2H07

Exploring "Big data" Applications for Disaster Management: a Scientific Keyword Word Co-occurrence Network Analysis

○ Santiago Ruiz Navas, Kumiko Miyazaki (Tokyo Institute of Technology)

Natural disasters are part of the planet natural course, these phenomena in our present society have become a matter of social interest due to their impact on people' s lives and the world' s economy, when they happen and in their aftermath. In 2015 the third UN world conference on disaster risk reduction was held in in Japan - Sendai. Among the measures to achieve the objectives of the risk reduction framework a new word for this field was used in the proceedings, this word was "big data".

The big data technologies offer to enhance humans in their decision making process which translates into improved efficiencies and cost reductions, this cost depending on the field of where the big data is being applied, in disaster prevention it is expected to be translated in more lives saved.

The aim of this work is to explore the applications of big data in disaster prevention. To limit the scope of the paper we propose the following research questions

- What are the more researched applications?
- Which countries are researching about it?

To answer these questions this research implements network analysis on scientific journal paper' s keyword. As results applications in the whole chain of disaster risk mitigation were detected and as an unexpected result also the applications on improvement of algorithms, platforms and data for disaster management applications.

Keywords: big data, disaster management, NHK, Disaster big data, keyword analysis

INTRODUCTION

Natural disasters have been part of the history of our planet since its creation, now when these happen their consecuences on society are considerable in different aspects, human, social, economic and environmental. Past efforts have been done to mitigate the impacts of the natural disasters but with recent changes in the world policies about climate change and inclusion and technological advance, the Sendai framework proposed an update approach to mitigate disaster impacts. Among the mentioned technological tools, big data was presented as one of the tools to help in these efforts.

Data has been used before to help decision making, recently with new sources of data have been reported as useful in disaster mitigation efforts and decision making, like social networks data. This usefulness extends to disaster management and crisis response. The synergy of institutions, volunteers and technological platforms have allowed the usage of big data to save lives and improve the whole cycle of disaster management. The typical example of the benefits of big data were seen after Haiti earthquake, in which crowd sourced big data helped to improve relief delivery and impact assessments among other activities.

For the institutions involved in disaster prevention the identification of these applications and big data utility is important, yet is difficult to revise all the published information about them. To explore great amounts of published data, different methods for mapping science have been used, citations and roadmaps, keywords analysis and the combination of them. It is perceived that citations are good to map strong ties of scientific progress in some scientific sectors, this method can sometimes lead to ignoring emerging patterns or weakly cited relevant research.

Words analysis on the other hand, focuses on semantic relationships between documents generating a more broad connection of topics, resulting in possible more noisy results but also allowing the identification of potential emerging topics.
The aim of this work is to explore the applications of big data in disaster prevention. To limit the scope of the paper we propose the following research questions

- What are the more researched applications?
- Which countries are researching about it?

BACKGROUND THEORY

Disaster risk management: disaster is defined as "A serious disruption of the functioning of a community or a society involving widespread human, material, economic or environmental losses and impacts, which exceeds the ability of the affected community or society to cope using its own resources" (UNISDR 2009). Disaster implications for society are broad, they range from economic cost to social and environmental ones (UNISDR et al. 2015). From long time the United Nations have been organizing to reduce the negative impacts that disasters have on the world. Recently in the last meeting of the United Nations on SENDAI-Japan, an update on the framework for disaster risk management was published and in this document two points were highlighted, political interest on climate change and social inclusion and the advent of new technologies like big data and IoT gave to this document a new face and identity.

Big data and disaster: in the context of disaster management, previously, big data was not a familiar term until its application on Haiti earthquake where big data and the power of the crowd and volunteers played an important role to help humanitarian efforts, other applications of this nature were reported for Philippines after the 2013 Haiyan Typhoon and April 2015 Nepal's earthquake. The amount of data these project involved was huge, as thousands of users were uploading, updating and tagging of information and the system had to process, deliver and visually report it so the humanitarian ground support could use it (Marr 2015). In Japan after the 2011 Tohuku, earthquake and tsunami the project 311 was developed, which involved universities, private institutions, government, volunteers and NGOs, the objective of this project was to use social network data to learn about the disaster (Google 2012) some of the developed big data based solutions were reported by NHK documentaries special on disaster big data (NHK 2013).

Science and technology quantitative mapping methods: with the increasing number of publications, scientific, technological and others, the capacity for a single human to revise all the information related to an specific topic has become limited. Quantitative methods have been proposed as tools to help researchers to analyze this sea of information. Among these methods, the citation analysis is widely used (Small 2006), for mapping of scientific fields, the detection of emerging topics or technologies (Ávila-Robinson & Miyazaki 2013), knowledge flows and diffusion (Geum et al. 2012). Other approach is the use of words' semantic relationships, to detect knowledge topics (Van-RAAN 1988). The combination of the two proposed approach is also used to get the best of the two worlds for example improve the detection of clusters in co-citation analysis (Yasutomo et al. 2016) or to link technology solutions to social issues (Ittipanuvat et al. 2014). In the previously mentioned examples, the citation methods are always used as base to construct the knowledge field or network of related documents, it has been said that citations based methods include the dynamics of scientific citation behaviors, as it is self-citation or elite scientific groups' citation clusters. Also the dynamics in time of citations might leave behind emerging important topics out of the resulting maps. Words based methods on the other side, are broader on their reach of topics or semantic meaning, but they tend to include words which might not be related to the scientific topic that the researcher wants to study, for example scape words (the, or, by, study…). In summary for applications where the scientific structure, the citations of the papers to study is of importance for the researcher and the topics' life cycle is long to medium the citation analysis offer a stable solution; for the case where the focus is on exploration of topics more than on the citation quality of papers and the life cycle of the topic is short and might not be correctly captured by citations then the words analysis can be used.

One improvement for word analysis is the use of dictionary or bag of words (Turney & Pantel 2010) that relate to the topic that one wants to study.

METHOD

The method proposed in this paper consist of 3 steps, 1) data collection, 2) keyword co-occurrence network construction and 3) Potential application analysis. The flow of the method can be seen in FIG 1.

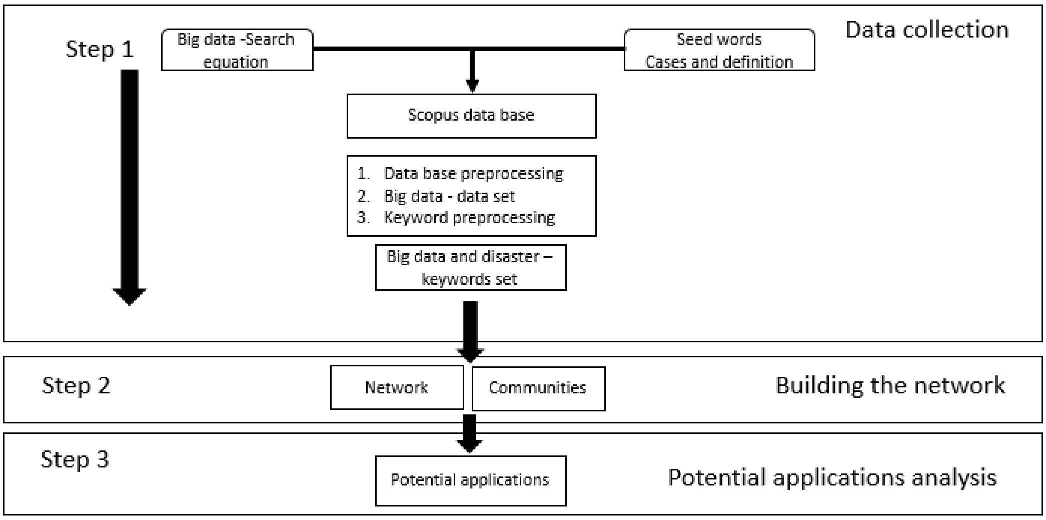

FIG1. Method flowchart.

1) **Data collection**: the objective is to obtain words related to big data and disaster prevention to create a search equation that will return papers related to both topics. The selection of big data words is obtained by referencing the work of Huang et al. 2015 where a search equation for big data is proposed. For the case of disaster prevention the terms for disaster risk management published by the UNISDR, 2009 is used as a reference and from them specific keyword/s are/is selected. In this step also basic descriptive statistics of the data set will be done to answer the second research questions.
2) **Building the network**: after obtained the big data and disaster prevention paper set, the keywords from the author keyword fields of these papers are selected to create a co-occurrence matrix, which then is used as the adjacency matrix to create a keyword co-occurrence network.
3) **Potential applications analysis**: After creating the network, the Louvain algorithm for network community detection will be run and the most important or central words of these will be selected as indicators of the topics they represent. After analyzing these more representative words and revising the contents of the abstract of the papers in each community, the authors will explain the potential applications found in each of the detected word communities taking into account the SENDAI framework for disaster risk mitigation.

RESULTS

Data collection: the search equation is composed of the two parts, firs part is big data related

words and concepts as explained in Huang et al. 2015 the selected search equation was adapted to be run in Scopus API and also filtered to include only journal articles, normal and in press and conference papers, from 2008-2016.
For the second part, after reading the terms included in the UNISRD terminology on disaster risk reduction document, the selected word is disaster, which showed a generalization power to represent the topic.
After combining these two equations the final search equation is created.
"TITLE-ABS-KEY ({Big Data} OR big?data OR map?reduce OR hadoop OR hbase OR nosql OR newsql AND (disaster)) AND (DOCTYPE (ar) OR DOCTYPE (ip) OR DOCTYPE (cp))", 185 hits.
The countries with more papers are china, united states, japan, the United Kingdom and South Korea. As seen in FIG2.

Country/Territory	
China	(52)
United States	(35)
Japan	(28)
United Kingdom	(14)
South Korea	(13)

FIG2. Report of country affiliations of the papers in the data set. Source: obtained from Scopus web search interface.

Building the network: by using the R scripts, a keyword co-occurrence network is created from the papers extracted. The network had 569 nodes or unique keywords and 1860 edges or co-occurrence between these keywords. Using the radial axis layout in FIG 3 the general network and its 40 communities can be seen.

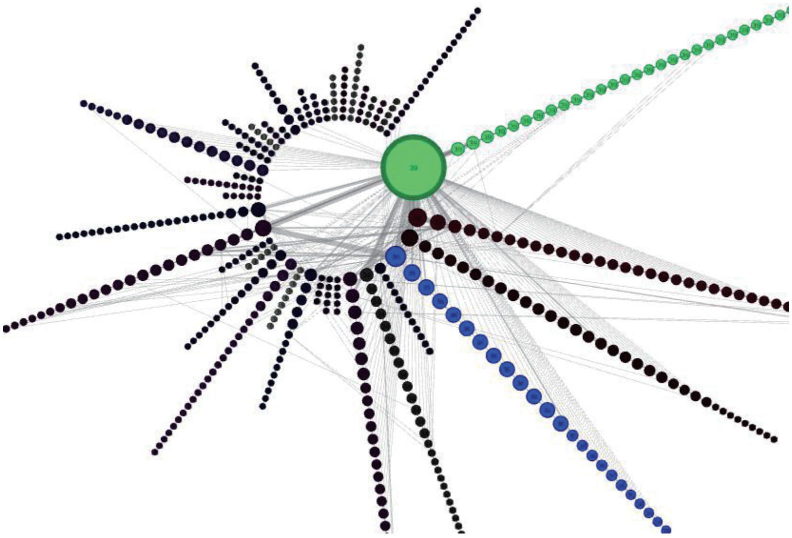

FIG 3. Radial axis layout of the network, generated in GEPHI v.0.8.2, and its 40 communities the two in color are the 2th biggest and more central communities selected for the analysis.

Potential applications analysis: After analyzing the papers involved in the two communities selected for analysis the following two types of applications were defined
Disaster management systems and Improvements of disaster data, processing algorithms and architecture.

- Disaster management and their design: The systems cover the different stages explained by Sendai disaster risk reduction framework.
- Improvements of disaster data processing algorithms and architecture: consist in applications that modify data or the ways to analyze it to make them more useful for

disaster management applications.

Community 36: the most in average central community.

- **Disaster management systems and their design:** tsunami early warning systems for Indonesia using social network data, to resource allocation for oil spills using ship traffic data, disaster evaluation using matching of geospatial simulations against real data, among others.
- **Improvements of disaster data processing algorithms and architecture**: algorithm improvements examples are the addition of location data to mobile phones in India, other is the adaptation of probabilistic methods to discover the causal relationships of earthquake related data to explain disaster impacts. Architectural improvement examples are a knowledge management platform for tagging data from third party sensors, which allows the integration of different providers' sensor information to the system in case of disaster. The other example is a system to integrate different formats and compression versions of Geographic information system (GIS) data.

Community 39: the biggest and second more in average central community.

- **Disaster management systems and their design:** These mainly help as instruments to improve decision making in different stages for example, evacuation, planning, reaction and prediction. Applied in different context mines, urban, firefighting, disease disasters, drug safety, power systems, smog and patient care also in different levels i.e. coastal regions, city and country levels.

  Planning: using previous disaster data, models and simulations were constructed to help the communications infrastructure planning, healthcare patient's survival models in low resource hospitals after earthquakes and even hypothetical nuclear detonations in urban areas. Monitoring people's behavior (emotions and movements) during special events like disaster, political movements, New Year's Eve or evacuation situations.

  Data used in these applications was social network data, GPS data, meteorological and geophysical data from previous events.

  Early warning: traffic incident detection, smart-city's urban incident detection, crowd mobile sensors to help on warning and evacuation.

  Disaster reaction: human monitoring in time of disaster, how to manage information during urban crisis to avoid spreading panic, evaluating disaster impacts on road and using geo-tagged tweets.

- **Improvements of disaster data processing algorithms and architecture**: algorithm improvements to work with disaster data i.e. real time data processing of emergency data o cloud infrastructure, generation of association rules between geophysical data and disasters, algorithm to classify news during disaster, studying patterns of rumors on tweets during disaster. Architectural improvements examples are a platform to crowd-improve disaster related data, platform to discover data from diverse sources related to disaster and data storage systems for geotagged data.

CONCLUSIONS AND DISCUSSION

The conventional data sources for disaster management detected were and previous disaster data and geological or weather related data, emerging new data sources detected were, social media data, in specific twitter, crowdsourced data and Geospatial data.

Applications in the whole chain of disaster risk mitigation were detected as an unexpected result the applications on improvement of algorithms, platforms and data for disaster management applications.

Big data publications related to disaster are growing in time, Japan is still one of the leading countries in the development of these publications to the side of United States and China. The community with the major number of keywords and biggest average of centrality described, in general terms, the structure of the papers in the network. Yet when the other most central cluster was analyzed more specific applications were detected.

This study was limited to scientific paper publications, this left out other potential applications that are not reported in the scientific literature.
Further study could be done on to explore more in deep the nature of the data used for disaster management studies, it could be possible to study the data sources implemented in these new applications to inspire the design and development of other service and applications.
Big data is having an impact on disaster management and the number of scientific articles related to it shows an incremental trend. Looking at this evidence the authors share their hope that more companies and institutions join getting the benefits and solving the challenges that big data, cloud technologies, IoT among others phenomena are generating for disaster risk mitigation.

REFERENCES

Ávila-Robinson, A. & Miyazaki, K., 2013. Dynamics of scientific knowledge bases as proxies for discerning technological emergence - The case of MEMS/NEMS technologies. *Technological Forecasting and Social Change*, 80(6), pp.1071-1084. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0040162512001813 [Accessed June 1, 2014].

Geum, Y. et al., 2012. Technological convergence of IT and BT: Evidence from patent analysis. *ETRI Journal*, 34(3), pp.439-449.

Google, 2012. 東日本大震災ビッグデータワークショップ - Project 311 -. *Project 311*. Available at: https://sites.google.com/site/prj311/ [Accessed September 23, 2016].

Huang, Y. et al., 2015. A systematic method to create search strategies for emerging technologies based on the Web of Science: illustrated for "Big Data." *Scientometrics*, 105(3), pp.2005-2022. Available at: http://link.springer.com/10.1007/s11192-015-1638-y.

Ittipanuvat, V. et al., 2014. Finding linkage between technology and social issue: A Literature Based Discovery approach. *Journal of Engineering and Technology Management - JET-M*, 32, pp.160-184. Available at: http://dx.doi.org/10.1016/j.jengtecman.2013.05.006.

Kajikawa, Y. et al., 2010. Multiscale analysis of interfirm networks in regional clusters. *Technovation*, 30(3), pp.168-180. Available at: http://dx.doi.org/10.1016/j.technovation.2009.12.004.

Marr, B., 2015. Using Big Data In A Crisis: Nepal Earthquake. *Forbes*. Available at: http://www.forbes.com/sites/bernardmarr/2015/04/28/nepal-earthquake-using-big-data-in-a-crisis/#72617359532f [Accessed August 10, 2016].

NHK, 2013. NHK Special | ～ Disaster Big Data ～ . *NHK special*. Available at: http://www.nhk.or.jp/datajournalism/about/index_en.html [Accessed August 25, 2016].

Small, H., 2006. Tracking and predicting growth areas in science. *Scientometrics*, 68(3), pp.595-610. Available at: http://www.akademiai.com/index/711P28W6178K22X0.pdf [Accessed May 16, 2014].

Turney, P.D. & Pantel, P., 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37, pp.141-188.

UNISDR, 2009. 2009 UNISDR Terminology on Disaster Risk Reduction. *International Stratergy for Disaster Reduction (ISDR)*, pp.1-30. Available at: www.unisdr.org/publications.

UNISDR, CRED & EM-DAT, 2015. *The human cost of weather-related disasters 1995-2015*,

Van-RAAN, A.F.J. Van, 1988. *Handbook of Quantitative Studies of Science and Technology* A. F. J. VAN-RAAN, ed., Amsterdam: Elsevier B.V. Available at: http://www.sciencedirect.com/science/article/pii/B9780444705372500143.

Yasutomo, T., Cristian, M. & Yuya, K., 2016. Unconnected Component Inclusion Technique for a Patent Network Analysis Case Study of Internet of Things-related Technologies. *Journal of Informetrics*, 10(4), pp.967-980. Available at: http://dx.doi.org/10.1016/j.joi.2016.05.004.