Data and Information Management 7 (2023) 100038

Contents lists available at ScienceDirect

Data and Information Management

journal homepage: www.journals.elsevier.com/data-and-information-management

The impact of big data on research methods in information science

Jin Zhang [a], Dietmar Wolfram [a], Feicheng Ma [b,*]

[a] School of Information Studies, University of Wisconsin Milwaukee, Milwaukee, WI, USA
[b] School of Information Management, Wuhan University, Wuhan, PR China

ARTICLE INFO

Keywords:
Big data
Research methods
Research design
Inferential analysis
Data processing
Information science

ABSTRACT

Social media platforms, search engine transaction logs, Web portal transaction logs, large text corpora, and other data sources offer users a variety of invaluable data sources. Data generated based on social media platforms and other data sources have become important sources of big data for researchers in information science. Big data provide not only challenges but also opportunities for information science. Emerging big data trends inevitably have an impact on research methods in information science. The authors of this paper discuss the impact of big data on research methods in information science.

This paper addresses these challenges and opportunities through the lens of research methods, ranging from data processing, to sampling, to information visualization, to temporal analysis, to sentiment analysis, to correlation, to cause-effect relationship, to data accessibility, to data privacy, and data ethics issues.

The discussions on related aspects of research methods provoke a healthy reflection and debate for the information science research community, which can help researchers in the field produce sound research designs for big data-oriented research studies and assist information practitioners in solving problems they face in the big data age.

Research methods are fundamental and essential for any research studies. Big data analysis has a natural relationship with research methods. The paper discusses an emerging and important research topic big data from the unique angle of research methods.

1. Introduction

Big data are impressive in terms of digital ubiquity because a great range of human behaviors and natural phenomena are captured digitally, and the pursuit of big data is driven by the epistemic assumption that expansive data sets offer superior forms of intelligence and erudition (Mills, 2018). Big data are characterized by the famous five *Vs: Volume, Velocity, Variety, Veracity,* and *Value*. They refer to the gigantic quantity of generated and stored data; the exponential growth speed of the data; the diverse types and forms of the data; the quality of the captured data; and the utility extracted from the data, respectively. These data characteristics have presented both unprecedented challenges and opportunities for data use and data analysis methods in information science. Big data call for sophisticated research methods (boyd & Crawford, 2012).

Any empirical research study consists of essential elements: research aims/objectives, research methods, data, and their analysis. They are intertwined with one another. Research methods are roadmaps, techniques, and procedures employed in a study to collect data, process data, analyze data, yield findings, and draw a conclusion to achieve the research aims. To a large degree the availability, nature, and size of a dataset can affect the selection of the research methods, even the research topics. A big dataset is not self-explanatory and it needs proper methodologies for analysis and processing. A content rich, type diverse, and size large dataset would not only broaden the horizon of research topic selections and but also give researchers a great deal of leverage to adapt a wide spectrum of research methods to the research topics. The emerging big data trend has impacted data-driven empirical research studies in information science. It is necessary to conduct a research study to systematically analyze and discuss the impact of the trend on research methods in information science. The significance of the study is multifold. It provokes a healthy discussion and thoughts about challenges and opportunities brought by the big data trend in the information science research community, it helps researchers in the field produce sound research designs for big data-oriented research studies, and it assists information practitioners in solving problems they face in the big data age.

* Corresponding author.
E-mail addresses: jzhang@uwm.edu (J. Zhang), dwolfram@uwm.edu (D. Wolfram), fchma@whu.edu.cn (F. Ma).

https://doi.org/10.1016/j.dim.2023.100038
Received 24 March 2023; Accepted 2 April 2023

2. Information science and big data

Definitions of information science have changed over time (Borko, 1968; Williams, 1987/1988; Saracevic, 2009; Stock & Stock, 2013). However, the core parts or fundamental elements of information science remain the same. Information science focuses on the selection, organization, retrieval, dissemination, and use of information. The interactions between information systems and users of information are crucial for the field and attract a lot of research studies.

Thanks to information technology (Internet technology, in particular), users can tap into rich online information sources in various ways. Search engines, social media, and Web portals have become important information sources for public users. Users employ search engines to search related information for their information needs on the Internet. Search engine Google receives over 63,000 searches per second on any given day (SEM Knowledge Base, 2018). Users explore Web portals to discover and browse information of interest. A Web portal is a specially designed website that offers information synthesized from various sources on a specific topic/area. It is customized to help users navigate a well-integrated information sources with finding aids like a roadmap, subject directory, internal search engine and other advanced features and provide the users with services. For instance, the CDC Website, a public health portal, received more than 3 billion page views in 2021 (CDC Digital Media Metrics, 2022). Users also utilize a variety of social media platforms to seek information, exchange information, and communicate with each other. Social media, supported by Web 2.0 Internet-based applications, are interactive platforms that allow users to create information, share information, and exchange information in virtual communities. There are a variety of social media platforms that serve different user communities. Social media can be classified into the following categories: Social networks, blogs, microblogs, social news, social bookmarking, media sharing, wikis, question-and-answer sites, and review sites (Barbier & Liu, 2011). Openness and interactivities of the social media platforms contribute a huge amount of data. User-generated content and the relationships and interactions between network entities (e.g., people, organizations) are the two important sources of information on social media (Gandomi & Haider, 2015). Facebook had over 2.7 billion monthly active users as of the second quarter of 2020 (Statista, 2020). As of May 2019, more than 500 h of video were uploaded to YouTube every minute, and more than one billion hours of video were watched every day (Tankovska, 2021).

All queries submitted to a search engine are maintained in transaction logs. Footprints of users who navigate in a Web portal are constantly captured and stored. The interactions or activities of users on social media platforms are recorded, kept, and finally aggregated to open archived huge data reservoirs. It is apparent that these transaction logs and reservoirs are user-generated data, growing big data, and important sources of big data sources for researchers in information science. These abundant big datasets are valuable for researchers in the field of information science because they faithfully document detailed traces of the users' information seeking activities, and widen the spectrum of research topics. Available data range from search queries, to traversal and browsing paths in a Web portal, to conversions on a topic of common interest, to users' reactions to a post like comments, likes, and dislikes to posts, to responses to events, to user-generated hashtags, and to other users' activities.

In addition to data generated by social media and general search sites, big data methods lend themselves to the analysis of scientometric datasets, which has also grown dramatically over time. These include traditional bibliographic and citation databases, and a growing number of data sources that provide access to bibliographic records, their citation linkages and altmetrics data (e.g., Google Scholar, Crossref, Microsoft Academic, Dimensions Database, Semantic Scholar, Altmetric). With access to potentially hundreds of millions of bibliographic records, billions of citation linkages and large full text corpora of scientific research that permit all areas of human knowledge to be studied simultaneously, opportunities have opened to study how disciplines relate to one another at both a high and more detailed level.

Without a doubt, the magnitude of the emerging big data generated from these big data, diversity of the data types, wide spectrum of topics, heterogeneity of user groups, variety of data sources, and complexity of users' information needs in new situations have created both challenges and opportunities to research methods applied in information science. Research findings derived from big datasets transcend the big datasets and facilitate the accumulation of knowledge in information science in an unprecedented way. Researchers in the field must revisit and reassess the traditional methods of data collection, data processing, and data analysis to address these new challenges. Understanding these challenges and opportunities and adapting suitable research methods are crucial for researchers to effectively conduct research studies in the new big data environment. It is not surprising that researchers in the field tap into the data bonanza to conduct research studies to understand users and their information use behaviors in many existing and emerging areas of information science thanks to big data.

3. Unstructured big data

Generally, most data in a big dataset, unless representing bibliographic or associated citation data, are uncleansed, unstructured (or semi-structured), and unorganized (or simi-organized). When data from a data source like a social media platform are harvested, the data are raw and uncleansed. Not every piece of information in the collected dataset is useful and useful information is not well organized to some degree. The raw data must be cleansed before any further processing can take place. Not all data are useful for a specific research study. The data useful for one research project may not be suitable for another project. Data partitioning, data separation, data extraction, and data tokenization are performed so that useful data are distilled and selected for a specific research topic. It is apparent that the distilled and selected dataset is a subset of the original big dataset. This process is guided by the researchers' perspective on the dataset while the perspective depends on research objectives and aims of a research study. In other words, different research projects may generate different distilled and selected datasets. If researchers are interested in queries in a transaction log, all query-related data are extracted. When researchers are interested in consumer search behavior on a specific disease in a public question and answer forum, only questions, answers, responses, and other information on that specific disease are selected. If a hot topic of a social event like COVID-19 focuses on a social media platform, all posts, reposts, likes/dislikes, comments, and shares on the topic are pulled out for analysis.

Defining a logical data unit is a fundamental step. A logical data unit is a complete record of an item which includes all possible useful attributes/characteristic of the item. The items are objects of interest in a study. The attributes or characteristics of an item are not only multidimensional but also hidden in most cases. Identification of meaningful and useful attributes from the item are challenging due to the unstructured and hidden nature. Defining a meaningful unit has proved a difficult task. For instance, if a visiting user is defined as a unit in a transaction log, identifying a unique user in the context is challenging. Since multiple users may use the same computer in a public place like a library to seek for information, they may use the same IP address. As a result, a proper session method must be utilized to separate different users from the same IP address. A session is a group of user interactions that take place within a given time frame. It can contain multiple page views, searches, link jumps, and other interactions. It can be defined by a time frame or campaign change or other meaningful methods. Another example is about a study about networking analysis on a social media platform, a user is defined as a record and users' connections with others are the attributes of the record. If the node centrality is used to describe the connections, the corresponding centrality values are hidden, and they cannot be obtained directly from raw data. They must be calculated based on the raw data.

The attribute identification is determined by the research questions or problems to a large degree. These attributes may serve as independent variables and dependent variables in the proposed research. The various relationships among these variables can be explored and revealed and patterns and trends can be discovered.

Big data have created a wonderful serendipity. Researchers may experience the serendipity of unearthing unexpected, relevant, and useful information arriving just when they identify and define the units and attributes from the unstructured datasets.

4. Inferential statistical analysis methods

The primary aim of a research study is to explain and predict natural and social phenomena by using scientific research methods. Inferential statistical analysis helps researchers explain and predict the phenomenon. Inferential statistical analysis infers properties of a population based on samples. A sample is a representative subset of a large population, and it is utilized to present the characteristics of the population. Samples are used in inferential analysis when a population size is too large for a test to include all possible members or observations in the population. Sampling is a key component in inferential analysis. The epistemic concerns with some traditional inferential analysis methods remain crucial parts of big data-oriented research.

4.1. Sampling

It is arguably claimed that a big dataset generated based on a data source like social media and search engines does not need sampling and inferential statistics because the dataset includes complete data. Samples and inferential analysis may be necessary and useful. First, for a huge dataset, a sample from the dataset and inferential analysis process based on the sample may be computationally efficient and save a lot of resources. Inferential analysis results can be sound and reliable if the sampling method is reasonable and appropriate. By the same token, two sampling techniques just for big datasets have been introduced (Kim & Wang, 2019). The first method incorporates auxiliary information from external sources, and the second method combines the big data sample with an independent probability sample. Second, a collected big dataset may be a subset from a growing data source. For instance, if a dataset is collected from social media like public social question answer forums, Facebook, Twitter, YouTube, or search engine transaction logs, it only covers a certain time period. It can be one month, one year, ten years, or an even longer period. Although the dataset includes all complete raw data in that certain time period, it is only a subset of an ongoing and growing data source. In this sense, the collected dataset is always a section of the ongoing and growing dataset to some degree. In other words, new posts, comments, likes, dislikes, and tags on that social media or newly submitted queries on the transaction logs can be appended to existing topics and themes. In addition, newly emerging topics and themes can be created and added to the social media and search engine logs, which may not be reflected in the old dataset. As a result, it is impossible that the collected dataset is treated as an entire population due to the characteristic of growing data. The collected dataset is a special sample of the ongoing and growing dataset. That is, the collected dataset represents a section of an entire population in only a limited time period. We can call it the section sampling method. Generally speaking, sampling methods can be classified as probability sampling and non-probability sampling (Etikan & Bala, 2017). The former group includes stratified sampling, simple random sampling, cluster sampling, and systematic sampling while the latter group includes convenience sampling, and judgmental sampling. The major difference between these two groups is whether the items in a sample are selected randomly or not. It is apparent that the section sampling method is a non-probability sampling method because the items come from a data selection and they are not randomly selected. The random selection of sample items would ensure that the results are well representative of the population. Lack of the random selection runs the risk of systematic bias. It would increase the risk of over-representation of a sample or under-representation of a sample. Lack of the random selection is the weakness of the section sampling method. Finally, a study showed that the sample size affected inferential analysis results (Ajiferuke et al., 2006). A relatively large sample size from a big dataset leads to less biased and more accurate and robust results for inferential analysis.

4.2. P-value hacking issues in inferential analysis methods

In inferential statistical analysis, the p-value is an extremely important concept. It refers to the probability that describes how likely a sample of observations is founded if a null hypothesis is true. The p-value from an analysis result is used to reject (or not reject) the null hypothesis at a significance level. Since a null hypothesis is derived from a research question/problem in a research study, the p-value directly affects the answer to the research question/problem. In the field of information science, in most of the cases, the significance level is 0.05.

P-value hacking, which is also called data-dredging and significance chasing, refers to the phenomenon that researchers achieve a satisfactory p-value in a study by using multiple means. Notorious p-value hacking can result in bias in an inferential study. For instance, to gain a desired p-value, people can keep changing samples, applying different inferential analysis methods, excluding unfitted sample items, using different measurements, changing data types, altering experimental environments, etc. A study showed that p-value hacking and publication bias had a close relationship, and p-value hacking can severely increase the rate of false positives (Friese & Frankenbach, 2020). The p-hacking issue is widespread throughout science (Head et al., 2015). It is clear that the p-value hacking issue stems from the inferential analysis processes. If a research study is based on a big dataset that covers the entire population, the study does not need any samples and inferential analysis. Consequently, the p-value hacking issue can be avoided if a complete big dataset is used in a research study.

5. Exploratory and statistical methods

Big data analysis relies heavily on different types of quantitative analysis methods that may be applied to numeric, textual or media formats. Although many researchers in library and information science traditionally bring background preparation in a range of disciplines, unless this preparation includes a strong technical background in computing or statistical methods, some of the techniques used may be new to researchers. Data mining techniques cover a broad range of exploratory and statistical methods to identify patterns that exist in datasets. There are too many to be covered in this brief treatment, but the methods that may be used in information science research include correlation and regression analysis, natural language processing and other language-based methods such as sentiment analysis, temporal analysis and visualization methods.

5.1. Correlation and regression

Correlation analysis is a statistical method used to evaluate the strength of the relationship between quantitative variables. Correlation analysis methods really shine in the context of big data and predictions based on correlation analysis lie at the heart of big data (Mayer-Schönberger & Cukier, 2013). That is because a big dataset provides multidimensional and sophisticated data which allow researchers identify and define diverse quantitative variables and conduct correlation analyses for these variables. Predictions for variables have been widely used in many domains. They are motivations for many research studies in information science. For instance, if information needs of health consumers in a virus breakout like Covid-19 on social media are discovered, the findings can be used to raise awareness of the public about a pandemic and help people reduce its impact on the society. If information seeking patterns of

people on asthma in the four seasons are found, they can be employed for health consumers to prevent the possible risks caused by asthma.

Regression analysis is an approach for modeling the relationship between variables. One of the most salient features of regression analysis methods is that it can be employed to predict what may happen to one variable in the future given another variable. Correlation analysis shows the relationship between the variables while regression analysis demonstrates how one variable affects the other variable. For instance, a regression analysis between users' preference and the role of presenters who present in a video, gender of the presenters, settings, or topics on diabetes related video on YouTube can be conducted (Wang & Zhang, 2020). The attitudes of the viewers to the identified characteristics can be discovered through regression analysis. The results would benefit health consumers, health professionals and YouTubers in multiple ways. Correlation analysis and regression analysis have a natural relationship and the two analyses can be conducted on the same dataset. It is worth pointing out that a variety of regression methods such as linear regression, logistic regression, ridge regression, lasso regression, polynomial regression, Bayesian regression, negative binomial regression, and other regressions can be applied to big datasets thanks to data relationship richness of big data.

5.2. Natural language processing and text mining methods

The availability of large text corpora for research offers opportunities for the application of Natural Language Processing (NLP) methods to the datasets for text and data mining purposes. Parts of speech, text types and token data for large corpora such the HathiTrust Digital Library, available through the HathiTrust Research Center, provide access to more than 17 million volumes, more than 6 billion pages, and more than 2.9 trillion tokens (https://analytics.hathitrust.org/datasets). Although NLP methods are also applied to smaller datasets, the insights gained from linguistic analysis and text mining on a large scale can lead to a better understanding of differences in language use over time or based on geography, as exemplified by Google Trends (https://trends.google.com/trends/) and Google Ngram Viewer (https://books.google.com/ngrams#).

The application of topic modeling, which uses NLP and machine learning methods to reduce the textual relationships with text corpora to a much smaller number of identified "topics", has become common in information science research. It has been used to study the evolution of library and information science (Figuerola et al., 2017) and to study topical areas in social media such as Twitter (Hong & Davison, 2010). Latent Dirichlet Allocation (LDA) was applied to discover autism themes on Facebook (Zhao et al., 2019). It can also be used as a complementary way to study the relationships of authors based on content of their works as a complement to more traditional approaches based on citations or term co-occurrence (Lu & Wolfram, 2012).

5.3. Sentiment analysis method

Sentiment analysisis a data analysis method that combines text analysis, computational linguistic analysis, and statistical analysis to systematically identify, quantify, and reveal affective states and degree of a text of interest. The text can be defined as a word, sentence, paragraph, or a large section. Sentiment analysis result from a text can be represented by a sentiment score, which indicates positive, negative, or neutral emotion for the text. The sentiment score is usually standardized on a -1 to $+1$ scale or other scales. Here -1, 0 and $+1$ stand for negative emotion, neutral emotion, and positive emotion, respectively. It is the score that provides researchers with a rational measurement for quantitative sentiment analysis. It is widely used to analyze opinion, attitude, feedback, or reaction to an event or a figure. It is quite clear that sentiment analysis is subjective in nature. For instance, a social media platform is open for the public and its users can interact with each other. It covers a wide spectrum of topics. Some topics like public health, politics, news, and so on can be quite sensitive to emotion. It is not surprising that people post items, make comments, and present other forms of expressions on these topics with strong emotions. Traditional information text processing methods focus on subject terms from a text; stop words and other insignificant words are eliminated. During the processing non-subject terms like sentiment words usually are excluded because they bear little or no subject meanings. These terms cannot be used as indexing terms to represent themes of the text. Therefore, they cannot be utilized for subject analysis and information retrieval. Sentiment analysis adds a unique dimension to social media related data analysis. It helps people better understand the context of the analyzed text. Sentiment analysis aids researchers in revealing user's information seeking behavior on social media from a unique perspective. It puts a "colorful sentiment layer" on the user's information behavior patterns. For instance, a study (Zhang et al., 2020) found that negative sentiment scores from the terms related to panic (e.g., danger, worried, concern, sad, scared, epidemic, and kill) on a social media platform suggested that the Zika virus outbreak caused public panic.

5.4. Temporal analysis method

Temporal analysisis used to examine and model the change patterns of a specific variable in a dataset over a period. The salient uniqueness of a big dataset generated by computer systems is its temporal characteristic. Notice that when data are generated in a computer-based system, time is always associated with the generated data. In most data transaction logs, a time stamp is used to automatically record exact time of a data item creation caused by an online action. It is composed of the year, month, date, hour, minute, and second. It is the temporal characteristic that allows researchers analyze the dataset from the time perspective. It can demonstrate accurate duration of a meaningful activity based on the time stamp data, helping researchers connect the dots of an event of interest from beginning to end. That is, it provides the researchers with a new time dimension to examine, evaluate, and model the behavior change of specific variables over a period to observe their change trends and patterns. The longitudinal or temporal analysis method can be utilized as a natural and effective means. For instance, the parallel coordinate analysis technique enables researchers to effectively observe the evolution of multiple variables over a time period (Stevenson & Zhang, 2015). Using the temporal analysis method, the theme changes during Zika virus outbreaks on a public question and answer forum were revealed (Zhang et al., 2020). Thanks to the time information associated with activities in a transaction log, the duration that a user stays on a meaningful object like a webpage or node of a subject directory, and path that a user navigates in a Web portal can be specified and defined as variables for studies related to users' browsing behaviors.

5.5. Data and information visualization methods

Data and information visualization methods have a unique way to handle a large dataset and abstract relationships among items in the dataset. Information visualization methods are used not only to reduce cognitive burden to understand the data in the dataset, but also become effective data analysis means (Zhang, 2008). When a proper visualization method is applied to a large dataset, the data space dimensionality is reduced, sophistical data relationships are simplified, salient features are emphasized, meaningful dots are connected, and finally the distilled data are projected onto visual spaces. Consequently, invisible hierarchy relationships, network relationships, and other relationships are presented in an intuitive way for exploration and analysis. Holistic overall pictures of a dataset at different levels are demonstrated and various perspectives of the dataset are provided. For instance, for search terms, term clusters are yielded and displayed based on their subject connections. The patterns and trends of the dataset merge for analysis. The visualization methods and other clustering methods are widely used in subject analysis. Visualization methods includes, but are not limited to, multidimensional scaling (*MDS*), self-organized maps (*SOM*), and social network

analysis (*SNA*). For instance, the MDS method was utilized to reveal the subject themes from transcripts of malaria related clips on YouTube (Omwando & Zhang, 2021), the SOM method was used to investigate data mining related topics on Wikipedia (Wang & Zhang, 2017), and the SOM method was employed to optimize the relationships among subject directory nodes on a public health portal (Zhang et al., 2016). However, how to effectively visualize a huge amount of information in a limited display space, to interpret and explain the patterns generated from the different visualization models, to project objects in a high-dimensional space to a low dimensional visual space, and to evaluate the visual environments remain challenging issues for researchers.

5.6. Link analysis methods

Link analysis is a data analysis technique used to evaluate relationships (connections) between nodes in a network environment. Nodes can be websites, people, organizations, events, references/citations, and objects of interest in a network. Relationships may be identified and defined based on the nodes (objects). There are a variety of relationships. Link analysis allows researchers to quantitatively analyze relationships and the importance of nodes, and discover patterns of a network. For instance, in a social media environment following and followed relationships of people form a virtual network. Influential figures in the social network can be effectively identified through link analysis. The diffusion analysis of the link analysis method includes diffusion breadth, diffusion time, diffusion speed, diffusion acceleration, and diffusion delay. It enables researchers to trace an event like a piece of misinformation from the beginning to the end, size up the scope of the event, and monitor spread speed of the event in a community. On a Web portal users' visit log data can generate a network. When a user jumps from one website to another website within the portal, it creates a link between the two websites. Accumulated visit data yield a user visit network. Using the link analysis method important webpages on the portal can be specified.

5.7. Clustering methods

An important function of data exploration of large datasets is to be able to identify similar characteristics within the data to identify groupings of data observations or records. The groupings may be used to reduce the dimensionality of the data relationships, which in turn reduce the computational overhead associated with data processing. Related statistical techniques used for this purpose include cluster analysis, factor analysis and principal component analysis, and multidimensional scaling. Clustering techniques have been widely adopted in information science research including information retrieval, where it can be used to identify related documents and topic modeling in IR systems to reduce the dimensionality of the vocabulary in determining document relationships. These relationships can also extend to search log analysis, where groups of session patterns may be identified (Chen & Cooper, 2001; Wolfram et al., 2009) or relationships in the co-occurrent of posts may be visualized using multidimensional scaling (Zhang et al., 2020). Forms of clustering are also frequently used in scientometric research to identify groups of entities of interest, whether authors, publications, journals or other units of measurement. Notably, the links created between entities of interest based on citations, co-authorship, or the co-occurrence of terms allows link-based methods to also be used to identify clusters (Zhao & Strotmann, 2014). Language-based groupings can include the use of topic modeling, as mentioned above to cluster scientific documents (Yau et al., 2014).

6. Experimental methods

An experimental study is utilized to ascertain a cause-effect relationship between variables. It is fundamental and essential for any fields because a true experimental research study enables researchers to randomly assign treatments to subjects, fully control independent variables in a research design, collect data from dependent variables, analyze the data, and draw a conclusion. The randomization of the treatment assignments is the power of a true experimental research study. Unfortunately, in a big data situation, researchers cannot randomly assign treatments to subjects from different test groups to gain data because the data in a big dataset pre-exist. As a result, in the big data situation researchers cannot manipulate and control independent variables to get data and test deliberately framed hypotheses at will. Big data encourage passive data collection, as opposed to experimentation and hypothesis testing (Frické, 2015).

Quasi-experimental research also aims to examine a cause-and-effect relationship between an independent variable and dependent variable. However, the way of assigning treatments to subjects in quasi-experimental research differs from true experimental research (Cook & Campbell, 1979). The treatment assignment is based on a non-random method in quasi-experimental research by manipulating/selecting subjects with(without) pre-existing treatment characteristics in experimental subject groups. In other words, the subjects receive the treatments after the fact. This characteristic just dovetails with the big data situation because the statuses of an independent variable and the scores of a dependent variable pre-exist in a big dataset before the independent variable and dependent variable are defined and their relationship is established by researchers. Researchers don't generate new data through an experiment and they only utilize pre-existing data. The manipulated independent variables can be gender, pre-existing medical condition, education level, age, user groups, etc. For instance, in a local school group on Facebook the attitudes to the Covid-19 vaccination among students, parents, and teachers can be compared and analyzed based on their posts. The valid levels of the independent variable are student, parent, and teacher while the dependent variable is their attitude to the vaccination.

Definition of a dependent variable is indispensable for a quantitative research study and finding a proper measurement for the dependent variable is key to its research design. Observe that in open public platforms like social media, users are allowed to add personal reaction information to an existing data item like a video, post, tweet, or answer to a question. It is important part of interaction on the social media platforms. Big datasets resulted from social media contain rich user feedback information such as the number of likes/dislikes, the number of reads/views, the number of retweets/reposts, the number of the comments, the number of followers, the number of the responses, and other meaningful numbers. The first-hand and ample information is good indicators of users' attitude or reaction to an item of interest. They can be used as potential measurements for possible dependent variables if research studies are related to user-oriented studies like users' satisfaction to a specific topic, or preference to certain features and characteristics of a social media platform, or influential people in an online public community. Researchers can directly take advantage of the pre-existing data and associate with defined independent variables without conducting an experimental study to obtain the data. The advantages of using these user-generated data in real environments as scores of the dependent variable are not only the sheer amount of the available data but also less bias than the data generated from an experimental study.

7. Other big data issues

7.1. Assessment of data

Large amounts of data are not an indicator of the quality of the data. Similarly, the exploratory and analytical techniques applied to large datasets, although identifying patterns that may be hidden within the data, may not be conclusive. Prathap (2014) has warned that large datasets may not conform to the relationships found in smaller datasets and that "brute force computation with big data may lead to false discoveries and spurious correlations" (p. 1422). Also, the use of different analytical and visualization techniques applied to the same datasets can

produce different outcomes or interpretations. Can researchers conclude that one technique is preferable to another, or more correct? boyd and Crawford (2012) argue that "In reality, working with Big Data is still subjective, and what it quantifies does not necessarily have a closer claim on objective truth" (p. 667).

7.2. Data availability and accessibility

The availability and accessibility of large datasets has increased with larger capacity storage devices and high-speed transmission methods. However, the existence of big datasets does not guarantee their availability to all interested audiences.

Due the concerns of users' privacy, more and more social media platforms tighten their data access policies. The tightening of data access policies can have a significant impact on the availability and accessibility of data for researchers who want to collect data from these platforms. These changes can make it more challenging for researchers to access the data they need to conduct their studies, and may limit the types of research questions that can be answered using social media data. One of the primary effects is that it can become more difficult to obtain data from social media platforms. Platforms may require researchers to go through a more rigorous application process, provide more detailed information about their research. These changes can create additional barriers for researchers, especially those who are not familiar with the platform or its policies. Another impact of tighter data access policies is that the data that are available may be more limited in scope or quality. Platforms may limit the amount of data that can be accessed or the types of data that can be collected.

Furthermore, access to high-speed networks for efficient data transfer and very large capacity storage media can limit what researchers are able to collect and analyze.

7.3. Ethical and privacy issues

Raw scientific data that are not tied to individuals or population may not lend themselves to ethical concerns, unless the data were collected and analyzed in such a way that outcomes are intentionally misleading, which may then erroneously inform decision-making. In information science, much of the data we study have a direct or indirect human element that can raise privacy concerns. Just because large volumes of data from information system transaction logs or social media platforms are accessible, should we take full advantage of these opportunities? Should the individuals the data represent also have a say in what is made available. Although good-intentioned, privacy can be compromised with enough de-identified data. A classic example was the AOL search log debacle in 2006, where, in the interests of supporting the research community's wish to better understand user search behavior, AOL released a search log containing 20 million de-identified searches from 650,000 users. With sufficient data that could contain personally identifiable information, researchers were able to identify several individuals based on their submitted searches. Similar concerns exist with social media data (Smith et al. 2012), leading to calls for the development of ethical guidelines in the use of these data (Neuhaus & Webmoor, 2012; Taylor & Pagliari, 2018).

8. Conclusion

Issues associated with the collection and analysis of big data impact how we engage in information science research. With the potential for access to larger datasets and the application of established and novel analytical techniques to reveal patterns in large amounts of data, along with tools to visual complex relationships both statically and over time, many opportunities exist for the information science research community to understand more fully the phenomena we study. Still, challenges remain and reflection on the implications of big data is needed. Larger datasets are not necessarily better datasets. Furthermore, the existence of larger datasets does not mean they are readily available. Technical or policy reasons may limit their accessibility. Even if available, the outcomes of what large scale analyses reveal must be tempered with the ethics of using personal data. With the availability of greater amounts of data comes greater responsibility for how we use data. Finally, the analytical methods we apply to large datasets may produce different outcomes, and outcomes that may be subjectively interpreted, raising questions about the objectivity and validity of any conclusions drawn.

In this paper, we have explored how big data impacts various research methods, particularly within the field of information science. The influence of big data on research methods cannot be overstated, as it has a profound impact on the way research is conducted. However, the application of these research methods to big data-related studies ultimately depends on the nature, emphasis, and perspective of the specific research study at hand. By taking into account these key factors, researchers can better tailor their methods to studying big data and unlock new insights and knowledge in the field.

The examination of the impact of big data on research methods within the realm of information science has significant theoretical and practical implications. It not only enriches research methods in this field, but also contributes to the overall knowledge system in information science. The insights gained from this exploration can aid researchers in selecting appropriate research methods, expanding their research design options, and exploring a wider range of big data-related research topics. Additionally, this knowledge can assist information professionals in honing their skills in effectively solving problems within the realm of big data.

Established and emerging research methods and data analysis techniques in information science hold much potential for the analysis of big data in multiple areas of the field, in particular those that rely on usage and information production data including transaction logs, social media, textual corpora, and document data used in scientometrics.

Declaration of competing interests

The authors have no competing interests to declare.

References

boyd, d., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society, 15*(5), 662–679.

Ajiferuke, I., Wolfram, D., & Famoye, F. (2006). Sample size and informetric model goodness-of-fit outcomes: A search engine log case study. *Journal of Information Science, 32*(3), 212–222.

Barbier, G., & Liu, H. (2011). Data mining in social media. In C. C. Aggarwal (Ed.), *Social network data analytics* (pp. 327–352). Springer.

Borko, H. (1968). Information science: What is it? *American Documentation, 19*(1), 3–5.

CDC Digital Media Metrics. (2022). CDC.gov 2022 monthly page views. Available at https://www.cdc.gov/digital-social-media-tools/metrics/cdcgov/index.htm. (Accessed 28 January 2023).

Chen, H. M., & Cooper, M. D. (2001). Using clustering techniques to detect usage patterns in a Web-based information system. *Journal of the American Society for Information Science and Technology, 52*(11), 888–904.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues in field settings.* Boston, MA: Houghton Mifflin.

Etikan, I., & Bala, K. (2017). Sampling and sampling methods. *Biometrics & Biostatistics International Journal, 5*(6), Article 00149.

Figuerola, C. G., Marco, F. J. G., & Pinto, M. (2017). Mapping the evolution of library and information science (1978–2014) using topic modeling on LISA. *Scientometrics, 112*(3), 1507–1535.

Frické, M. (2015). Big data and its epistemology. *Journal of the Association for Information Science and Technology, 66*(4), 651–661.

Friese, M., & Frankenbach, J. (2020). P-hacking and publication bias interact to distort meta-analytic effect size estimates. *Psychological Methods, 25*(4), 456–471.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management, 35*(2), 137–144.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biology, 13*(3), Article e1002106.

Hong, L., & Davison, B. D. (2010). Empirical study of topic modeling in twitter. In *SOMA '10 Proceedings of the first workshop on social media analytics* (pp. 80–88). https://doi.org/10.1145/1964858.1964870

Kim, K., & Wang, Z. (2019). Sampling techniques for big data analysis. *International Statistical Review, 87*(260), 177–191.

Lu, K., & Wolfram, D. (2012). Measuring author research relatedness: A comparison of word-based, topic-based, and author cocitation approaches. *Journal of the American Society for Information Science and Technology, 63*(10), 1973–1986.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

Mills, K. A. (2018). What are the threats and potentials of big data for qualitative research? *Qualitative Research, 18*(6), 591–603.

Neuhaus, F., & Webmoor, T. (2012). Agile ethics for massified research and visualization. *Information, Communication & Society, 15*(1), 43–65.

Omwando, B., & Zhang, J. (2021). Analysis of malaria information on a social media platform, HCI international 2021. *Lecture Notes in Computer Science, 12796*, 298–316.

Prathap, G. (2014). Big data and false discovery: Analyses of bibliometric indicators from large data sets. *Scientometrics, 98*(2), 1421–1422.

Saracevic, T. (2009). Information science. In M. J. Bates (Ed.), *Encyclopedia of library and information sciences* (3rd ed., pp. 2570–2585). New York: Taylor and Francis.

SEM Knowledge Base. (2018). 63 fascinating google search statistics. Available at https://seotribunal.com/blog/google-stats-and-facts/#:∼:text=1.,second%20on%20any%20given%20day.&text=That's%20the%20average%20figure%20of,5.6%20billion%20searches%20per%20day. (Accessed 28 January 2023).

Smith, M., Szongott, C., Henne, B., & Von Voigt, G. (2012). Big data privacy issues in public social media. In *2012 6th IEEE international Conference on digital Ecosystems and technologies (DEST)* (pp. 1–6). IEEE.

Stevenson, J., & Zhang, J. (2015). A temporal analysis of institutional repository research. *Scientometrics, 105*, 1491–1525.

Stock, W. G., & Stock, M. (2013). *Handbook of information science*. Boston, MA: De Gruyter Saur.

Tankovska, H. (2021). Hours of video uploaded to YouTube every minute 2007-2019. available at https://www.statista.com/statistics/259477/hours-of-video-uploaded-to-youtube-every-minute/#:∼:text=As%20of%20May%202019%2C%20more,newly%20uploaded%20content%20per%20hour. (Accessed 28 January 2023).

Taylor, J., & Pagliari, C. (2018). Mining social media data: How are research sponsors and researchers addressing the ethical challenges? *Research Ethics, 14*(2), 1–39.

Wang, Y. Y., & Zhang, J. (2017). Exploring topics related to data mining on Wikipedia. *The Electronic Library, 35*(4), 667–688.

Wang, Y. Y., & Zhang, J. (2020). Investigation of women's health on Wikipedia - a temporal analysis of women's health topic. *Informatics, 7*(22), 1–26.

Williams, M. E. (1987/1988). Defining information science and the role of ASIS. *Bulletin of the American Society for Information Science, 14*(2), 17–19.

Wolfram, D., Wang, P., & Zhang, J. (2009). Identifying web search session patterns using cluster analysis: A comparison of three search environments. *Journal of the American Society for Information Science and Technology, 60*(5), 896–910.

Yau, C. K., Porter, A., Newman, N., & Suominen, A. (2014). Clustering scientific documents with topic modeling. *Scientometrics, 100*(3), 767–786.

Zhang, J. (2008). *Visualization for information retrieval*. Springer.

Zhang, J., Chen, Y., Zhao, Y. H., Wolfram, D., & Ma, F. C. (2020). Public health and social media: A study of zika virus-related posts on Yahoo! Answers. *Journal of the Association for Information Science and Technology, 71*(3), 282–299.

Zhang, J., Zhai, S., Liu, H., & Stevenson, J. (2016). Social network analysis on a topic based navigation guidance system in a public health portal. *Journal of the Association for Information Science and Technology, 67*(5), 1068–1088.

Zhao, D., & Strotmann, A. (2014). The knowledge base and research front of information science 2006–2010: An author cocitation and bibliographic coupling analysis. *Journal of the Association for Information Science and Technology, 65*(5), 995–1006.

Zhao, Y. H., Zhang, J., & Wu, M. (2019). Finding users' voice on social media: An investigation of online support groups for autism-affected users on Facebook. *International Journal of Environmental Research and Public Health, 16*(23), 4804.