8 plays as evidence to this social phenomenon; in other words, young people feel comfortable or happy with single life as seen from the sentiment data. It also gives a hint that the single youth household is more for women, though the tendency is getting clearer both for men and women.

〈Fig. 7〉 Top: Crude marriage rate in Korea (Source: http://www.index.go.kr). Bottom: Google trend pattern with respect to the keyword of single life[f].



[f] In the top image, the horizontal axis represents the year, and the vertical axis represents the crude marriage rate which is calculated as the number of marriages registered during a calendar year per 1,000 estimated resident population at July, 1st of the same year.

〈Fig. 8〉 Keyword map obtained from the search with respect to the query of single life.

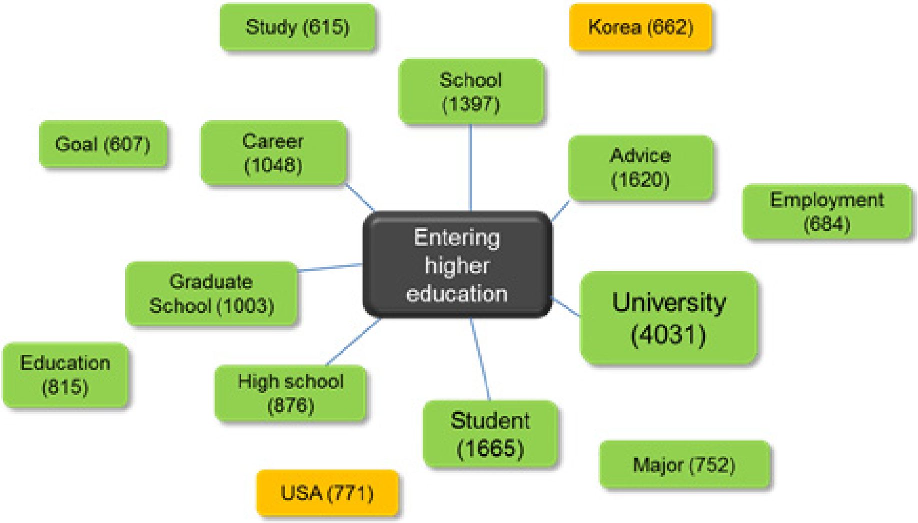


3. Living conditions of single youth households from an educational aspect

According to the survey conducted by the National Statistics Office of Korea in the year of 2010, 60% or more students aged 15 or more want to get a college education or higher, and 28.9% of them expect a higher education from graduate schools.

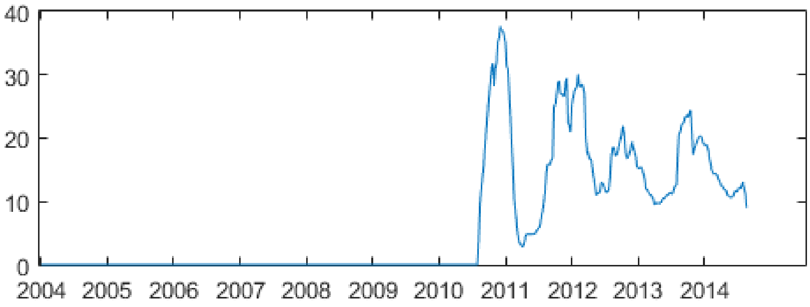
This tendency can also be confirmed by social METRICS and Google trends. In this case, we supplied a search inquiry of entering higher education. The search result from the keyword map reveals that university is ranked at the top position (4,031) of the number of search hits, followed by student (1665) (see Fig. 9). Also, Google trends show that the same keyword returns a growing number of search hits from since 2009 (see Fig. 10). This pattern is evidence that young people are getting more interested in studying at higher educational organizations, and again indicates that there would be a growing number of single youth households to finish their higher education courses. This case shows a search data pattern (virtually zero search volume before 2010) similar to that of single life in Section III.2, again explaining the weak relationship between the search volume via the keyword of higher education and the behavior of single youth households.

〈Fig. 9〉 Keyword map obtained from the search with respect to the query of entering higher education[9].



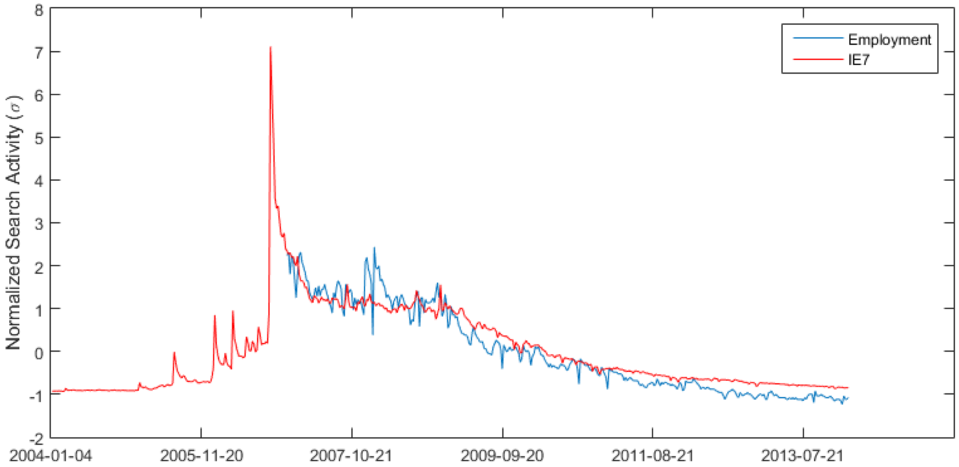
[9] Each box represents a key word associated with the given search query along with the search volume of the corresponding query.

〈Fig. 10〉 Google trend pattern with respect to the keyword of higher education[h].



[h] As the figure shows, the interest in the higher education is growing rapidly for the last few years. The yearly fluctuation may be regarded to be originating from seasonal factors (such as graduation or start of new school year).

〈Fig. 11〉 Google trend pattern showing a high correlation with the search pattern of employment provided by the NAVER trends service[i].



[i] In this particular case, the Pearson correlation coefficient is 0.9677 and can be anticipated from the significant resemblance between the two search patterns, though the search data before 2007 are missing in the NAVER trends service.

Ⅳ. Discussion and Concluding Remarks

We could determine that the single youth households are closely related to young people's employment status as explained in Section III.1. And this fact can be again confirmed by finding Google trends which exhibits high correlation with the search patterns of the keyword *employment*. The Google Correlate service provides a list of relevant Google trends with respect to a particular pattern of data provided by the user. In our case, it was the search pattern of the keyword employment and surprisingly, the resultant Google trends with high correlation coefficients were almost related to high-tech terms, including windows xp pro (0.9726), Microsoft activesync (0.9702), download (0.9678), web templates (0.9673), and the like. As can be noticed from Fig. 11, the close relationship between the employment and the single youth household can be further reinforced by the fact that those who are interested in the employment issue and prefer to live alone are also highly interested in the high-tech issues.

Though we found that the correlation between the single youth households and marriage or education-related issues is relatively small, they are still valuable research variables in a sense that demographic change and variation of the value system can be properly checked through a couple of social monitoring tools and trends services as described in Section III. 2 and 3.

In this study, we introduced a methodology using the Internet search data as a new means to social studies. As opposed to traditional approaches to social studies, big data analytics based on the search queries provides an instant and up-to-date social activity result. This is a great opportunity for a social study utilizing demographic data if one notices that the National Statistics Office of Korea conducts population census every five years.

In the previous work right before this study, we extracted research variables based on literature survey and analyzed the living conditions of single youth households based mostly on statistical data provided by the National Statistics Office of Korea. In that case, to extract appropriate variables, one needs to have a deep insight against social activities, and sometimes it is difficult to find meaningful variables, with which the whole research activities are conducted. On the other hand, if one wants to have confidence in the extracted research variables, consulting the big data, which is the search data in this case, is a new choice, and this study demonstrated the possibility of using the search data as a means to immediately probe into the people's opinion and interests.

There are drawbacks, however, in the search data analysis. Unless the source of search data is tightly linked to personal information, it is often difficult to reduce typical buzz observed in the Internet data. Since this research relies on the general keyword map and trend pattern from the general public, one should exert caution to select appropriate research variables.

Suppose the search data are properly linked to personal information. Even for that case, a carefully designed clustering, regression analysis or association analysis should be performed to find out relative importance of selected research variables. In this research, we did not carry out the big data analytics meant for making decisions since the original data are all coming from the unknown public, which limits the degree-of-freedom of applying analytics techniques. If search data are supported with personal profiles, subsequent data analysis and policy design can be performed relatively easily and with confidence. At the same time, this may cause a security issue in applying search data to social studies or policy making. In this sense, research on methods for mining user-generated contents is worthy of attention, which can be used for demographic prediction (for example, age and gender) of owners of generated contents. For example, if the demographic prediction data can be combined with the corresponding location information, local administrative services can be designed and/or monitored in an effective manner.

Now the importance of big data generated from social activities is getting more and more important. As expected from various sources of research reports related to big data, application range of big data is huge (Lee, 2013). So far, most of research works related to big data are limited to introduction to characteristics of big data and potential area thereof, only a few actually demonstrate the power of social monitoring tools to solve actual social problems. We demonstrated how search trends tools can be actually applied to social problems. To this end, we re-considered the living conditions of single youth households and demonstrated that search data well reveals the hidden opinions of the public, which cannot be easily obtained from traditional survey and interviews. One of the future works will be to establish a model describing the trend of single youth households in terms of research variables extracted from the search volume of the input keywords. To this end, degree of significance of each input keyword determined from this research should be determined.

References

Kim, Dohee. (2013), Analysis of Living Conditions According to Increase of Single-Youth Households and Government Policy Issues. *Korean Public Administration Quarterly*, 25(3): 717-742.

Kim, Hyung-Sung. (2012). Paradigm change of information technology in a big data era and utilization of public information of smart government. *Modern Society and Administration*, 22(3): 277-302.

Kim, Jongpyo. (2013). Crime prevention system in Busan utilizing big data. *Local Information Magazine*, 83(2): 23-30.

Ko, Hanseok. The science of winning with big data, *easyspub*, Seoul, 2013.

Lee, EunJoo. (2015). A Study on Big Data Convergence for Analyzing Preferences of Tourists in Jeju Island. *Local Information Magazine*, 90(1/2):9-13.

Lee, Jaeho. (2013), Utilization of big data for realization of Gov. 3.0. *KIPA Research Report*.

Lee, Keumsook, Park, Jong Soo, & Jeong, Mi Seon. (2014). Accessibility Changes in the Metropolitan Seoul Subway System: Time-distance Algorithm based on the T-card Big Data and an Accessibility Measurement Model for Un-fixed Transportation Networks. *Journal of the Economic Geographical Society of Korea*, 17(1): 98-113.

Lee, Young-Jin, & Yoon, Ji-Hwan. (2014), A Study on Utilizing SNS Big Data in the Tourism Studies, *International Journal of Tourism and Hospitality Research*, 28(3): 5-14.

Lim, Yong-Jae, Baik, Sun-Kyung, Jung, Sung-In, & Won, Hee-Sun. (2013). Cloud & big data for the smart Internet services. *PM Issue Report*, 3(1).

Ferreira, Nivan, Poco, Jorge, Vo, Huy T., Freire, Juliana, & Silva, Cludio T. (2013). Visual exploration of big spatio-temporal urban data: a study of New York City taxi trips, *IEEE Trans. Visualization and Computer Graphics*, 19(12): 2149-2158.

Ginsberg, Jeremy, Mohebbi, Matthew H., Patel, Rajan S., Brammer, Lynnette, Smolinski, Mark S., & Brilliant, Larry. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457: 1012-1014.

Hu, Han, Wen, Yonggang, Chua, Tat-seng, & Li, Xuelong. (2014). Toward scalable systems for big data analytics: a technological tutorial, *IEEE Access*, 2: 652-687.

Jardak, Christine, Mhnen, Petri, & RiihiJrvi, Janne. (2014). Spatial big data and wireless networks: experiences, application, and research challenges. *IEEE Network*, 28(4): 26-31.

Kim, Gang-Hoon, Trimi, Silvana, Chung, Ji-Hyong. (2014). Big-data applications in the government sector, *Communications of the ACM*, 57(3): 78-85.

McKinsey Global Institute. (2011). Big data: The next frontier for innovation, competition, and productivity.

Moon, Hyejung, & Cho, Hyunsuk. (2013). Big Data and Policy Design for Data Sovereignty: A Case Study on Copyright and CCL in South Korea. *International Conference on Social Computing (SocialCom)*, 1026-1029.

Noh, Kyoo-Sung. (2014). A Study on Utilization Strategy of Big Data for Local Administration by Analyzing Cases. *International Journal of Digital Convergence*, 12(1): 89-97.

O'Malley, Martin. (2014). Doing what works: Governing in the age of big data. Public Administrtion Review, Sep. Oct. 555-556.

Snijders, Chris, Matzat, Uwe, & Reips, Ulf-Dietrich. (2012). Big Data: Big gaps of knowledge in the field of Internet science. *International Journal of Internet Science*, 7(1): 1-5.

Srinivasan, Uma, & Arunasalam, Bavani. (2013). Leveraging big data analytics to reduce healthcare costs, *IT Professional*, 15(6): 21-28.

Vosen, Simeon & Schmidt, Torsten. (2011). Forecasting private consumption: survey-based indicators vs. Google trends. *Journal of Forecasting*, 30(6): 565-578.

Zhou, Xichuan, Ye, Jieping, & Feng, Yujie. (2011). Tuberculosis Surveillance by Analyzing Google Trends. *IEEE Trans. Biomedical Engineering*, 58(8):2247-2254.

김도희(金度希): 부산대학교 대학원에서 “도시비선호시설입지갈등의 갈등유발요인에 관한 연구”논문으로 행정학 박사학위(2001)를 취득하고, 현재 울산대학교 정책대학원에서 재직중이다. 주요 관심분야는 갈등관리, 지방행정, 복지행정, 여성행정 등이다. 주요논문으로는 울산광역시 남구의 문화도시정책 추진성과의 정책적 함의-남구의 문화도시 정책사례 분석을 중심으로(2012), 공공정책갈등의 제3자 중재개입의 역할과 한계-울주군청사 이전갈등사례를 중심으로(2013), 주민참여 정책사례 분석과 정책적 함의-울산광역시 북구를 중심으로(2014) 등이 있다 (dhkim5090@ulsan.ac.kr).

〈논문접수일: 2015. 03. 03 / 심사개시일: 2015. 03. 12 / 심사완료일: 2015. 05. 20〉

국문요약

검색 데이터 분석의 사회과학 연구에의 적용과 시사점: 청년 1인가구 생활실태 분석을 중심으로

김도희

본 연구에서는 최근 빅데이터의 대표적 사례인 인터넷 검색 데이터를 이용하여 청년일인가구의 생활실태분석을 시도한다. 또한 사회분석 도구로서 인터넷 검색데이터를 이용하는 실질적인 방법론을 제시하고자 한다. 더 구체적으로 본 연구에서는 인터넷 검색 데이터의 상관계수 분석을 통해 청년일인가구 분석에 필요한 연구변수를 도출한다. 이렇게 함으로써 사회현상 분석의 즉시성과 비용측면에서 기존의 방법론에 비해 제안한 새로운 방법론이 효과적임을 보이고자 한다. 본 연구의 중요성은 기존 연구의 경우 사회과학분야에 있어 빅데이터의 잠재성을 주로 다룬 데 비해, 실제 사회 문제를 분석하기 위해 인터넷 검색 데이터와 같은 빅데이터 수집 및 해석 등 방법론을 제시하였다는 것이다. 본 연구에서 제시한 방법론은 향후 보다 수요지향적인 정책수립, 연령대별 맞춤형 정책, 그리고 정책시행 후 신속한 분석에 활용될 수 있을 것이다.

주제어: 검색데이터분석, 청년1인가구, 생활실태