Technological Forecasting & Social Change 130 (2018) 69–87

Contents lists available at ScienceDirect

Technological Forecasting & Social Change

journal homepage: www.elsevier.com/locate/techfore

Ten years of research change using Google Trends: From the perspective of big data utilizations and applications

Seung-Pyo Jun[a,b,*], Hyoung Sun Yoo[a,b], San Choi[a,b]

[a] Technology Commercialization Analysis Center, Korea Institute of Science and Technology Information, 66, Hoegi-ro, Dongdaemun-gu, Seoul 130-741, South Korea
[b] Science & Technology Management & Policy, University of Science & Technology (UST), 66, Hoegi-ro, Dongdaemun-gu, Seoul 130-741, South Korea

ARTICLE INFO

Keyword:
Google Trends
Big data utilization
Big data application
Networks analysis
Author keyword
Science Journal Classification
Clustering

ABSTRACT

This study seeks to analyze the trends in research studies in the past decade which have utilized Google Trends, a new source of big data, to examine how the scope of research has expanded. Our purpose is to conduct a comprehensive and objective research into how the public use of Big Data from web searches has affected research, and furthermore, to discuss the implications of Google Trends in terms of Big Data utilization and application. To this end, we conducted a network analysis on 657 research papers that used Google Trends. We also identified the important nodes of the networks and reviewed the research directions of representative papers. The study reveals that Google Trends is used to analyze various variables in a wide range of areas, including IT, communications, medicine, health, business and economics. In addition, this study shows that research using Google Trends has increased dramatically in the last decade, and in the process, the focus of research has shifted to forecasting changes, whereas in the past the focus had been on merely describing and diagnosing research trends, such as surveillance and monitoring. This study also demonstrates that in recent years, there has been an expansion in analysis in linkage with other social Big Data sources, as researchers attempt to overcome the limitations of using only search information. Our study will provide various insights for researchers who utilize Google Trends as well as researchers who rely on various other sources of Big Data in their efforts to compare research trends and identify new areas for research.

1. Introduction

Recently, we have seen an unprecedented amount of data collected, stored and communicated within organizations and over the web. This data explosion has drawn attention to big data utilizations and analytics, and raised academic interest in the social transformation that will result from the use and application of big data. Big data is generally considered linkable information, with a large data volume and complex data structures (Khoury and Ioannidis, 2014). Examples include social media data, mobile phone call records, commercial website data, volunteering geographical information, search engine data, smart card data, and taxi trajectory data (Liu et al., 2016).

One of the most significant applications of big data utilizations is supporting various organizations and groups seeking to understand social changes and make predictions. In this study, we analyzed the trends in the development of research related to the analysis of web search engine data, which has been a leading example of the utilization and application of big data. Our goal was to better understand the directions in which the use and application of big data shall develop hereafter and identify any foreseeable challenges.

To analyze how search engine data has contributed to various areas of research, we referenced other previous studies that utilized Google Trends. Google Trends is a public website that belongs to Google Inc. and offers data based on Google Search that shows how frequently a particular search term is entered in comparison with all other search terms in different regions and languages (Google, 2017). It has been ten years since Google Trends was launched in 2006. While its service is limited to providing simple data on Google Search usage, it is a significant source since some estimate that searches on Google Trends reached 2 trillion in 2016 (Sullivan, 2016). It would be fair to state that Ginsberg et al. (2009) opened the door for the use of Google Trends in research studies. Their research published in *Nature* demonstrated that Google Trends traced and predicted the spread of influenza earlier than the Centers for Disease Control and Prevention (CDC) (Ginsberg et al., 2009).

In this study, we analyze the research papers that cited Google Trends or were otherwise influenced by Google Trends since the company made this source of data available in 2006 and review the areas of

* Corresponding author.
E-mail addresses: spjun@kisti.re.kr, spjunn@msn.com (S.-P. Jun), hsyoo@kisti.re.kr (H.S. Yoo), soullives83@kisti.re.kr (S. Choi).

https://doi.org/10.1016/j.techfore.2017.11.009
Received 22 February 2017; Received in revised form 4 November 2017; Accepted 9 November 2017
Available online 24 November 2017

progress achieved in these papers. This outlook on the development of research using Google Trends will shed light on how the utilization and application of big data has opened new opportunities for analysis. Google Trends is already used in a wide range of areas, but it has been difficult to obtain an overarching review of this usage. We believe this study on the use of Google Trends, as an example of big data sources or their utilization, will have significant implications both academically and practically. We hope our findings will contribute to the theoretical developments in the use and application of big data by overseeing and clustering studies relating to Google Trends. Our more practical goal is to create new values in diverse areas through Google Trends by clarifying its potentials and limitations and promoting a better understanding of its characteristics.

Within the body of literature that cites Google Trends, our study focuses particularly on research papers. In regards to the technological life cycle and the trend of developments in related literature, in general, paper publications have a strong tendency to precede patents, while patents have the characteristic of being temporally closer to commercialization (Daim et al., 2005; Martin and Daim, 2007; Watts and Porter, 1997). Considering these characteristics, we concluded that research activities related to Google Trends in the socio-technical system (paper publications) can be exhibited in the patents (Jun, 2012a), and we conducted a simplified comparison by adopting the patents as a proxy indicator for the socio-technical impact of the papers.

This study used network analysis (or social network analysis, SNA) to analyze directions of research related to Google Trends. SNA is a research method for a system approach that visualizes the invisible flow in the network by identifying the types of interactions, correlations and roles among actors in the network (Scott, 2012). In general, simple meta-analysis, such as the analysis of the frequency of keywords, can be used to examine the changes in research in specific areas, but in this approach, it remains difficult to learn how various issues are linked together. In addition, SNA provides further implications by allowing us to identify more specific and objective development patterns. In particular, we used SNA to perform clustering for identifying the academic disciplines in which the research using Google Trends was taking place and we also employed SNA to identify the areas that can be expected to play a key role in the structure of linkage between various fields or keywords.

With this goal in mind, Section 2 reviews theories and literatures relating to online search and Google Trends. Section 3 deals with research methodologies, explaining how we collected and analyzed research papers related to Google Trends. Section 4 introduces statistical data on relevant research papers and the outcome of the network and cluster analysis we performed on this data. In addition, to reinforce our clais regarding the development directions, we also provide the results of patent analysis and SNA. In Section 5, we take a closer look of each paper to analyze the trends of major research clusters. In Section 6, we propose some insights gained from reviewing the scope and development process of research studies relevant to Google Trends, in regards to big data utilizations and applications.

2. Theoretical background and prior research

There are several reasons why Google Trends has become such a popular source for big data research and applications. First, Google Search provides an excellent platform for observing consumers' information seeking activities. It offers instant reflection of the needs, wants, demands and interests of its users. Second, Google Trends is easy to use because Google not only collects data but also provides a variety of options for comparison (Jun et al., 2014a). In this light, this section considers two major theories about information seeking, through Google searches. The section first highlights the characteristics of information seeking for purchasing decisions in marketing and business context and then for technology adoption in the economy or in the context of technological innovation. Even prior to the launch of Google Trends, in fact, there had been a lot of preceding research which utilized the search information in the fields of economics or technological innovation. It is important to understand this context (Ettredge et al., 2005; Moe, 2003; Zimmer et al., 2007). Next, we introduce Google Trends, which offers extensive statistics on search data, and briefly review earlier studies on Google Trends from the perspective of Google, Inc.

2.1. Consumer search behavior and consumer adoption model

Naturally, consumers act to maximize their satisfaction by reducing any uncertainties, be it psychological or financial, and avoid risks involved in the purchasing process. This prompts consumers to engage in information seeking behaviour, which these days often involves searching the web, thanks to the heightened Internet accessibility made possible by the rise of smartphones. A consumer's online search often leads to online product purchasing (Shim et al., 2001). This positive correlation between one's willingness to search the web and to purchase a product online has been described thoroughly in To et al., (2007).

The information seeking behaviour of consumers are often affected by the characteristics of the relevant product itself and those of the consumer who conducts the search. One's pre-existent knowledge plays an important role throughout the search process and eventually in the decision-making on whether or not to purchase a certain product (Raju et al., 1995). In general, as a consumer's knowledge improves, so does one's information seeking activities. However, the correlation is valid only up to a certain point, as a consumer with sufficient knowledge would have little need to perform additional searches, resulting in a decrease in searching activities. A graph of the relations would be an inverse U-shape one, with the peak in the middle with a consumer with an intermediate level of knowledge engaging in searches more actively than those with higher or lower levels of knowledge (Bettman and Park, 1980; Rao and Sieben, 1992). The type of the relevant product is another factor that affects a consumer's information seeking activities. Consumers tend to be more active when they associate higher level of risk with the product, as shown by Beatty and Smith (1987). The same tendency is observed with products with higher price or higher involvement. Similarly, consumers rely more on searches when information is lacking on a specific product due to the novelty of its brand or the wide range of available products of the kind (Assael, 1992).

The consumer adoption model in the field of technological innovation is central in explaining consumer behavior. Here, any goods, services or ideas are deemed as innovative if people perceive them as something new even if they have been around for a long time. As Rogers (2003), put it, diffusion of innovation is "*the dissemination of a new idea generated by invention or creation among the final users or adopters*." In this regard, the process of consumer adoption closely concerns an individual's line of thoughts from the initial perception of the innovation up to the point of final adoption. The process can be broken down to five stages: awareness, interest, evaluation, trial and adoption (Kotler and Keller, 2008). This classification focuses on an individual's mental process, observing the process from the perspective of the consumers. On the other hand, the conventional approach has focused on the life cycle, assessing the process from the perspective of the producers. Online search, the subject of the current study, belongs to the Interest stage under Rogers' classification.

As we have seen, search is an act that accompanies purchase of a product or acceptance of innovation. Thus, statistics on search activities are good resources for monitoring, analyzing, and even predicting acceptance of a product or an innovative new technology by individuals or the society. Due to this possibility, before Google fully released its search data and before the publication by Ginsberg et al. (2009), researchers had already conducted studies using data from search engines other than Google. Moe (2003), who researched consumers' online information search activities, categorized such activities into browsing and searching according to the motivation and their outcome. Also,

Ettredge et al. (2005) argued that we can identify people's needs, wants, interests, and concerns through search engine data, and used the search information to compare the behavior of men and women during job searches based on the difference in search terms (Ettredge et al., 2005). According to Zimmer et al. (2007), through search engines, users can view many search results for the keyword that interests them and access a wide range of information sources. Given the numerous search results, consumers do not utilize all of these available information sources and, instead, seek out specific sources to acquire the information they need. In other words, information searching by consumers involves making decisions regarding which of the sources of information to select and use (Zimmer et al., 2007). Fenn and Raskino (2008), who systematically conceptualized the hype cycle, also predicted that the hype cycle that had they contributed to theorizing might be empirically demonstrated through the analysis of data from search engines such as Google (Fenn and Raskino, 2008).

There has thus been ongoing research focused on search activities and as the use of search engines increased, empirical studies that attempted to utilize the search engine data continued to multiply. The Google Trends service became fully available in 2008 and from then on, there were significant changes in the related research.

2.2. Google Trends and beginning of research

Google Trends was first introduced on May 11, 2006. Google then released Google Insights for Search on August 5, 2008, an advanced and more detailed service that provided data on search trends to the users. On September 27, 2012, Google combined Google Insights for Search with Google Trends (Google, 2017). Google Trends merely provides the data on searches on Google Search, but the latter's popularity has been explosive: it has recorded 1.2 trillion searches per year in 2012, and the searches are estimated to have reached about two trillion in 2016 (Sullivan, 2016). This translates to 5.5 billion searches per day and 167 billion searches per month, leaving little doubt that we are dealing with a typical big data source. Of course, there are limitations to the uses of Google search. For example, although there are slightly different results depending on the research, in some countries such as China, Japan, South Korea, and Russia Google is not the dominant search engine due to political or linguistic issues. However, the global statistics on search engine market share shows that Google has maintained 90% of the market share from 2010 onward and therefore still retains its representative dominance as a search engine (StatCounter, 2017).

This has been the case since the launch of Google Insights for Search, a free service introduced on August 5, 2008. It was an extended version of the earlier Google Trends and popular research subject not only for marketing managers but also for other researchers. The service allowed the users to trace various terms and phrases entered on Google Search. Now, with the current Google Trends, combined with Google Insights, users can not only categorize and organize statistical data on searches but also sort it by geographical areas (Google, 2017). Currently, Google Trends offers data on the search volume for specific search terms since 2004. We have access to relative traffic data, converted with the maximum search volume for a search term set to 100, tracked by various time intervals (minutes to months). This traffic data is available specifically classified by category (e.g. health, game, etc.), region, search site, etc., along with information on related topics and search terms. However, as explained above, one limitation is that Google only provides a relative search value (i.e., a normalized index) and does not provide exact search volume.

Google itself was the most effective voice in publicizing the great applicability of Google Trends. Ginsberg and his colleagues, working for Google, showed that diseases similar to influenza in population groups can be traced using Google Trends in a research paper published in Nature, prompting other research studies to follow suit (Ginsberg et al., 2009). While the paper was not without prior studies, it has become one of the most cited studies among those that utilized Google Trends. The study has three points on the use of statistics of search data. The search traffic provided by Google Trends is useful to promptly detect a certain phenomenon and is therefore an excellent monitoring tool. It is representative, as more than 5.5 billion searches are run on Google every day. Its search traffic exhibits high correlation with social phenomena, which is indicative of its high potential to be applied in a wide range of areas. Google then went on to suggest fields other than medicine that can benefit from Google Trends. Choi and Varian, also working for Google, argued that Google Trends can be used to predict unemployment rate (Choi and Varian, 2009). Their later research showed that Google Trends has a significant correlation to car or house sales (Choi and Varian, 2012). As such, research using Google Trends was bolstered by researchers at Google in the early stage. Their papers and studies reported that Google Trends has a “now-casting” function and may become a great forecasting tool in various fields.

The initial studies on Google Trends, largely led by Google, has been through academic scrutiny and thus evolved thanks to the efforts of scientists in various fields across the world. In the below, we will review how these studies have made progresses in which areas.

3. Research methodology

3.1. Analysis methodology

To analyze the changes in research related to the Google Trends, we analyzed the relevant literature published over the past 10 years. Specifically, we selected the author keywords and the ASJC (All Science Journal Classification) code as variables to analyze the changes in the literature. Author keywords are determined by the author, and are intended to reflect the theme and characteristics of the study. The SCOPUS database provided by Elsevier contains 18,000 journals published by around 5,000 publishers. Each is assigned one or more ASJC codes classified by SCOPUS. This information can thus be considered to provide an objective viewpoint regarding the area of the journal in which the paper was published. Our hypothesis was that the temporal frequency of these two types of information and the analysis of changes in their interrelationships can explain the changes in the directions of research (Kim et al., 2013; Romanelli and Feldman, 2007; Scott, 2012).

In this study, we analyse 657 Google Trends related documents using simple descriptive statistical analysis, and utilize network analysis for the analysis of specific research directions. Network analysis has evolved as a methodology for the analysis of social structure. The most significant feature of network analysis is the use of relational data. The people, organizations, and objects that form a network are referred to as nodes, and a network structure is that which expresses the position and relation between nodes (Scott, 2012). For network analysis, it is important to determine the nodes and links of the network. In this study, we consider the author keyword and ASJC code as nodes and we propose a network model in which the relationships of each bibliographical information are considered to be links and we perform co-word or co-classification network analysis.

In these one-mode networks, we utilized association strength as the method for analyzing the similarity of co-occurrences. The methods of analysing the properties of similarity measures for co-occurrence data include the association strength, the cosine, the inclusion index, and the Jaccard index. Among these, the association strength, which constitutes a probabilistic measure, was chosen because it is suitable for scientometric research (Van Eck and Waltman, 2009). Specifically, we used the method above to conduct community analysis (that is, cluster analysis) and visualization, and for this purpose we used the VOS viewer (version 1.6.5) software (Van Eck and Waltman, 2010). Here, community analysis is an analysis that detects the community structure at the substructure level of the network. A community structure in a network is a subgroup (community or module) in which the relationships among specific nodes (local communities) are dense internally but not externally (i.e. with other communities) (Newman, 2010). In the network

Table 1
Detailed conditions of the data collection and results.

Descriptions	Conditions/results
Search target DB	SCOPUS.COM
Search expression	"google trends" or "google insights" or "google search queries" or "google search query" or "google search volume" or "google search data"
Search field	Title, Key words, Abstract
Search period	2006～2017. Jan.
Last search date	2017. 1. 20.
Search results	657 documents
Total references of searched documents	13,906 references
Total kinds of author keywords of searched documents	1,434 kinds of words
Total types of ASJC types of searched documents	168 types of codes

analysis, the meaning of the community structure is the identification of conceptual groups that form a contextual cluster (Paranyushkin, 2011). VOS viewer calculates the community, or cluster, based on modularity (Newman and Girvan, 2004). VOS viewer also improves the accuracy by adding a smart local moving algorithm (Waltman and van Eck, 2013), and adjusts the number of clusters by adjusting the resolution parameter γ in the Modularity function (Yan et al., 2012).

The second network model proposed in this study is two-mode networks showing the relationship between heterogeneous nodes. In this study, we examine the relationship between the author keywords and different kinds of nodes such as research area (that is ASJC code) and time period to examine the research trends. Ultimately, this is an attempt to derive and compare the keywords that are considered important for research in each field and period. We also used Gephi (version 0.9.1) to analyze the similarity and community of co-occurrences in these two-mode networks (Blondel et al., 2008; Knuth, 1993). Graphs are usually laid out with "Force-based" algorithms (Bastian et al., 2009). Before analyzing the above network, we pre-processed the thesaurus of the dataset, using the KnowledgeMatrix program (Lee et al., 2008).[1]

Also, to find the nodes that play a key role in the networks, we comparatively analysed centralities of the nodes. Centrality is an indicator of the degree to which a node (or actor) is centrally located in the network, and can be measured in a variety of ways: the leading examples of such measures are degree, closeness and betweenness centrality. Degree centrality, in which importance is given to the degree of connection with other nodes, indicates the sum of the nodes connected to a specific node in the network. Degree centrality can be defined as "*the number of ties incident upon a node* (Borgatti, 2005)." Closeness centrality is the concept of how close a node is to another node, focuses on the geodesic (or path) distance between nodes in a network. Closeness centrality is defined as "*the sum of graph-theoretic distances from all other nodes, where the distance from a node to another is defined as the length* (*in links*) *of the shortest path from one to the other* (Borgatti, 2005)." Finally, betweenness centrality refers to the degree to which a node is located between other nodes in the network, and measures the degree to which a specific node plays a broker role (Freeman, 1979). Betweenness centrality is defined as "*the share of times that a node i needs a node k* (*whose centrality is being measured*) *in order to reach a node j via the shortest path* (Borgatti, 2005)." Freeman (1979) explained about the three centralities of the network analysis: "*the centrality of a point may be determined by reference to any of three different structural attributes of that point: its degree, its betweenness, or its closeness. The choice of a structural attribute and its associated measure depends upon the context of the substantive application intended. Concern with communication activity suggests a degree-based measure. Interest in control of communication requires a measure based upon betweenness. And concern with either independence or efficiency leads to the choice of a measured based upon closeness.*" In this study, we examine these three types of centrality and objectively identify the nodes we need to pay attention to present and future, namely the keywords and ASJC codes.

In addition, to understand the implications of the directions of future development of research papers using the Google Trends we identified from our network analysis results we also performed simple time series analysis. We additionally applied a methodology that compares the annual trends of related keywords or categories to the trends in other types of documents (patents) (Daim et al., 2007; Watts and Porter, 1997) and compared this to the SNA results and performed re-verification.

3.2. Research dataset

As mentioned in the introduction, Google Trends was launched in 2006, and based on various studies, results have been released in articles, conferences, papers, and books. To collect these research results and to build the research dataset, we used the SCOPUS data, as explained in Table 1. Also, to ensure the accuracy of the searches, we used phrase search expressions, and in all the search phrases we included Google. Therefore, we inevitably omitted the cases in which only comprehensive search information was mentioned without specifying Google. The reason we limited the search expressions to those specifying Google was because if we were to include search terms that are too general (e.g. search traffic), there would be the problem that the analysis dataset will include many papers about the general technology of the search engines rather than about the use of Google Trends. The targeted search fields were limited to the title, keywords, and abstract and the search period was set from 2006 to the present, taking into account the period Google Trends has been in service.

We collected 657 documents by applying the search conditions above. Based on the number of citations, the collected papers were compared to the paper search results provided by the Google Scholar service, and we confirmed that the majority of the highly cited papers had been included. The top 20 papers in terms of citation, included in 657 papers, are listed in Table 2. By performing an analysis of the citations among the 657 papers, we proceeded to perform another review of the appropriateness of literature. Adopting the reference dataset of 13,906 that had been cited by the 657 papers, we performed the citation analysis once again. Table 2 shows a comparison of the citation rankings for all of the literature on SCOPUS and the citation rankings within the dataset adopted in our study (we presented them in Table 2 as 'Rank in DB' which refers to the ranking among the 13,906 references). In most cases, a paper with a high rank of SCOPUS citations was also found to be highly ranked in terms of the citation ranking within the reference dataset as well, but some general papers that were low in relevance were found to have low ranking.[2] We once again confirmed that the 657 cases consisted of papers that have a relatively high degree of relevance to Google Trends. Also, as described above, the study by Ginsberg et al. (2009) showed an overwhelmingly high number of citations and was ranked the highest in terms of the number (209 citations) of citations even among the 657 cases, once again demonstrating that it is one of the most influential papers among the studies that used

[1] The thesaurus processing for Google, Google Trends, and Google Insights was done only for capitalized/non-capitalized and singular/plural forms. The reason for not processing thesaurus of these three words is that each word may not be used in the documentations in the same meaning. It is also to see if there is a difference in clustering (including time series).

[2] The rank correlation analysis result for the two ranks presented in Table 2 was Spearman correlation coefficient (rho) of 0.435 and the one-side p-value of 0.028. Thus, at the significance level of 0.05, the relationship of two rankings was significant (Kendall's tau test also showed a significant relationship at the same significance level).

Table 2
Status of the highly ranked papers in the dataset, ranked in terms of the number of citations.

Rank in SCOPUS	Title	Publication year	ASJC code	Number of citation in SCOPUS	Document type	Author name	Source title	Rank in DB
1	Detecting influenza epidemics using search engine query data	2009	1000	1,222	Article	Ginsberg, Jeremy et al.	Nature	1
2	Predicting the present with Google Trends	2012	2002	222	Article	Choi, Hyunyoung & Varian, Hal	Economic Record	3
3	The web-wide world	2006	1000	217	Short Survey	Butler, Declan	Nature	1266
4	In search of attention	2011	2003	205	Article	Da, Zhi et al.	Journal of Finance	5
5	Google trends: a web-based tool for real-time surveillance of disease outbreaks	2009	2725	182	Article	Carneiro, Herman et al.	Clinical Infectious Diseases	2
6	Quantifying trading behavior in financial markets using google trends	2013	1000	131	Article	Preis, Tobias et al.	Scientific Reports	17
7	The influence of task and gender on search and evaluation behavior using Google	2006	1706	112	Article	Lorigo, Lori et al.	Information Processing and Management	1266
8	Assessing Google Flu Trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic	2011	1100	98	Article	Cook, Samantha et al.	PLoS ONE	10
9	Complex dynamics of our economic life on different scales	2010	2600	96	Article	Preis, Tobias et al.	Philosophical Transactions of the Royal Society A	15
10	Automatic evaluation of topic coherence	2010	1203	95	Conference Paper	Newman, David et al.	NAACL HLT 2010 - Human Language Technologies	None
11	Reprint of: the anatomy of a large-scale hypertextual web search engine	2012	1705	76	Article	Brin, Sergey & Page, Lawrence	Computer Networks	None
	A study of results overlap and uniqueness among major web search engines	2006	1706	76	Article	Spink, Amanda et al.	Information Processing and Management	270
13	Google Flu Trends: Correlation with emergency department influenza rates and crowding metrics	2012	2725	74	Article	Dugas, Andrea et al.	Clinical Infectious Diseases	18
	More diseases tracked by using google trends	2009	2713	74	Letter	Pelat, Camille et al.	Emerging Infectious Diseases	8
15	Using web search query data to monitor dengue epidemics	2011	2725	69	Article	Chan, Emily H et al.	PLoS Neglected Tropical Diseases	43
16	Google Glass in pediatric surgery: an exploratory study	2014	2746	65	Article	Muensterer, Oliver J et al.	International Journal of Surgery	None
17	Prediction of dengue incidence using search query surveillance	2011	2725	64	Article	Althouse, Benjamin M et al.	PLoS Neglected Tropical Diseases	20
18	Forecasting private consumption: survey-based indicators vs. google trends	2011	2611	61	Article	Vosen, Simeon & Schmidt, Torsten	Journal of Forecasting	14
	Reassessing Google Flu Trends data for detection of seasonal and pandemic influenza	2013	2804	61	Article	Olson, Donald R et al.	PLoS Computational Biology	33
20	The utility of "Google Trends" for epidemiological research	2010	2701	60	Article	Seifter, Ari et al.	Geospatial Health	13

Note: ASJC means All Science Journal Classification, DB refers to research dataset.

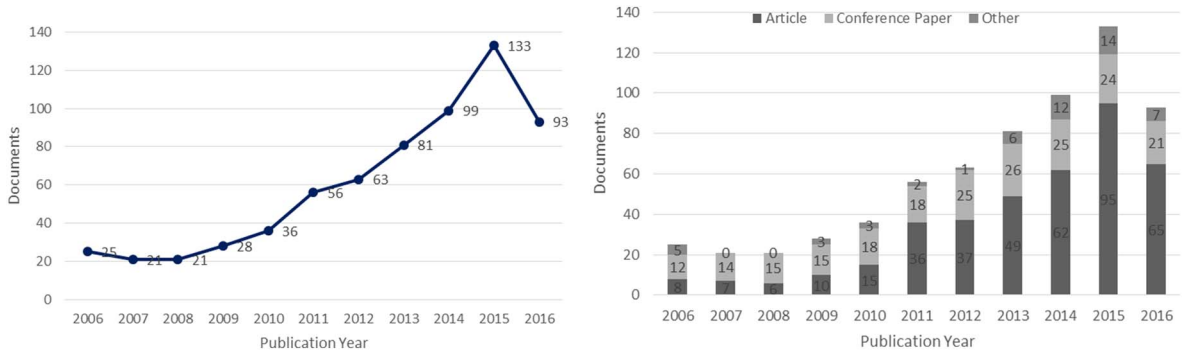

Fig. 1. Publication trends of literature related to Google Trends, by publication year (left) and literature type (right)
Note: the "Other" category of documents include chapters, erratum, letters, notes, reviews, and short surveys.
Some of the other categories of documents have duplicate elements or relatively low importance (volume), but they are not excluded from the dataset due to their low impact. In this study, 1,434 kinds of keywords were shown 2,250 times and 168 kinds of ASJC appeared 1,479 times in 657 documents. Among them, five of the ASJC were included in the three documents of erratum. Keywords did not exist. In addition, notes were found to contain 8 keywords and 6 ASJCs in 5 documents, all of which showed little effect on the analysis.

Google Trends. Also, Table 2 shows that many highly ranked papers are related to medical studies, and we should take into account the fact that the medical field has a practice of making citations relatively frequently (Althouse et al., 2009; Seglen, 1997).

4. Analysis results

4.1. Descriptive statistical analysis results

First, the graph on the left of Fig. 1 presents the publishing status of the 657 cases by year (excluding one that was published in 2017). When we track the data from 2008, the year in which the service was launched in earnest, we can see that the related literature expanded with a high growth rate while receiving high attention from 2009 onwards. Although the related literature showed a slight decline in 2016, overall, this rapid increase has continued until recently.[3] In this regard, this study is very timely. On the right side of Fig. 1, the graph shows annual publication trends according to the specific types of documents. The number of conference presentations was relatively steady from the beginning, but the number of articles started to increase after 2011, and it was only recently that this number has greatly increased. Major achievements have been observed since 2011, when the importance of "big data" began to receive much attention.[4]

Table 3 shows the top rankers in terms of ASJC, first author country, and affiliation. As shown in Table 3, the most common categories are those related to computer science, followed by the medical and biology sciences, and then followed by economics and business related fields. As for the first author country, the United States was overwhelmingly the most common, and it was followed by China, Germany, and the United Kingdom. Meanwhile, affiliation, unlike other bibliographic categories, did not show any overwhelmingly dominant trend. The largest number was affiliated with Harvard University, followed by Google Inc. The majority of the remainder were universities, indicating that university-led research has been prominent.

4.2. Results of the cluster analysis of Google trends literature

Here, we examine the results of our analysis of co-author keyword to find out how author keywords co-occur in the Google Trends-related literature. Through this network analysis, we examined the results of the community (cluster) analysis of the literature on Google Trends. Fig. 2 shows the density visualization results based on author keywords. When we visually examine Fig. 2, we can see that there are 3 to 4 clusters. In the density view, the color of a point in a map is determined based on the item density of the point. Beginning on the right, we notice that Google Trends itself forms one cluster, followed by the clustering of economic fields such as forecasting, tourism, unemployment, and global financial crisis. Next, in the center, we see a gathering of keywords such as Big data, Twitter, social network, influenza, and epidemic, forming a cluster of big data or health related issues. Finally, on the left, we see author keywords such as Google, search engine, internet search, Bing, and page rank in proximity to one another, and this also indicates a collection of issues related to Google search engine itself.

Examining the network analysis results of the author keywords shown above, we confirmed that qualitatively, there are at least 4 - 5 research clusters. We have adjusted the minimum cluster size for clustering the author keywords to generate several clustering cases, and from among these, we selected the four clusters that are presented in Table 4 and Fig. 3, selected as a relatively significant examples of clustering results (minimum cluster size: 150). Google Trends was, of course, found to be central among the keywords in the largest cluster. We noted the presence of keywords related to economic fields such as forecasting, nowcasting, time series analysis, and volatility. The next cluster includes keywords related to medical services and health such as influenza, Google flu trends, infodemiology, suicide, etc. This was followed by Cluster 3 and 4, which mainly consisted of keywords that are highly relevant to information systems or computer science and contain a number of keywords related to big data utilizations and applications.

Fig. 3 presents the results of network visualization analysis of the network distinguished into four clusters. We present the four clustering results in respectively different colors. In VOS viewer, when you create a network picture (also known as Label view), circles or labels of important items (in this case, keywords with many simultaneous links) appear large. In Fig. 3, the link strength, which calculates the distance between nodes, is extracted by the full counting method. As the similarity of the items increases, the distances appear closer, but the minimum distances are maintained to prevent complete overlap (Van Eck and Waltman, 2010).

Although the results are relatively concentrated in the center, Clusters 2 and 4 in Table 4 are mixed with other clusters and are respectively indicated by yellow and green. On the other hand, the red cluster (Cluster 1 in Table 4) and the blue cluster (Cluster 3 in Table 4) are clearly distinguishable. This confirms that the economic studies shown as the red cluster and the information-related studies shown in blue form relatively distinct research clusters. This leads us to expect that the research in these two fields will yield relatively independent and unique performances. It is characterized by divergence of diverse

[3] We shall defer drawing any further conclusions regarding changes in the growth rate at the present, since the dataset for this study was formed in January 2017 and 2016 was the only year that exhibited a decline amid the overall trend of growth.

[4] According to Liu et al. (2016) (Liu et al., 2016), the publication of the academic literature with the topic field of "big data" started to show a noticeable increase in 2011, which led to a dramatic increase from 2012 onward.

Table 3
Comparison of the number of documents by ASJC code, First Author Country and Affiliation.

Rank	ASJC Code		First author country		Affiliation	
	Description	Documents	Description	Documents	Description	Documents
1	1705: Computer networks and communications	42	United States	192	Harvard University |USA	14
2	1700: Computer science(all)	40	China	50	Google Inc. |USA	12
3	1710: Information systems	35	Germany	38	Chinese Academy of Sciences |CHN	10
4	1706: Computer science applications	30	United Kingdom	34	Children's Hospital Boston |USA	9
5	1100: Agricultural and biological sciences(all)	28	Italy	24	Johns Hopkins University |USA	8
6	2002: Economics and econometrics	27	Australia	22	University of Pennsylvania |USA	7
7	2700: Medicine(all)	26	Spain	21	University of Sydney |AUS	
8	1712: Software	24	Canada	19	Carnegie Mellon University |USA	
9	2739: Public health, environmental and occupational health	20	South Korea		George Washington University |USA	6
10	1702: Artificial intelligence	15	Taiwan		National University of Singapore |SGP	
11	2718: Health informatics		India	15	University of Michigan |USA	
12	1000: General		Greece	12	University of Melbourne |AUS	
13	2725: Infectious diseases	12	France	11	Virginia Tech |USA	
14	1403: Business and international management	11	Netherlands	10	University of California San Diego |USA	
15	1709: Human-computer interaction		Japan	9	University of Regensburg |DEU	
16	1703: Computational theory and mathematics	10	Ireland	8	University di Verona |ITA	
17	2200: Engineering(all)		Austria	7	City University of Hong Kong |HKG	5
18	2728: Clinical neurology	9	Czech Republic		Vanderbilt University Hospital |USA	
19	1704: Computer graphics and computer-aided design	8	Thailand		University of Washington |USA	
20	2741: Radiology nuclear medicine and imaging		Turkey	6	San Diego State University |USA	
	2746: Surgery				University of Warwick |GBR	
	3315: Communication				University of Genoa |ITA	
					University of Alberta |CAN	
					Santa Fe Institute |USA	

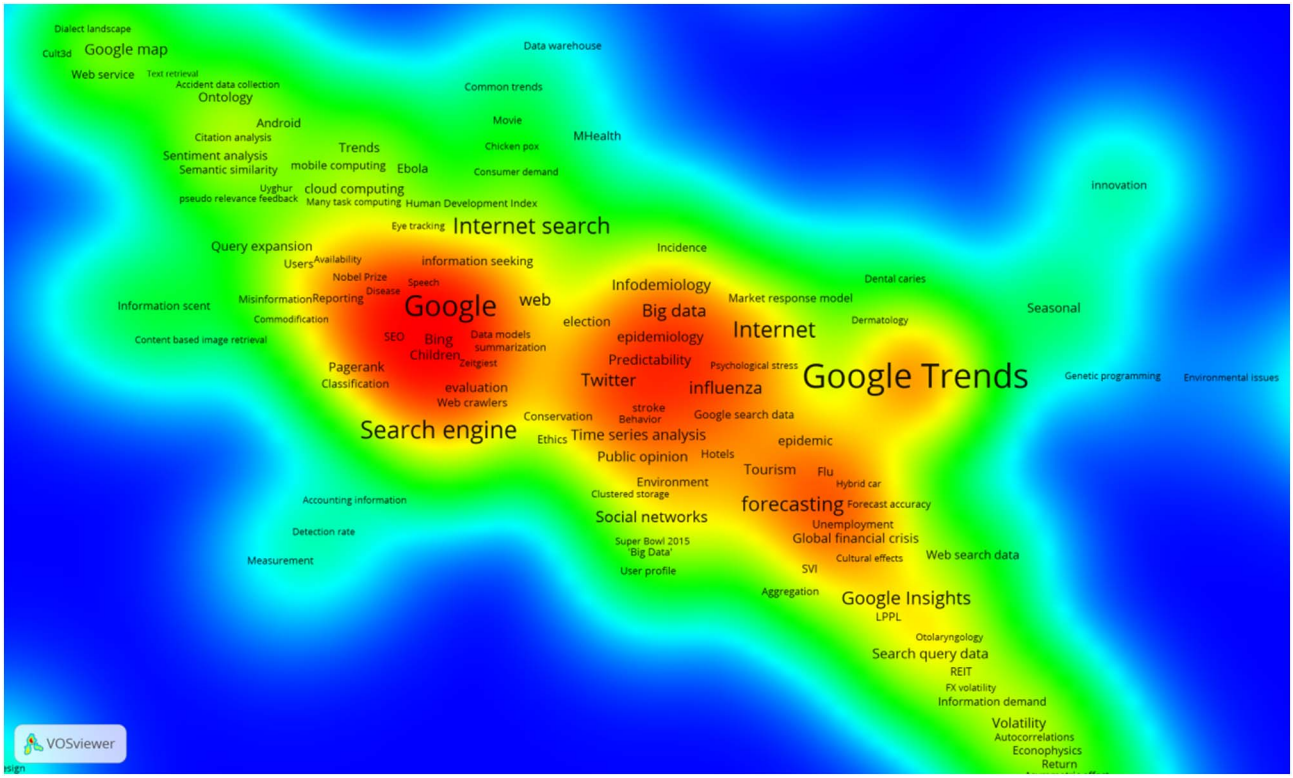

Fig. 2. Density visualization of author keywords.

themes (keywords) centered on the economic studies. Meanwhile, information-related studies are characterized by a combination of relatively smaller and more diverse studies related to Google or search rather than Google Trends. On the other hand, the two other fields which are located in mixed dispersion around the center can be explained as fields in which new opportunities are expected to be created through convergence. What we should pay attention to here is the existence of keywords that are close to other clusters. For example, in Fig. 3 we see that "user profile" exists in Cluster 4, but it is also very close to Cluster 1 which indicates that related technology is also being used in the economic field.

Examining the one-mode network analysis above, we analysed the co-occurrence of author keywords. Although we were able to find the meaning for the results, however, it was somewhat challenging to identify the comprehensive features. Therefore, on only authors, we proceeded to perform a combined examination of the scientific

Table 4
Clustering of the author keywords for literatures related to Google Trends.

Cluster rank	Cluster 1	Cluster 2	Cluster 3	Cluster 4
1	Google Trends	Google	data mining	Search engine
2	forecasting	Internet	Information retrieval	Internet search
3	Google Insights	influenza	queries	Social media
4	Twitter	Big data	Google Earth	World Wide Web
5	nowcasting	web	Google map	overlap
6	Surveillance	Google Flu trends	Text mining	Privacy
7	Time series analysis	Social networks	Page rank	web mining
8	Bing	Infodemiology	Relevance ranking	AOL
9	Search query data	Suicide	semantic	anonymity
10	Volatility	Predictability	Ontology	information disclosure
Total weight	6,994	5,534	3,984	2,422
Cluster name	Economy +	Medical +	Information system +	Information general +

Note: the weight means the number of links of an item.

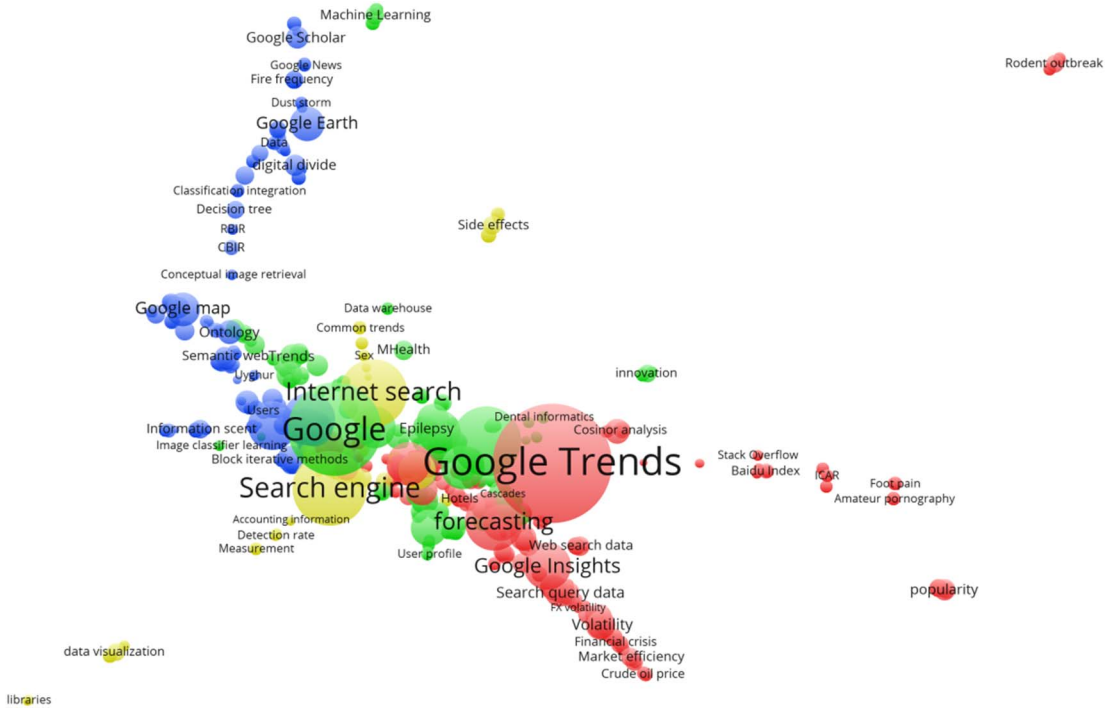

Fig. 3. Network visualization of author keywords.

classification (ASJC codes).

Fig. 4 presents the network analysis results for the keywords and ASJC codes. According to Fig. 4, the field that covers the widest and most noticeable area are the fields of “information systems” and “computer science,” with a particular focus on 1710 (Information Systems). With the Google Trends keyword at the center, this area in question was located between keywords such as internet search and Twitter and 1710 and 1706 (Computer Science Applications) and 1705 (Computer Networks and Communications) shared the keyword “search engine.” 3309 (Library and Information Sciences), which is located between 1710 and 1706, had the characteristic of sharing “information retrieval” and it was particularly notable that 1706 and 1802 (Information Systems and Management) were also in relatively close distance. 1705 was found to have “data mining” as a relatively close keyword. The relatively distant 1712 (Software) was characterized by its closeness to “web mining” or “query expansion.”

The second area that is of interest is the medical studies related field, with a focus on 2700 (Medicine). Fig. 4 shows that this area is focused on the 2700 field and includes the fields of 2733 (Otorhinolaryngology), 2739 (Public Health, Environmental and Occupational Health) and 2746 (Surgery) that share keywords such as influenza and surveillance. Among these, the 2739 field is shown to be the center of another clustering and shares the keyword “internet” with the field of 2728 (Clinical Neurology). The 2728 field shares Google Flu trends with the 2700 field, which is the core center of the related group.

The last area that deserves our attention are the fields related to 2002 (Economic & Econometrics), which is focused on Google Trends and shares forecasting, social media, and Big data with the field of 2700 (Medicine). The 2002 field classification was close to keywords such as Google search volume and volatility while the 2003 (finance) classification was close to keywords such as volatility and investor attention. The 2003 classification shared search query data keywords with the 2000 (Economics, Econometrics and Finance) classification, and were close to the keywords market efficiency and recession. In addition, the 1400 (Business, Management and Accounting) classification was close to the keyword GARCH and the 3315 (Communication) classification was close to search traffic.

To perform the community analysis of the network that we had been analysing qualitatively so far, we followed modularity-based algorithms (Blondel et al., 2008). When the modularity value is larger, it indicates that the network is well divided, and in this case the modularity value was found to be 0.471 and the modularity with resolution was 1.327

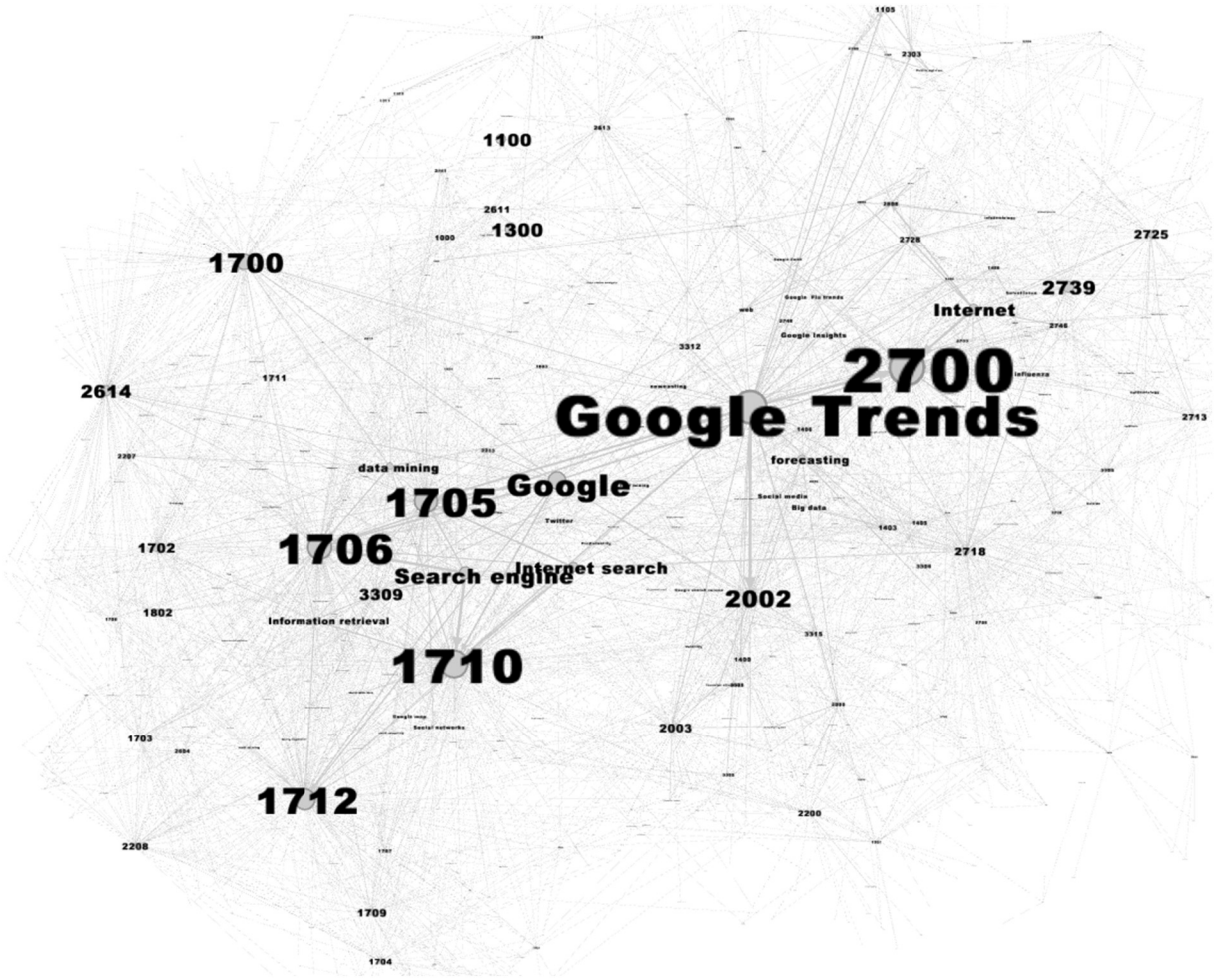

Fig. 4. Two-mode network visualization of the author keyword and ASJC codes.
Note: the key ASJC codes are 1710 (Information Systems), 2700 (Medicine) and 2002 (Economic & Econometrics).

Table 5
Results of the author keywords & ASJC community analysis.

Class	ASJC codes	Author Keyword	Total weight	Cluster Name
1	1710, 1706, 1705, 1712, 1100	Google, Search engine, Internet search	1,483	Information Systems and Computer Science Applications
2	2700, 2739, 2725, 2718, 2713, 2728, 1000, 2746	Internet, influenza, web, Google Flu trends	829	Medicine, Public Health, Health Informatics
3	2002, 2003, 1105, 2303, 1403, 1400, 1405, 1406, 2000	Google Trends, forecasting, Big data, Google Insights, Twitter, nowcasting	726	Economics, Business, Finance and Ecology
4	1700, 1300, 2614, 3312, 3315, 3300, 3304	queries	421	Etc.
5	2200, 2613, 1804, 2308	Social media, Google Earth, Facebook	199	Statistics, Probability, Management, Monitoring, Policy and Law

Note: the weight means the number of links of an item.

while the number of communities was 19. To obtain the suitable communities, we adjusted the resolution to 2.0 (Lambiotte et al., 2008). Among the derived communities, we identified the classes that had a distribution size of 10 or more: these are presented in Table 5 below.[5] The top three classes were nearly identical to the content of the analysis above, and among the other fields, the notable ones were Statistics, Policy and Law.

[5] In the Gephi program used for two-mode analysis, the distribution size is a filter for adjusting the degree range, which means that the degree distribution is limited to 10 or more. In other words, the community is defined only when the degree of the node is 10 or more.

Reviewing the above results of our one-mode and two-mode cluster analysis through SNA, we see that there are three major areas that are distinguished from the other areas. Specifically, as shown in Table 5, these three areas are 1) Information Systems and Computer Science Applications, 2) Medicine, Public Health, Health Informatics, 3) Economics, Business, Finance and Ecology.

4.3. Identifying core ASJC and keywords

Based on the above results of network analysis, we identified key areas and keywords in a decade of research related to Google Trends and confirmed the clustering results. In this study, to draw specific

comparisons in the nodes, we analysed degree centrality, closeness centrality and betweenness centrality (Freeman, 1979). Degree centrality of a node is for measuring how many nodes are directly connected to the node in the network. The degree of centrality can thus be used to determine the direct influence of each node. In Fig. 3, the size of each node represents the weighted degree centrality of the node. As described in the previous section, nodes with high levels of centrality are the nodes that remain central in the research to date, and which play a central role in clustering research areas. Closeness centrality is defined as the reciprocal of the mean of the shortest distance to other nodes. That is, a node with a short average distance from other nodes has a high closeness centrality, and may be the fastest influence on other nodes dispersed throughout the network. In the case of this study, scientific innovation in areas of high closeness centrality has the fastest impact on other areas, or inversely, is likely to have the fastest exposure to innovations in other sectors. Degree centrality is concerned only with directly connected nodes and it only gives us a view of the roles in the local range. The rankings of degree centrality (hereinafter referred to as the Rank in DC) therefore show the node with a large degree of influence, with many direct connections. On the other hand, betweenness centrality is an indicator of the control over flow in the network. Betweenness centrality of a node represents the probability of which the node stands in the shortest paths between pairs of other nodes. A node with higher betweenness centrality would have more control over the network, because more knowledge will pass through that node. Although betweenness centrality does not appear as distinctly as degree centrality in the visualization above, it is a meaningful way of identifying the nodes that are highly likely to affect the flow of communication within the network.

Table 6 shows the results of centrality analysis according to ASJC codes. The ASJC with the highest degree of centrality is 2700 (Medicine), which is currently most directly related to the surrounding areas. However, the ASJC codes in the field of computer science such as 1706, 1710, 1705, 1712, 1702, 1703 have a high degree centrality overall, which indicates that this is the field currently being studied most actively. In addition to the field of computer science, ASJC codes with a high closeness centrality were found to be in the fields of social science such as 3315, 3308, 3305 and mathematics such as 2611, 2604 and 2613. Therefore, if innovations occur in these fields, we can anticipate that the impact on other disciplines will be most rapid. The majority of the nodes that ranked high in betweenness centrality were found to have a high ranking in degree centrality, but an interesting finding was that a few nodes had a low degree centrality ranking. This was the case for 3305, which was ranked 3rd (the DC ranking: 42nd), 3315, which was ranked 11th (44th), 1406, which was ranked 12th (41st), 2748, which was ranked 14th (114th), 1901, which was jointly ranked 19th (89th), 1903 (89th), and 2307 (62nd). Each of these codes respectively corresponds to the following classifications: Geography (3305), Planning & Development, Communication (3315), Marketing (1406), Urology (2748), Earth & Planetary Sciences (1901), Computers in Earth Sciences (1903), Health, Toxicology & Mutagenesis (2307), Tourism, Leisure & Hospitality Management (1409), and Management, Monitoring, Policy & Law (2308). Such cases in which the betweenness centrality was found to be higher rank compared to that of the degree centrality within the science and technology field were cases which had received hardly any attention in the network visualization results. Although these fields are not the ones that are the most immediately noticeable when we observe the degree centrality results, we can regard them as areas that perform the role of bridge among other fields and thus areas in which we expect to see future development of research using Google Trends.

Table 7 shows the top 20 keywords with the highest centralities out of the 1,434 author keywords. In addition to Google Trends, Google, and Google Insight, the keywords with the greatest relevance to the topic of the study such as Search engine, Internet, and Internet search were found to have high centrality values. Other keywords such as forecasting, Nowcasting, Big data, Twitter, data mining, and text mining are all in the top rankings in terms of centralities. In other words, we expect that the research efforts to perform forecasting by mining big data will continue attracting attention in the present and future. Keywords with particularly high degrees of centrality included Google Flu Trends, Infodemiology (23rd), and Surveillance (24th), which are related to communicable diseases: these were found to be areas of continuing high interest, in linkage with other keywords. Meanwhile, time series analysis, election, tourism, and suicide were some of the keywords with high closeness centrality, and these

Table 6
Centrality analysis results for ASJC.

Rank	Weighted degree centrality	Closeness centrality	Betweenness centrality	
			ASJC codes	Rank in DC
1	2700	2700	2700	1
2	1706	1706	1706	2
3	1710	2611	3305	42
4	1705	1712	2002	9
5	1712	3315	1710	3
6	2611	2213	2613	15
7	1702	2604	2739	11
8	1703	2613	1712	5
9	2002	1710	2611	6
10	2604	1703	2725	18
11	2739	3308	3315	44
12	3312	3305	1406	41
13	1802	1704	2213	16
14	2208	1705	2748	114
15	2613	1312	1705	4
16	2213	2718	3312	32
17	2614	2204	2003	35
18	1711	2805	2303	15
19	1704	1702	2307	89
20	2207	1201	1903	89

Note: DC is an abbreviation for “Degree Centrality.”

Table 7
Centrality analysis results for author keywords.

Rank	Weighted degree centrality	Closeness centrality	Betweenness centrality	
			Author keywords	Rank in DC
1	Google Trends	Google Trends	Google Trends	1
2	Google	Google	Google	2
3	Search engine	Internet search	Internet search	5
4	Internet	Search engine	Search engine	3
5	Internet search	Internet	Internet	4
6	forecasting	forecasting	data mining	7
7	data mining	data mining	forecasting	6
8	Information retrieval	Twitter	Google Earth	19
9	Google Insights	Google Insights	Google Insights	9
10	Twitter	Text mining	Information retrieval	8
11	influenza	Social media	Social networks	17
12	Big data	Big data	influenza	10
13	web	Time series analysis	cloud computing	39
14	Google Flu trends	influenza	Google map	20
15	Social media	election	Service Oriented Architecture	91
16	nowcasting	web	Big data	12
17	Social networks	nowcasting	Text mining	21
18	queries	epidemiology	Investor attention	30
19	Google Earth	Tourism	Twitter	10
20	Google map	Suicide	Volatility	26

Note: DC is an abbreviation for “Degree Centrality.”

correspond well with the mathematics and social science fields mentioned in Table 6. Specifically, if there is an innovation in the technique of time series analysis for search traffic or in the methodology of using search traffic for social issues such as election, tourism, and suicide, such innovation can be expected to have the fastest ripple effect on other research subjects. As with ASJC, most of the nodes with high ranking in betweenness centrality were also ranked high degree centrality, but there were a few nodes that had a low ranking for degree centrality. Service Oriented Architecture, which is ranked 15th (91st) is a leading example, and although this isn't shown in Table 7, other cases which were notable because the betweenness centrality was relatively high included the following: Information search which was ranked 24th (364th), cross correlations, ranked 26th (139th), Google search volume index, ranked 28th (139th), CBIR (Content-based image retrieval), ranked 29th (215th), and Web 2.0, ranked 31st (139th). The keywords that have this characteristic can be divided into two types. One type consists of keywords such as Service Oriented Architecture, CBIR, and Web 2.0 represent new technologies, in the case of which the subject of analysis, such as influenza, received much attention as the search keyword. Another type consists of keywords that are related to the features presented in Google Trends, such as information search, cross correlations, and Google search volume index. The nature of the statistics provided by Google Trends, as explained previously in Section 2, is that the results of the activities related to the consumer's information search come to constitute the factor that links various studies. Furthermore, the attempt to explain society by organizing these results into an index also becomes an intermediary factor that connects research. From these attempts, we derive the insight that many studies come to take into account the cross correlations to explain social phenomena. Research on the retrieval behavior will continue in the future, and our findings underscore that research such as index development that utilizes the unique characteristics of Google Trends present new possibilities.

We identified the directions of development in research related to Google Trends through the SNA results presented in Tables 6 and 7. We attempted to improve the appropriateness of these prediction results by comparing them with those of a conventional meta-analysis-based approach (time series). Therefore, we compared our analysis of centralities used to determine the ASJC and keywords that are likely to develop in the future with the conventional time series analysis results. First, we analyzed the growth rates of 168 ASJCs used in classification in 657 papers. We analyzed the growth rate of the research in a given field based on the time series data on the number of paper publications in each ASJC. The growth rate was determined by the slope of the regression line of the time series data.

The results showed that there was a rapid growth in the number of papers published in the fields of 2700 (Medicine), 2002 (Economics and Econometrics), 2739 (Public Health, Environmental and Occupational Health), and 1705 (Computer Networks and Communications). Table 8 shows the correlation between the growth rates and the various centralities derived through SNA. The growth rate was found to have the highest correlation with betweenness centrality. As described above, a node with a high degree of betweenness centrality acts as a bridge in the network and has strong control over the flow. This means that such nodes facilitate the flow of knowledge by linking various disciplines and helps them to create a new field of converged research. We can thus understand why research in these fields is more active and the growth rate is high.

Table 8
Correlation between the growth rate of the number of articles and the centralities according to ASJC.

	1	2	3	4	5
1. The growth rate	1				
2. Degree centrality	0.679**	1			
3. Weighted degree centrality	0.669**	0.915**	1		
4. Closeness centrality	0.344**	0.588**	0.487**	1	
5. Betweenness centrality	0.806**	0.788**	0.678**	0.346**	1

** $p < 0.01$

Table 9
Correlation between the growth rate of the number of articles and the centralities according to keywords.

	1	2	3	4	5
1. The growth rate	1				
2. Degree centrality	.827**	1			
3. Weighted degree centrality	.834**	.997**	1		
4. Closeness centrality	.168**	.218**	.206**	1	
5. Betweenness centrality	.872**	.943**	.947**	.135**	1

** $p < 0.01$

This tendency can be confirmed by comparative analysis of the author keywords. Next, we analyzed the time series data of the number of papers which included the 1,463 keywords mentioned in 657 papers and analyzed the growth rate of research related to the keywords. Table 9 shows the correlation between the centralities and the growth rate of the number of articles for the respective author keywords. We observed the same tendencies as found in the case of ASJC. The keywords with high betweenness centrality such as Google Trends, internet, forecasting, Google, data mining, and internet search showed high growth rates. Aside from words that are generally used when dealing with the subject of this study, such as Google Trends, Internet, and Google, the most notable keywords were forecasting, data mining, and internet search. These research keywords are expected to gain prominence across all disciplines, including information system and computer science, medicine and bio science, economics and business.

4.4. Results of the time-series analysis of literature on Google trends

4.4.1. Changes in the keywords over time

Fig. 5 presents the two-mode network analysis results based on year of publication and author keyword, and with the exception of the major keywords that are shared in the center, we observed that by year, various smaller keywords existed in abundance. The keywords located in the center included Google Trends, search engine, Google, internet, data mining, social media, information retrial, forecasting, nowcasting, Big data, influenza, and Twitter. These search words were those that already occupied the top rankings in frequency.

Table 10 shows the results of community analysis performed to quantitatively examine the changes in the keywords by year and the number of communities was found to be 12 (including the year 2017).[6] The derived communities were organized by the publication year, and Table 10 shows the keywords that were found to have a class weight of 3 or more. In the early stages, there were many keywords related to the information system or computer science, but from 2011 onwards, there was a rise in keywords for areas of application such as epidemiology, public health, children, evaluation, popularity, stock return, and unemployment. From 2013 onward, the range of areas further expanded and we began to see keywords such as tourism, election, environment, and global financial crisis in key positions. In 2014, there were more additional keywords including investor attention, public opinion, and climate change. The keywords that gained new attention since 2016 are those related to big data utilization, applications and analysis, including Big data, Twitter, text mining, social media, cloud computing, ontology, and Android. These results allow us to deduce that as studies that utilize Google Trends expanded in the range of fields, and then combined with big data applications, there has been an emergence of analyses that are

[6] 12 communities were derived from the results of the two-mode analysis of 12 years (with the default resolution set at 1.0).

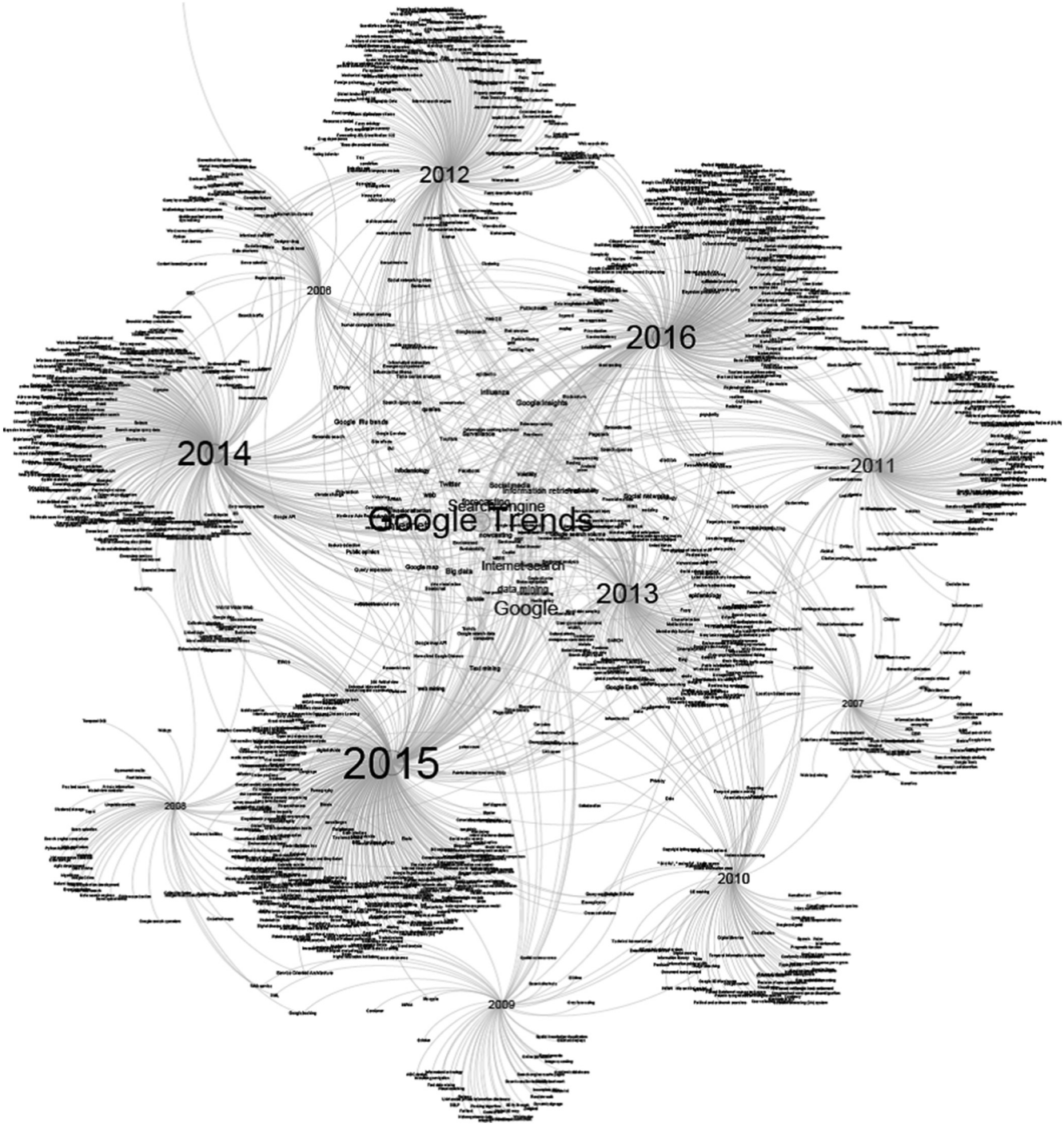

Fig. 5. Two-mode network visualization of author keyword and publication years.

interlinked with other media.

The total weight, the number of linked keywords, in Table 10 shows the time-series characteristics associated with Google Trends. The trend of the publication of related literature by year in Fig.1 and the change of total weight of year in Table 10 shows a similar tendency, while the decrease in total weight in 2016 in Table 10 is relatively small. We have shown that a new big data source, Google Trends, has been successfully expanded to various fields over the past decade.

4.4.2. Patent analysis and SNA

Most commonly, in the basic research stage researchers focus on analysing paper publications while in the development stage, patents are the target of analysis (Jun, 2012a; Watts and Porter, 1997). Our study also builds on the results of these preceding bibliographical studies (Daim et al., 2007; Martin and Daim, 2007; Winthrop et al., 2002), and analyzes patents, which are representative indicators in the development stage. As outlined above, the purpose of this analysis is to derive implications regarding the impact of research (papers) related to Google Trends, by comparing research activity and patent activity related to Google Trends in the socio-technical system.

We conducted a simple analysis of 76 U.S. patents that cite Google Trends (USPTO, 2017).[7] First, we derived some observations from the

[7] Patent information was collected from the USPTO (USPTO, 2017) and the patents

Table 10
Community analysis results for author keyword & publication year.

Year (community class)	Author keywords	Total weight
2006	Queries, page rank, data analysis, Google API, human computer interaction, information seeking, semantic web, clustering, data integration	103
2007	Google Earth, location based service, privacy	101
2008	Google map, query expansion, World Wide Web, semantic, page rank	91
2009	Feature selection, Google map API, Google Scholar, service oriented architecture	96
2010	Time series, digital divide, Information search	134
2011	Information retrieval, epidemiology, public health, search queries, bing, children, evaluation, personalization, popularity, stock return, unemployment, Web 2.0	228
2012	Search engine, epidemic, Google search, search query data, influenza like illness, information extraction, semantic search, sentiment, social networking sites	314
2013	Data mining, Google Insights, nowcasting, Google search volume, predictability, web mining, tourism, election, environment, GARCH, global financial crisis, information seeking behavior, sentiment analysis, sustainability, user generated content, webometrics	330
2014	Internet, influenza, Google flu trends, surveillance, infodemiology, investor attention, public opinion, time series analysis, Facebook, prevention, ARIMA, climate change, epilepsy, search traffic	437
2015	Google, internet search, forecasting, big data, Twitter, text mining, suicide, seasonal, computers, ethics, google search data, trends	580
2016	Social media, web, social networks, volatility, cloud computing, ontology, android, Bayesian prediction	482

Note: the weight means the number of links of an item.

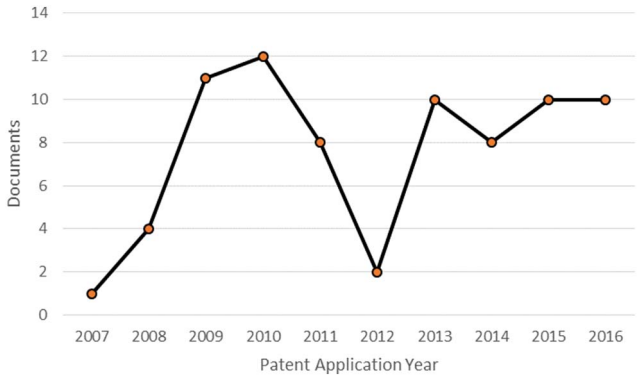

Fig. 6. Application trends of patents related to Google Trends.

time series changes (the number of applications). The trend in the applications for patents showed a slight difference in temporal trend from the time series data for paper publications shown in Fig. 1, Fig. 6 shows that the time series trend of patent applications corresponds to a typical hype cycle (Fenn and Raskino, 2008). In science, technology and innovation studies, expectations can be defined as "real time representations of future technological situations and capabilities" (Borup et al., 2006; van Lente et al., 2013). Expectations play an important role in determining the direction of technological changes and the speed of innovation adoption (Alkemade and Suurs, 2012; Berkhout, 2006; Mercer, 1997). Such expectations can also set the agenda for innovation and may establish legitimacy for investment. The performative character of expectations has important implications for the study of hypes (Alkemade and Suurs, 2012; van Lente et al., 2013). Meanwhile hypes are often seen as something deceptive, incorrectly exaggerating the impact and outcome of an independent technological development (Ruef and Markard, 2010; van Lente et al., 2013). In general, hypes inevitably lead to disappointment. When high expectations fail to be fulfilled by actual performance accomplished by innovative efforts, the disappointment has the effect of reducing the positive expectation (Brown, 2003; Konrad, 2006). Once expectations are thus lowered, it will gradually recover, in accordance with the speed of the actual realization of innovations (Fenn and Raskino, 2008). This cycle of hope and disappointment is so common that Gartner have given it a name, the hype cycle, because all the initial enthusiasm is built mainly on expectation and hype (Fenn and Raskino, 2008; Jun, 2012a; van Lente et al., 2013). The hype cycle consists of five phases: the Technology Trigger phase, the Peak of Inflated Expectations phase, the Trough of Disillusionment phase, the Slope of Enlightenment phase, the Plateau of Productivity phase (Fenn and Raskino, 2008).

Unlike the cross-sectional trend of the paper publications shown in Fig. 1, patents applications have been filed intensively since 2008. Fig. 6 can be divided into two phases in the manner of the hype cycle. The first phase is the cycle from 2007 to 2011 and the second phase is the period after 2012. The first period is similar to the period in which the hype cycle unfolds through the Technology Trigger and Peak of Inflated Expectations and Trough of Disillusionment, while the second phase is similar to the cycle leading through the Trough of Disillusionment, Slope of Enlightenment and Plateau of Productivity (Fenn and Raskino, 2008). If this phenomenon is the result of basic research activities such as the results obtained in preceding studies, however, then the first phase in which patent applications soared following the launch of Google Trends will be difficult to explain as result of research related to Google Trends shown in Fig. 1. As mentioned in Section 2.1, for decades before Google Trends was released, there had already been numerous studies on consumer search behavior and consumer adoption models (Kotler et al., 2014; Kotler and Keller, 2008; Rogers, 2003) and researchers had already discussed the possibility of analysis using search data once search engines become popularized (Ettredge et al., 2005; Moe, 2003; Zimmer et al., 2007). Therefore, in Phase 1, it appears more appropriate to explain that such basic research were the result of ideas or development studies that were stimulated by the launch of Google Trends. Also, the second phase of patent applications from 2013 can be explained as the result of the expansion of research reflected in the previous SNA results, which led to the identification of new commercial possibilities. Fig. 1 shows the increased research activity from 2011 and a surge in basic research outcomes from 2013, and this can explain the increase in patent applications from 2013, which is visible in Fig. 6.[8] We also need to pay attention to the downturn between the two phases: as for the period of decline between 2011 and 2013, it may be seen in light of how the Trough of Disillusionment is usually explained in the hype cycle: it is likely that interest in the results

(*footnote continued*)
included in this study were those that mentioned Google Trends in the title, abstract, or claims and description. We also limited our analysis of patents to USPTO, because, as shown in Table 3, the research related to Google Trends are concentrated predominantly in the U.S. Also, this restriction allows us to avoid the potential problem of duplicate applications that may exist in the patent offices of other countries.

[8] As for the identity of the patent applicants, in Phase 1, the overwhelming majority of the applications were filed by the brothers, Chad and Ryan Steelberg (2009～2010). The Steelbergs founded dMarc Broadcasting, which is said to have been acquired by Google in 2006 (Goolge_Press, 2006). In Phase 2, companies such as Affectiva Inc. and Sirius XM Radio Inc., which are involved with not only Google but also Google Android Apps, applied for multiple patents from 2011.

Table 11
Results of patent CPC first appearance timing analysis.

CPC sections	CPC first application date	Documents	Code description
G06Q 10/04	2007.09.14	1	Forecasting or optimisation, e.g. linear programming, "travelling salesman problem" or "cutting stock problem"
G06Q 30/02	2008.01.18	28	Marketing, e.g. market research and analysis, surveying, promotions, advertising, buyer profiling, customer management or rewards; Price estimation or determination
G06F 17/30	2008.10.06	25	Information retrieval; Database structures therefore
G06F 15/173	2009.02.09	1	Using an interconnection network, e.g. matrix, shuffle, pyramid, star, snowflake
G06Q 10/063	2009.10.22	2	Operations research or analysis
H04H 60/46	2010.05.17	1	for recognizing users' preferences
G06Q 99/00	2010.07.12	1	Subject matter not provided for in other groups of this subclass
H04N 21/4826	2010.11.03	1	Using recommendation lists, e.g. of programs or channels sorted out according to their score
H04N 21/4223	2011.07.21	1	Cameras
G06Q 30/00	2011.07.27	1	Commerce, e.g. marketing, shopping, billing, auctions or e-commerce
H04L 43/10	2011.11.14	2	using active monitoring, e.g. heartbeat protocols, polling, ping, trace-route
G06T 11/60	2013.12.06	1	Editing figures and text; Combining figures or text
G06K 9/00302	2014.03.12	1	Facial expression recognition
G06N 5/04	2014.03.12	2	Inference methods or devices
G06Q 30/0185	2014.04.09	1	Product, service or business identity fraud
H04L 51/32	2014.07.29	1	Messaging within social networks
H04L 51/28	2014.09.12	1	Details regarding addressing issues (arrangements and protocols for addressing and naming
G06Q 20/145	2015.05.11	2	Payments according to the detected use or quantity
H04N 21/4668	2015.08.10	1	for recommending content, e.g. movies
G06Q 30/04	2016.02.22	1	Billing or invoicing
A61B5/165	2016.12.16	1	Evaluating the state of mind, e.g. depression, anxiety

of the development studies based on the existing basic research "waned as experiments and implementations failed to deliver" or that the "producers of the technology failed."

These claims can be verified through an analysis of the timing of the first appearance of the cooperative patent classification (CPC) of Google Trends related patents.[9] Table 11 shows the results of the analysis of the timing of the first appearance of CPC for the 76 patents related to Google Trends. In the beginning, there are two major areas that led to patent applications, namely Marketing (e.g. market research and analysis) and Information Retrieval. As described above, in these two areas researchers had been actively engaged in basic studies regarding information search behavior, consumer adoption, and search engines. Therefore, we can plausibly explain that the preparatory foundation set by such basic research allowed development stage technology (patent) to be achieved quickly after the launch of Google Trends. This view is corroborated by the case of Medical & Bio Science: we introduced in Section 2.1 that prior to the launch of Google Trends, there were only few studies related to information search behavior in the field of Medical & Bio Science. Table 11 shows that such a field only emerged gradually after 2012, in the second phase indicated in Fig. 6. Since 2014, various methods of emotional analysis and ideas about linked projects have been introduced, and patents also began to be filed in the field of medical & bio science studies, which have been actively researched for paper publications (H04L 51/32 field). CPC related to epidemiology, influenza, sentiment, and social networking, which were the prominent keywords prior to 2012, are found increasingly in commercialization attempts from 2014 onwards, as shown in Table 11, which can be compared with Table 10 in particular (relevance of individual patents was confirmed based on titles and abstracts). In addition, the keywords mentioned above are relatively high in terms of betweenness centrality according to Table 7. This confirms the importance of these keywords and the future-oriented tendency of betweenness centrality. The patent analysis thus corroborated the claims presented in this study regarding the changes in the research direction predicted by SNA and the possible corresponding socio technical variables such as patents.

[9] The analysis of the timing of the appearance of cooperative patent classifications refers to the method of analysing when a classification first emerged in the related dataset.

5. Status of research in each major field and implications

When we synthesize the results in Fig. 4 and Tables 4–5, we find three areas in which the distinction in the direction of research (keyword and scientific classification) has been notable thus far: 1) Information system or computer science, 2) Medicine & bio science, 3) Economy & Business and other fields such as Finance & Ecology Management and Policy & Law. Specifically, when we observe Fig. 4 in relation to Table 10, of course, we find that the early studies related to Google Trends tended to be in fields related to information systems or computer science but more recently there has been more activity in studies utilizing big data, and accordingly, the range of fields has gradually expanded to medicine and economics. Accordingly, in the following, we will review the important research results yielded from the three areas discussed above, and derive insights related to big data utilizations and applications.

5.1. Information system and computer science fields

Considering that Google Trends is data regarding search engine utilization, we can obviously understand why this field is highly relevant. Research related to Google Trends became active as soon as the Trends service was launched, in continuation of the research on search engine related technology development and web search service development.

Early on, Waller (2011) noted the potential uses for the regional and category information provided by Google Trends and argued that the search engine is not only an interface for information or a shortcut to accessing websites, but will also become a site of leisure. It was argued that there was no observed statistically significant difference in search type or topics for different segments of the online population (Waller, 2011). Prakash et al. (2012) claimed that when there are two competing products spreading across a given network, the information provided by Google Insights can be used to predict which product will win in terms of market share (Prakash et al., 2012). Vaughan and Romero-Frias (2014) used the information supplied by Google Trends in U.S. English and Spanish to test whether the data can predict academic fame. This study showed that there is a significant correlation between the search volume for a university's name and its academic reputation or renown (Vaughan and Romero-Frías, 2014). In addition, Vaughan and Chen (2015) extended his research beyond the earlier comparative study of English and Spanish using Google to a comparison

of the search volumes of the search engines that handle different web query data sources. A comparison was made of the search results of the Baidu Index with those of Google Trends, which boasts the world's highest usage. Despite the differences between the two services, such as different language processing methods, the search volume data of the two languages were highly correlated and combining the two data sources did not improve the predictability of the data. It was pointed out, however, that one limit of Google Trends is that it provides less information about search volume than the Baidu Index. According to Vaughan and Chen, this because the number of people who use Google in China is relatively small (Vaughan and Chen, 2015). In the same vein, there is a study that provides a comparison between Google Trends and Naver Trend, which is the statistics provided by one of the main search engines used in Korea. Jun et al. (2017) introduced the analogical method for demand forecasting which compares the data from Google Trends and Naver trend and performs analogies. In this process, there was also a significant correlation between the search volume of Google Trends and the search volume of Naver Trend for an identical product, and it was argued that there was a time lag in the search volume for some of the products (Jun et al., 2017).

Not all of the discussion about Google Trends focused only on its positive aspects. Vaughan and Yang (2013) selected groups of universities and businesses from the U.S. and China and collected their web traffic data from three sources: Alexa Internet, Google Trends for Websites, and Compete. They found that the web traffic data indicated a significant correlation between a university's academic quality and corporate financial variables. These researchers reviewed the characteristics of the three data sources and compared their usefulness, and concluded that the web search data from Alexa Internet was the most useful (Vaughan and Yang, 2013). In other words, the results indicated that depending on certain conditions, Google Trends did not perform as well at predicting these variables compared to other sites.

Looking at the above studies, we see that rather than focusing only on utilizing Google Trends data, many researchers have evaluated and compared Google's performance as a search engine and identified areas for improvement. Some studies have focused on the search engines and compared Google with other sites in terms of the search data they provided and actively compared it to search engines used in other languages and countries. These studies thus assess the limits and potentials of the search information provided by Google Trends from various perspectives. The most significant feature of the information system and computer science fields is that users of Google Trends are more interested in Google search than any other specific services or sources and are more interested in technologies as such than in utilizing the data.

5.2. Medicine and bio science fields

As mentioned in previous studies, it was the study by Ginsberg et al. (2009) that initiated the interest of researchers in the potentials of utilizing Google Trends and this study was closely related to medical science. Ginsberg et al. (2009) presented a model that can predict current flu levels based on an analysis of data from Google's past search queries, and this stimulated further research using Google's search information. This study by Ginsberg et al. developed a computer model that used the raw search data before processing and converted it into a real-time monitoring system that predicts the activity of influenza viruses one week to two weeks earlier than the report released by the Centers for Disease Prevention and Control (CDC) (Ginsberg et al., 2009). By presenting the potential uses of Google search information, this study also influenced the use of other SNS information such as Twitter (Cha et al., 2010). Carneiro and Mylonakis (2009) also acknowledged that Google Trends shows great promise as a timely, robust, and sensitive surveillance system, argued that it will have a higher potential for use in advanced countries where there is a larger population of search users (Carneiro and Mylonakis, 2009). Carneiro and Mylonakis (2009) thus drew our attention to an important point, namely that when utilizing search traffic, it is important to understand the characteristics of the users, such as the size of the user population. The ability to monitor disease using Google Flu Trends proved to have a very powerful influence: Pelat et al. (2009) were also inspired to investigate the possibility of monitoring other diseases other than influenza in France. They found that for each of three infectious diseases, just one well-chosen query was sufficient to provide a time series of searches highly correlated with incidence. They also demonstrated that once the internet search keyword is chosen well, it was even possible to perform a surveillance of acute diarrhea and chickenpox in a non–English-speaking country (Pelat et al., 2009). Furthermore, Seifter (2010) found Google Trends to approximate certain trends previously identified in the epidemiology of Lyme disease (Seifter et al., 2010). There were also studies showing that Google search query data can be used to monitor dengue activity in various countries including Bolivia, Brazil, India, Indonesia and Singapore (Althouse et al., 2011; Chan et al., 2011).

As we have seen, there was an avid and positive response from the medical field to the feasibility of surveillance and monitoring using Google Trends, but in 2013, a paper was published that refuted the claims of Ginsberg's monumental study. According to Butler (2013), compared to the data from the Centers for Disease Control and Prevention (CDC) based on laboratory surveillance reports from all across the U.S., Google Flu Trends (GFT) reported the percentage of visits to influenza-like illness (ILI) to be more than two times greater (Butler, 2013). The fact that this happened even though GFT was created to predict the CDC report indicates that forecasts based on Google search data may not be accurate, according to this study. A series of studies began to point to these limitations (Olson et al., 2013), and Lazer et al. (2014) further pointed out that this problem is not limited to GFT, and raised fundamental questions about what kinds of phenomena search data or social media can predict (Lazer et al., 2014). It was argued that although previous studies have shown some of the values of search information, the analysis results of social big data (including web search data) are not reliable enough to replace conventional methods, which remain valuable. They pointed out that search information on granular views ought to be measured and provided for more accurate predictions and that we need to understand the process by which the provided search information as generated. The algorithms underlying Google, Twitter, and Facebook help determine what we find out about our health, politics, and friends (Lazer et al., 2014). These scholars acknowledge the strengths of analysis and utilizations of big data such as Google Trends but warned against hubris in its utilization.

Among the three main areas for Google Trends that we identified in this study, the medicine and bio-science field is a relatively new one in terms of search information and utilization, and as such, we can regard it as a field that has been strongly inspired by Ginsberg et al. (2009). This influence is shown in the trends visible in Table 10 and Table 11, where we can see a surge in related studies after 2011. This field also Fig. 1 contributed significantly to the increase in research related to Google Trends since 2011. In the Medicine and bio science fields, attempts are made to perform various analyses using the Google Tends service or the data itself, but they are relatively averse to errors due to the nature of the field.

5.3. Economy, business and other fields

In the economy and business sectors, as we examined above in preceding studies, research on information search behavior was active even before Google Trends was launched. Unlike the information system and computer science field, however, the research here was more actively focused on the application of the data rather than the search engine, and as an extension of such efforts to utilize and apply Google Trends data, research in this field became active right away following the launch of the Trends service.

The possibility of monitoring macroeconomic phenomena has also been actively studied in many papers focusing on the unemployment rate and economic indices. Askitas and Zimmermann (2009) were interested in Google search information as a means of quickly monitoring economic crises (Askitas and Zimmermann, 2009). They used monthly German-language data and found a strong correlation between keyword search and unemployment, and proposed a way to monitor and forecast changes in economic conditions. Similar research has been performed by Google as well (Choi and Varian, 2009). Da et al. (2011) proposed a new method of directly measuring investor interest using search frequency (Da et al., 2011). They proposed the Google Search Volume Index (SVI): SVI proved to be a strong indicator of investors' interest, and allowed us to quickly identify investors' interest. In particular, SVI performed well in predicting the stock price trend for the next two weeks and the initial profit of IPO (initial public offering) stocks. Vosen and Schmidt (2011) compared the MCSI (University of Michigan Consumer Sentiment Index) and the Consumer Confidence Index, which indicate trends among U.S. consumers, with the search traffic results provided by Google Trends and demonstrated that the predictive power of search traffic provided by Google can be higher than the survey-based index (Vosen and Schmidt, 2011). This study provided empirical proof that Search traffic may show a forecasting power comparable to a survey-based microanalysis in personal consumption forecasting. Tefft (2011) claimed that Google Insights (Trends) could be used to predict unemployment related information (Tefft, 2011). Preis et al. (2013) analyzed the changes in the number of Google searches using financial search terms and suggested patterns that could be interpreted as "early warning signals" of stock market movements (Preis et al., 2013). Such forecasting of unemployment or on economic crises using Google Trends remains an attractive theme for research (Vicente et al., 2015).

In addition to these macroeconomic studies, there has also been active research in the more microscopic management fields. In particular, there have been many studies on direct forecasting based on nowcasting capabilities. Goel et al. (2010) noted that consumers searching online may allow us to predict their collective future behaviour (Goel et al., 2010). They argued that search query volume could be used to forecast the opening weekend box-office revenue for feature films, first-month sales of video games, and the rank of songs on the Billboard Hot 100 chart. They concluded that the search query provides a useful guide to the near future especially if there is no other data source or forecasting performance has not been significantly improved otherwise. Choi and Varian (2012) also demonstrated the possibility of using search traffic for direct predictions (Choi and Varian, 2012). They found that Google Trend is especially useful for forecasting economic activity. They describe car sales, home sales, retail, and travel as examples of economic activities that can be predicted using Google Trends. They argue that this data is useful to predicting the closer events rather than the distant future, because this helps identify the "turning point" that occurs suddenly in the market. If there is a surge in searches for "real estate agents" in a particular location, it is possible to predict that housing sales will increase in this location in the near future. Jun and his colleagues also argued that using search traffic information provided by Google Trends could help determine the likelihood of a consumer's adoption of technology or the purchasing a product, and could even help analyze the preferred technology or product specifications (Jun et al., 2014a; Jun et al., 2014b; Jun and Park, 2016). Jun et al. (2017) also suggested that by performing a network analysis of search words that are concurrently searched, researchers can draw up a position map that shows how consumers think about a brand or product (Jun and Park, 2017).

In addition to these areas of business management and economics, there have also been attempts to use Google Trends in political fields. Jun et al. (2016) argued that the public's interest in public policy can be analyzed using Google Trends information (Jun et al., 2016). Mavragani et al. (2016) also claimed that public interest in micro pollution policy could be measured using search information (Mavragani et al., 2016).

As we have seen, among the studies in the fields of economics, business and policy that use Google Trends data, many studies attempted to understand people or society through search activities and thereby predict behavior. Not all of the results, of course, yielded optimism. The study by Lui et al. (2011) used Google Trends search traffic to analyze the winning chances of candidate running in US Congress elections in 2008 and 2010. Although this study was not in the field of economics or business it should be noted because the research results pointed out the limits of the forecasting power of Google search traffic. According to this research, the ability of Google search traffic to predict the victory of a candidate was not relatively higher compared to the conventional NYT polling survey (Lui et al., 2011). The decline in the forecasting ability of Google Trends in the competitive area was explained by the authors as attributable to the fact that there may be a lot of negative information searches for contenders in such contested areas. Interest reflected in the searchers can be both positive and negative.

To overcome these limitations, there have been more attempts to create an index to make relative comparisons that take account of other variables or attempts to use this data in conjunction with other sources of big data (Mavragani and Tsagarakis, 2016; Vosen and Schmidt, 2011). These developments in economics, business and other fields have contributed to the increase in research related to Google Trends, but one feature that distinguishes these fields from others is that they had a greater influence on application attempts, in development technologies and business initiatives.

6. Discussion

One purpose of this study was to review the trends in the development of papers related to Google Trends over the past decade and thus contribute to related research, but more importantly, our goal was to provide implications regarding research on the utilization and application of big data. In achieving this goal, Google Trends has fully demonstrated its advantages in terms of economy, immediacy and objectivity, and there has been an expansion in research areas using this source. As of now, Google Trends data is provided free of charge although there is a daily limit to usage. The provided information is updated daily, and the Hot Trends are updated every hour (Google, 2017). In addition, compared to surveys, limitations such as cognitive dissonance and construal level theory are relatively less of a problem, and the objectivity is relatively high, since the number of users is close to that of the population (Jun et al., 2014a).

When we review these studies that have utilized Google Trends after 10 years of rapid growth, one common observation we can make is that Google Trends information provides once distinct advantage, which is that Google Trends makes it easy to identify the current interests of searchers. In the case of search activities for important events such as illness, it is relatively easy to grasp the intent of the searches and therefore the relevant research was conducted more actively and quickly than in other fields. Also, in these fields, it is relatively inevitable that there will be information searching activity reflecting purchases and concerns, and therefore there continued to be expansion of research in the related fields. However, as the application fields expanded, there was also an increase in the problems that emerged. In areas where the intent of the search activity is unclear, errors may arise when observing this interest and predicting business change. This points to two aspects that require our attention: the two advantages of Google Trends that were the basis of its appeal may also be the source of its drawbacks. (The differences among searchers, search languages, and search areas have been already been discussed often in previous studies (Jun et al., 2017; Vaughan and Chen, 2015) and will not be further discussed here.)

The first advantage was the ability to monitor the use of information searches, but it should be noted that the behavior of searching for information does not occur in the case of all products or events. As

explained above, when there is a higher perception of risk regarding the product (Beatty and Smith, 1987), or in cases of high involvement products or high priced products (Assael, 1992), there will be active searches. This is why, in cases involving medicine, health or the purchase of high priced products, there was a higher chance of successfully utilizing Google Trends. Of course, even for high-involvement products, we may observe behavior in which it is difficult to clearly grasp the intent of the search in Google Trends data. Jun (2012) used Google Trends to explain Gartner's hype cycle, claiming that Google Trends can account for the tremendous increase in users' interest and rapid shift to indifference in the process of adopting new technologies (Jun, 2012a; Jun, 2012b). Of course, this phenomenon does not always occur (Dedehayir and Steinert, 2016), and it is clear that overcoming the irrational aspect of search activity is an issue that must not be overlooked when expanding the use of Google Trends. Even when analyzing information provided by Google Trends, to obtain a clearer understanding of the intention of utilization, we will also need to be analyzed other social big data as mentioned above (Lazer et al., 2014).

Secondly, one of the advantages that made Google Trends highly appealing was that it was relatively free of problems such as cognitive dissonance that is the drawback of surveys (Jun et al., 2014a), but we should not overlook the risk that Google Trends data may reflect irrational tendencies. The speedy propagation of users' emotional responses and negative news (or gossip) is a typical example of irrational use of web searches, and it is also a limitation when utilizing Google Trends. Therefore, in many cases, such as elections, when the intent of the search is unclear or the search is driven by emotions (popular vote, etc.), there may be limits to the usefulness of this data. Of course, even though we acknowledge the limitations of using Google Trends data in cases of emotional involvement in searches or in the aforementioned case of low involvement products, we should also note that recent studies have sought various methods of overcoming these limitations. The study by Lui, et al. (2011) referenced above explained the limitations of Google Trends in election results forecasting (Lui et al., 2011), but the more recent research by Mavragani and Tsagarakis (2016), demonstrated outstanding forecasting power in regards to the 2015 Greek Referendum results by analyzing data from Google Trends on the 'YES' and 'NO' search terms (Mavragani and Tsagarakis, 2016). This underscores the importance of selecting the search terms: the analytic limitations can be overcome depending on how the search terms are selected. However, since Google Trends only provides processed information, it poses a limit on the selection and refinement of search terms, and this acts as a limitation on the utilization of Google Trends. Especially, as shown in Tables 10 and 11, sentiment analysis becomes all the more important when the application is expanded to sociology and political science.

From such research using Google Trends, we can draw some implications regarding the analysis of big data source such as Google Trends. First, regarding the fields of application, we have seen that this big data has already been used to analyse social variables across various fields including IT and communications, as well as medicine, health, business management, and economics and hereafter, we can anticipate seeing applications in additional fields such as geography, planning & development, communication, earth & planetary sciences and management, monitoring, policy & law. Second, whereas in the past the purpose of the analysis had mainly focused on describing and diagnosing research and business trends through surveillance or monitoring, the purpose is now shifting toward forecasting. Thirdly, various alternatives have been proposed, but to perform forecasting, additional analyses such as sentiment analysis and cross correlation will be required. Fourthly, rather than pursuing an infinite expansion of the fields of utilization, we should observe that there has already been a certain degree of exploration of the suitably applicable fields (e.g. high involvement products) and now we are rather seeing more specific subdivisions and diverse sources (SNS) are being analysed together to overcome various limitations. Fifthly, to conduct precise analysis using big data, access to raw data should be expanded. As examined above, it is impossible to utilize Google search information to its full advantage just by selecting a search term. For example, at present, it is impossible to find out what other search terms were used before or after a particular search term (only the top ranked ones are provided). This means that only Google itself has adequate access to all the utilizable Google search information. Ultimately, this points to the need to discuss the nature of big data as a public resource.

7. Conclusion

One great advantage of Google Trends is that it collects big data, processes the information to facilitate analysis, and even releases this information for free. Therefore, Google Trends is a prime candidate for showing the possibilities and limitations of utilizing big data. In addition, the trends in research using Google Trends provide significant insights on how big data utilizations and applications is evolving.

According to the results of Google Trends case studies, there are a large number of fields in which we can utilize big data. As we have seen above, aside from the field of computer science and information systems, which created search engines and Google Trends, the field that first used this data most actively was the pharmaceutical field. If we can predict the outbreak of an epidemic, such as the flu, the government will be able to respond by taking measures to strengthen public health, and pharmaceutical companies will be able to manage the inventory of drugs more efficiently, and this will give them the opportunity to increase market share and profitability. However, unlike research trends, the patent application trends indicated that the first attempts at commercial utilization occurred in the field of Marketing & Business, which is likely since researchers in these fields were aware of the potentials of using web search statistics even before the launch of Google Trends. Of course, there are also cases of big data applications using sources other than Google Trends. Telecom companies, banks, and credit card companies are using big data to develop marketing tools that recommend services based on an analysis of customer consumption patterns. Manufacturers use big data to better manage their equipment and increase their service life, which in turn reduces costs. Now, however, it is no longer enough to understand actual events that have occurred. Now, we are facing the pressure to predict future events to achieve optimal performance for business. In this regard, the research fields that use Google Trends provide us with an important insight. In early 2010s, surveillance and monitoring were the key words in research, but in recent years, the keyword has expanded to include forecasting. As more people recognized this potential, the field of application expanded to include medical services and health, economy and business and it now even encompasses sociology, politics and law.

Though we have seen rapid growth in research on Google Trends, recently there have been many challenges in all three main fields. This phenomenon can be explained as part of the process of the hype cycle: once we pass the 'Peak of Inflated Expectations,' during which positive outlooks were the majority, we enter the 'Trough of Disillusionment' period characterized by many negative challenges, and in the end, it is by addressing these challenges that successful technologies and theories will emerge. Observing the recent challenges, we can derive implications for the success of research and development using Google Trends. In some of these challenges, we see that while some have sought new applications that use Google Trends, there are also warning signs that show us the danger of drawing overgeneralizations based on big data hubris. The utilization or analysis of big data is neither a panacea nor a technology that is mandatory for everyone.

To successfully utilize big data, taking pre-emptive measures such as taking caution in collection and the refinement of the data through filtering are important, but it is also critical to properly understand the characteristics of big data and create and use new values based on this understanding. In this respect, the goal of this study was to enhance the understanding of the limitations, possibilities, and characteristics of the

data that Google Trends provides and we expect it to make a significant contribution for creating new values in various areas using Google Trends in the future. For scholars, this study provides an overview of the studies that have used Google Trends in the past 10 years and points out the directions for the future, which will help guide other studies in a robust direction. From the centrality analysis, we anticipate that index research on social phenomena will become active in fields such as Geography, Planning & Development, Communication, which have drawn less attention so far but which have shown a relatively high betweenness centrality. We forecast that research related to topics such as election, tourism, and suicide which showed a high degree of closeness centrality in these fields will have a strong impact on other research.

The limitation of this study is that there may be missing or unnecessarily added research because we utilized SCOPUS data and limited the search to certain specific bibliographies. For example, some of the papers published between 2006 and 2008 that were included in the analysis in this study were also papers that had weak relevance to the Google trends service (refer to Table 2). The percentage of such cases, however, is not very significant, and we believe that this minor omission and inclusion of noise will not make much of a difference in our effort to diagnose the overall direction of research.

Acknowledgements

Seung-Pyo Jun, the first and corresponding author, acknowledge support from the Research Program at Korea Institute of Science and Technology Information. Furthermore, he thanks Sunhee Jung, Chanhyeok Jun and KISTI Technology Commercialization Center for the support of the research. Especially, these authors thank anonymous reviewers for their constructive comments. They could not complete this paper without the devoted comments of the reviewers.

References

Alkemade, F., Suurs, R.A.A., 2012. Patterns of expectations for emerging sustainable technologies. Technol. Forecast. Soc. Chang. 79, 448–456.
Althouse, B.M., West, J.D., Bergstrom, C.T., Bergstrom, T., 2009. Differences in impact factor across fields and over time. Journal of the Association for Information Science and Technology 60, 27–34.
Althouse, B.M., Ng, Y.Y., Cummings, D.A., 2011. Prediction of dengue incidence using search query surveillance. PLoS Negl. Trop. Dis. 5, e1258.
Askitas, N., Zimmermann, K.F., 2009. Google econometrics and unemployment forecasting. Appl. Econ. Q. 55, 107–120.
Assael, H., 1992. Consumer Behavior and Marketing Action, 4th ed. PWS-KENT Pub.
Bastian, M., Heymann, S., Jacomy, M., 2009. Gephi: an open source software for exploring and manipulating networks. ICWSM 8, 361–362.
Beatty, S.E., Smith, S.M., 1987. External search effort: An investigation across several product categories. J. Consum. Res. 14, 83–95.
Berkhout, F., 2006. Normative expectations in systems innovation. Tech. Anal. Strat. Manag. 18, 299–311.
Bettman, J.R., Park, C.W., 1980. Effects of prior knowledge and experience and phase of the choice process on consumer decision processes: a protocol analysis. J. Consum. Res. 7, 234–248.
Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E., 2008. Fast unfolding of communities in large networks. J. Stat. Mech: Theory Exp. 2008, P10008.
Borgatti, S.P., 2005. Centrality and network flow. Soc. Networks 27, 55–71.
Borup, M., Brown, N., Konrad, K., Van Lente, H., 2006. The sociology of expectations in science and technology. Tech. Anal. Strat. Manag. 18, 285–298.
Brown, N., 2003. Hope against hype: accountability in biopasts, presents and futures. Sci. Stud. 16, 3–21.
Butler, D., 2013. When Google got flu wrong. Nature 494, 155.
Carneiro, H.A., Mylonakis, E., 2009. Google trends: a web-based tool for real-time surveillance of disease outbreaks. Clin. Infect. Dis. 49, 1557–1564.
Cha, M., Haddadi, H., Benevenuto, F., Gummadi, P.K., 2010. Measuring user influence in twitter: The million follower fallacy. ICWSM 10, 30.
Chan, E.H., Sahai, V., Conrad, C., Brownstein, J.S., 2011. Using web search query data to monitor dengue epidemics: a new model for neglected tropical disease surveillance. PLoS Negl. Trop. Dis. 5, e1206.
Choi, H., Varian, H., 2009. Predicting Initial Claims for Unemployment Benefits. Google Inc.
Choi, H., Varian, H., 2012. Predicting the present with google trends. Econ. Rec. 88, 2–9.
Da, Z., Engelberg, J., Gao, P., 2011. In search of attention. The Journal of Finance 66, 1461–1499.
Daim, T.U., Rueda, G.R., Martin, H.T., 2005. Technology forecasting using bibliometric analysis and system dynamics, in: Technology management: a unifying discipline for melting the boundaries. IEEE 112–122.
Daim, T., Monalisa, M., Dash, P., Brown, N., 2007. Time lag assessment between research funding and output in emerging technologies. Foresight 9, 33–44.
Dedehayir, O., Steinert, M., 2016. The hype cycle model: A review and future directions. Technol. Forecast. Soc. Chang. 108, 28–41.
Ettredge, M., Gerdes, J., Karuga, G., 2005. Using web-based search data to predict macroeconomic statistics. Commun. ACM 48, 87–92.
Fenn, J., Raskino, M., 2008. Mastering the Hype Cycle: How to Choose the Right Innovation at the Right Time. Harvard Business Press.
Freeman, L.C., 1979. Centrality in social networks conceptual clarification. Soc. Networks 1, 215–239.
Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., Brilliant, L., 2009. Detecting influenza epidemics using search engine query data. Nature 457, 1012–1014.
Goel, S., Hofman, J.M., Lahaie, S., Pennock, D.M., Watts, D.J., 2010. Predicting consumer behavior with Web search. Proc. Natl. Acad. Sci. 107, 17486–17490.
Google, 2017. Google Trends Help, in, Google Inc. https://support.google.com/trends/?hl=ko#topic=6248052.
Goolge_Press, 2006. Google to Acquire dMarc Broadcasting in: Google News from Google. Goolge Press. http://googlepress.blogspot.kr/2006/01/google-to-acquire-dmarc-broadcasting_17.html.
Jun, S.-P., 2012a. A comparative study of hype cycles among actors within the socio-technical system: With a focus on the case study of hybrid cars. Technol. Forecast. Soc. Chang. 79, 1413–1430.
Jun, S.-P., 2012b. An empirical study of users' hype cycle based on search traffic: the case study on hybrid cars. Scientometrics 91, 81–99.
Jun, S.-P., Park, D.-H., 2016. Consumer information search behavior and purchasing decisions: Empirical evidence from Korea. Technol. Forecast. Soc. Chang. 107, 97–111.
Jun, S.-P., Park, D.-H., 2017. Visualization of brand positioning based on consumer web search information: using social network analysis. Internet Res. 27, 381–407.
Jun, S.-P., Park, D.-H., Yeom, J., 2014a. The possibility of using search traffic information to explore consumer product attitudes and forecast consumer preference. Technol. Forecast. Soc. Chang. 86, 237–253.
Jun, S.-P., Yeom, J., Son, J.-K., 2014b. A study of the method using search traffic to analyze new technology adoption. Technol. Forecast. Soc. Chang. 81, 82–95.
Jun, S.-P., Yoo, H.S., Kim, J.-H., 2016. A study on the effects of the CAFE standard on consumers. Energ Policy 91, 148–160.
Jun, S.-P., Sung, T.-E., Park, H.-W., 2017. Forecasting by analogy using the web search traffic. Technol. Forecast. Soc. Chang. 115, 37–51.
Khoury, M.J., Ioannidis, J.P., 2014. Big data meets public health. Science 346, 1054–1055.
Kim, D., Kim, D., Rho, S., Hwang, E., 2013. Detecting trend and bursty keywords using characteristics of Twitter stream data. Int. J. Smart Home 7, 209–220.
Knuth, D.E., 1993. The Stanford GraphBase: a platform for combinatorial computing. 37 Addison-Wesley Reading, MA.
Konrad, K., 2006. The social dynamics of expectations: The interaction of collective and actor-specific expectations on electronic commerce and interactive television. Tech. Anal. Strat. Manag. 18, 429–444.
Kotler, P., Keller, K., 2008. Marketing Management, 13th ed. Pearson Prentice Hall, Upper Saddle River, NJ.
Kotler, P., Keller, K.L., Ancarani, F., Costabile, M., 2014. Marketing management 14/e. Pearson.
Lambiotte, R., Delvenne, J.-C., Barahona, M., 2008. Laplacian dynamics and Multiscale Modular Structure in Networks. arXiv preprint arXiv. 0812. pp. 1770.
Lazer, D., Kennedy, R., King, G., Vespignani, A., 2014. The parable of Google Flu: traps in big data analysis. Science 343, 1203–1205.
Lee, B.-R., Yeo, W.-D., Lee, J.-Y., Lee, C.-H., Kwon, O.-J., Moon, Y.-H., 2008. Development of the KnowledgeMatrix as an informetric analysis system. The Journal of the Korea Contents Association 8, 68–74.
van Lente, H., Spitters, C., Peine, A., 2013. Comparing technological hype cycles: towards a theory. Technol. Forecast. Soc. Chang. 80, 1615–1628.
Liu, J., Li, J., Li, W., Wu, J., 2016. Rethinking big data: a review on the data quality and usage issues. ISPRS J. Photogramm. Remote Sens. 115, 134–142.
Lui, C., Metaxas, P.T., Mustafaraj, E., 2011. On the predictability of the US elections through search volume activity. In: Proceedings of the IADIS International Conference on e-Society. Citeseer.
Martin, H., Daim, T., 2007. Technology roadmapping through intelligence analysis: nanotechnology, in: management of Engineering and Technology, Portland International Center for. IEEE 1613–1622.
Mavragani, A., Tsagarakis, K.P., 2016. YES or NO: Predicting the 2015 GReferendum results using Google Trends. Technol. Forecast. Soc. Chang. 109, 1–5.
Mavragani, A., Sypsa, K., Sampri, A., Tsagarakis, K.P., 2016. Quantifying the UK online interest in substances of the EU watchlist for water monitoring: diclofenac, estradiol, and the macrolide antibiotics. Water 8, 542.
Mercer, D., 1997. A general hypothesis of aggregated expectations. Technol. Forecast. Soc. Chang. 55, 145–154.
Moe, W.W., 2003. Buying, searching, or browsing: Differentiating between online shoppers using in-store navigational clickstream. J. Consum. Psychol. 13, 29–39.
Newman, M., 2010. Networks: an introduction. Oxford university press.
Newman, M.E., Girvan, M., 2004. Finding and evaluating community structure in networks. Phys. Rev. E 69, 026113.
Olson, D.R., Konty, K.J., Paladini, M., Viboud, C., Simonsen, L., 2013. Reassessing Google Flu Trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales. PLoS Comput. Biol. 9, e1003256.

P.-L. To, Liao, C., Lin, T.-H., 2007. Shopping motivations on Internet: a study based on utilitarian and hedonic value. Technovation 27, 774–787.
Paranyushkin, D., 2011. Identifying the pathways for meaning circulation using text network analysis, Berlin: Nodus Labs. Retrived at: http://noduslabs.com/research/pathways-meaning-circulation-text-network-analysis.
Pelat, C., Turbelin, C., Bar-Hen, A., Flahault, A., Valleron, A.-J., 2009. More diseases tracked by using Google Trends. Emerg. Infect. Dis. 15, 1327–1328.
Prakash, B.A., Beutel, A., Rosenfeld, R., Faloutsos, C., 2012. Winner takes all: competing viruses or ideas on fair-play networks. In: Proceedings of the 21st international conference on World Wide Web. ACM, pp. 1037–1046.
Preis, T., Moat, H.S., Stanley, H.E., 2013. Quantifying Trading Behavior in Financial Markets Using Google Trends.
Raju, P.S., Lonial, S.C., Glynn Mangold, W., 1995. Differential effects of subjective knowledge, objective knowledge, and usage experience on decision making: an exploratory investigation. J. Consum. Psychol. 4, 153–180.
Rao, A.R., Sieben, W.A., 1992. The effect of prior knowledge on price acceptability and the type of information examined. J. Consum. Res. 19, 256–270.
Rogers, E.M., 2003. Diffusion of innovations. Free Press, New York., NY.
Romanelli, E., Feldman, M., 2007. Anatomy of cluster development: emergence and convergence in the US human biotherapeutics, 1976-2003. In: Cluster Genesis: Technology-based Industrial Development. Oxford University Press.
Ruef, A., Markard, J., 2010. What happens after a hype? How changing expectations affected innovation activities in the case of stationary fuel cells. Tech. Anal. Strat. Manag. 22, 317–338.
Scott, J., 2012. Social Network Analysis. SAGE Publications Limited.
Seglen, P.O., 1997. Why the impact factor of journals should not be used for evaluating research. BMJ [Br. Med. J.] 314, 498.
Seifter, A., Schwarzwalder, A., Geis, K., Aucott, J., 2010. The utility of "Google Trends" for epidemiological research: Lyme disease as an example. Geospat. Health 4, 135–137.
Shim, S., Eastlick, M.A., Lotz, S.L., Warrington, P., 2001. An online prepurchase intentions model: The role of intention to search. J. Retail. 77, 397–416.
StatCounter, 2017. Search engine market share worldwide, in, StatCounter 1997–2017. http://gs.statcounter.com/search-engine-market-share#quarterly-200901-201702.
Sullivan, D., 2016. Google now handles at least 2 trillion searches per year. In: Sullivan, D. (Ed.), Search Engine Land, Danny Sullivan.
Tefft, N., 2011. Insights on unemployment, unemployment insurance, and mental health. J. Health Econ. 30, 258–264.
USPTO, 2017. USPTO Patent Full-Text and Image Database. http://patft.uspto.gov/netahtml/PTO/search-adv.htm.
Van Eck, N.J., Waltman, L., 2009. How to normalize cooccurrence data? An analysis of some well-known similarity measures. J. Am. Soc. Inf. Sci. Technol. 60, 1635–1651.
Van Eck, N.J., Waltman, L., 2010. Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics 84, 523–538.
Vaughan, L., Chen, Y., 2015. Data mining from web search queries: a comparison of google trends and baidu index. Journal of the Association for Information Science and Technology 66, 13–22.
Vaughan, L., Romero-Frías, E., 2014. Web search volume as a predictor of academic fame: an exploration of Google Trends. J. Assoc. Inform. Sci. Technol. 65, 707–720.
Vaughan, L., Yang, R., 2013. Web traffic and organization performance measures: Relationships and data sources examined. J. Informet. 7, 699–711.
Vicente, M.R., López-Menéndez, A.J., Pérez, R., 2015. Forecasting unemployment with internet search data: Does it help to improve predictions when job destruction is skyrocketing? Technol. Forecast. Soc. Chang. 92, 132–139.
Vosen, S., Schmidt, T., 2011. Forecasting private consumption: survey-based indicators vs. Google trends. J. Forecast. 30, 565–578.
Waller, V., 2011. Not just information: Who searches for what on the search engine Google? J. Am. Soc. Inf. Sci. Technol. 62, 761–775.
Waltman, L., van Eck, N.J., 2013. A smart local moving algorithm for large-scale modularity-based community detection. Eur. Phys. J. B 86, 471.
Watts, R.J., Porter, A.L., 1997. Innovation forecasting. Technol. Forecast. Soc. Chang. 56, 25–47.
Winthrop, M.F., Deckro, R.F., Kloeber Jr, J.M., Government, R., 2002. D expenditures and US technology advancement in the aerospace industry: a case study. J. Eng. Technol. Manag. 19, 287–305.
Yan, E., Ding, Y., Jacob, E.K., 2012. Overlaying communities and topics: an analysis on publication networks. Scientometrics 90, 499–513.
Zimmer, J.C., Henry, R.M., Butler, B.S., 2007. Determinants of the use of relational and nonrelational information sources. J. Manag. Inf. Syst. 24, 297–331.

Seung-Pyo Jun is a principal researcher and director of Technology Commercialization Analysis Center at the Korea Institute of Science and Technology Information (KISTI). Also, he is an associate professor of Science and Technology Policy in University of Science & Technology (UST). He received his Ph.D. from the Science and Technology Studies Program at Korea University. His area of interest includes the demand forecasting, emerging technology detecting and new technology adoption analysis. Recently he is focusing on the new technology adoption analysis using big-data.

Hyoung Sun Yoo is a senior researcher at the Korea Institute of Science and Technology Information (KISTI). Also, he is an assistant professor of Science and Technology Policy in University of Science & Technology (UST). His area of interest includes demand forecasting, science and technology policy, complex network, and agent-based modeling.

San Choi is pursuing Ph.D. of Science & Technology Management & Policy in University of Science& Technology (UST) and work as a student researcher at Korea Institute of Science and Technology Information (KISTI). His research interest lies in technology innovation and management, economics of innovation, S&T policy, knowledge economics, organization learning, and development economics.