e-ISSN: 2595-5527
10.32435/envsmoke.20214124-31
Volume 4, Number 1
2021

INFORMATION RETRIEVAL AND DATA MINING CASE STUDY: COVID-19 DASHBOARD ANALYSIS

Francisco Carlos Paletta[1]; *Luiz Wanderley Tavares*[2]

Abstract

Research data mining requires the use of information systems and complex methods of searching, accessing, retrieving, and appropriating information on the web of data. Platforms such as Google Scholar Google, Elsevier, JSTOR, ResearchGate, ScienceDirect and the ones provided by commonly used universities identified with search terms. The challenge is to ensure that researchers are being exposed to the state-of-the-art networked knowledge production. This study aims to analyze the process of searching and retrieving information, reflect on the role of information systems in the search result and the informational skills of the researcher in view of the quality of information retrieved. Case Study Covid-19 Dashboard Analysis.

Keywords: Complex Network Data Mining. Information Access. Seeking Engine. Browsing. Colleting. Rereading. Disseminating. Information Systems. COVID-19. Dashboard Analysis.

[1] University of São Paulo, Brasil. fcpaletta@usp.br

[2] University of São Paulo, Brasil. lwt@usp.br

Received: 19 Abr. 2021
Accepted: 23 Abr. 2021
Published: 30 Abr. 2021

1 Introduction

By the late 20th century, the Internet was already considered a revolution in the spread of knowledge. The potential for information integration and dissemination has made it the main source of data production, storage and distribution, and technological advances in information systems have led to the development of complex scientific data sharing networks.

The use of the Internet by science gained great encouragement in the early 1990s, when libraries began the processes of acquisition, cataloging, circulation, interlibrary lending, and reference services (LIU, 1995). In addition to libraries, newspapers and magazines began to make their catalogs available to increase visibility, credibility, and profits. With this new reality, the needs of users have been changing, generating less dependence on libraries and more databases available on the World Wide Web.

> The Internet has revolutionized the world of computers and communications like never. The invention of the telegraph, telephone, radio, and computer set the stage for this unprecedented integration of capabilities. The Internet is, at the same time, a form of worldwide transmission, a mechanism of information dissemination and a means of collaboration and interaction between individuals and their computers (LEINER, 2009, p.22).

Data retrieval is increasingly performed in academic search databases and Google (HEAD and EISENBERG, 2009). The database choices to query are related to the credibility of the content and the detailing of the retrieved information. Google's recent forays into academic content and mass digitization have blurred the unclear distinctions between libraries and commercial data storage services (DETLOR and LEWIS, 2006).

The query in these databases is performed through words and terms associated with the subject to be searched and the process are discontinuous, because users have no control over the search methods performed by platforms (ZHANG, 2007). This search process is totally dependent on the words and terms used, forcing the searcher to know more about the topic to be consulted for better results.

However, having no control over the outcome raises some questions about their answers. How can one be assured of being exposed to the best and most relevant articles? Regarding relevance, what is the metric adopted by the most used systems?

Therefore, the researcher needs to lean over on his searches. The need to consult mentors, nontrivial databases, such as those in university libraries, search by chaining references, and browse web pages on the subject increase the chances of finding the most relevant studies.

2 Information Research and Recovery

Data retrieval systems on the Web Data are gaining relevance with the emergence of online databases. ELLIS (1989) cited six search characteristics: start, chain, browse, differentiate, monitor, and extract. They were expanded by MEHO and TIBBO (2003) to include: access, network, verification, and information management.

All these characteristics were identified by analyzing the behavior of researchers before and after the advent of the Internet. However, the evolution of the digital universe has come to demand a more complex and intensely dependent approach to data mining and retrieval. PALMER et al. (2009) defined five core activities: searching, collecting, reading, writing and collaboration. These activities are acts performed by researchers when they are collecting information on the subjects covered in their research and all are dependent on the quantity and quality of information found.

> The information seeking patterns of a variety of social scientists have been divided into six

> characteristics: starting, chaining, browsing, differentiating, monitoring, and extracting. These characteristics constitute the main generic characteristics of the different individual patterns and together provide a flexible behavioral model for information retrieval system design (ELLIS, 1989, p.171).

In addition to this control, there are platforms and tools providing searches on all types of data, where the user can quickly identify similar documents. These tools help you better control collections by facilitating access to relevant documents already collected and creating a network of relationships between them - Linked Data. Classification methods using cluster techniques and supervised machine learning, such as Naive Bayes and Support Vector Machines (SVM), help the researcher create a fast and efficient way to support his research (MANNING, RAGHAVAN and SCHÜTZE, 2008).

Online search engines must deal with the Big Data phenomenon and the quality of data mining depends on information retrieval methods and tools. Typically, metadata and algorithms are transparent to users and affect their direct access to information. However, some platforms give the user greater control over sorted and selected material, allowing for better information organization.

2.1 Searching

The search process begins with defining how the search will be performed. Its activities are: direct search, chaining, browsing, polling and access. Basically, the process starts with searching for terms and keywords in search engines. This search is term-sensitive, requiring the searcher to search through their lists and select articles to search. Among the chosen, the chaining by references helps to identify other authors and texts relevant to the research.

2.2 Collecting

Collections are formed by the need for long-term access. Printed documents have been systematically replaced by digital documents. The amount of digitized material obliges researchers to gather their collections in an organized manner, because they are becoming larger and easier to collect. The organization of the material should consider the storage location of digital documents, where repositories on the Internet provide virtual space and policies for the protection and preservation of digital data.

2.3 Reading

The amount of material available on the Internet has been changing the ways researchers interact with texts. Exposure to more and more data sources is increasing the number of articles read. However, with the use of information visualization tools the reading time and data analysis has been significantly reduced. This practice has generated the ability of researchers to evaluate documents quickly to determine their relevance and usefulness. Still in this activity, the act of rereading the articles is common, reason for the creation of the collections. Rereading increases understanding and integrates research with work in progress.

2.4 Writing

Writing is part of the academic activity, where the researcher compiles the data, generating a new academic material. Co-authoring is common and has been expanded by the ease of communication generated by the Internet. The dissemination of works can be done through newspapers, conferences, digital repositories of schools and the Internet (blogs, social media, chat, etc.).

2.5 Collaborating

Collaboration among researchers is one of the most leveraged topics on the Internet. There are a multitude of collaborative platforms where participants can exchange data, text, images, etc. However, there are some difficulties regarding the distance and number of participants. However, the ease of consulting

partners and advisors is undeniable.

PALMER et al. (2009) points out how these activities increase the responsibility of librarians, responsible for collecting primary and secondary data and for the processing, preservation and archiving necessary for sharing and reuse.

Search results are increasingly dependent on the quality of search strategy, data mining and information retrieval. However, research is becoming increasingly data intensive. For Cragin and Shankar (2006), collective data management does not fit on existing library models, reinforcing the role of electronic repository searches and the need for procedures to ensure access to less publicized content.

2.6 Browsing and Searching

To have a worldwide dimension of the complexity of the World Wide Web, there are over 1 billion active sites with hundreds of libraries making their collections available online.

With this information capillarity, the researcher has the need to better organize his search actions. Knowing only the most used terms and words is no longer enough. Open-ended research on the subject can bring up words and unthought terms. This act of navigating the results and generating new research shows how complementary the two acts are (ZHANG, 2007). Browsing multiple sites is important for exploration and discovery while searches are more straightforward. The user performs the usability and relevance of the navigation result, while the user loses control over the choice in the case of search terms. Browsing is a laborious and time-consuming act compared to searches, users need to remember the navigated paths, understand the content, and decide new directions constantly. For BATES (2002), navigating involves successive acts of glimpsing, fixing a target to examine visually or manually more closely, examining, and moving on to restart the cycle again.

Thus, the creative process of research is cyclical and never-ending. Browsing, searching, examining, collecting, and restarting the process are repeated attitudes even after the research work is completed, as they can generate materials for new papers.

Navigation is the main activity in the search for content and research design, it must be done broadly, but it needs to be an organized procedure. The author should keep documented the procedures performed for any need to redo them.

The researcher also needs to seek “help” from his peers; other researchers may come up with new websites and ways of thinking unthought. Using collaborative tools and finding co-authors can help with the whole creation process by splitting activities and exploring by more than one person.

Other activities may also interact with the research. Reference chaining procedure provides information for new searches by restarting the collection extension process.

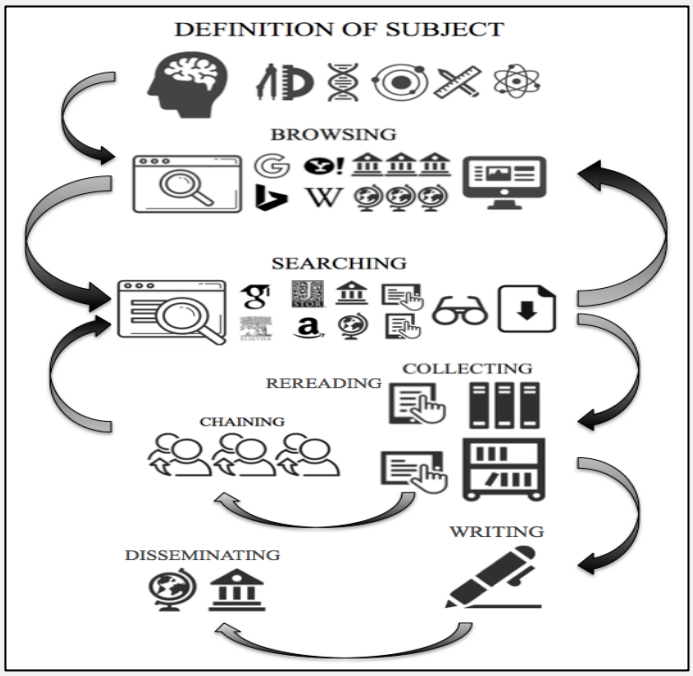

Fig. 1. Diagram of academic activities in reference research. Source: Author.

The activity diagram (see Fig. 1) exemplifies the process of academic creation using the Internet

and shows how navigation and research are intertwined and of different natures. While navigation is geared towards a broader search, the search is more objective in getting the results.

After a good exploration of data, the research work becomes easier, since the identification of terms and keywords, as well as authors and search places, has already been done. Thus, research becomes the time to select and catalog the materials to be worked and used in the process of knowledge production.

The created collection can be stored in computer systems capable of classifying them and performing a set of analyzes predefined by the researcher. The advantage of using this form of storage is in the guarantee against data loss as well as the recovery of the desired material.

The above concept shows ways to prevent recurrent myopia from searching in a few databases, generating an ability to further expand and consolidate work.

4 Citespace

Cheng's article (2006) presents the CiteSpace II program as a tool to operationalize information retrieval systems. The system is based on a platform for reading the data (metadata) of selected articles and relating them to each other. The most used database for creating data for CiteSpace II is the Web of Science.

> Emerging trends and abrupt changes in the scientific literature may be associated with internal and external causes. Typical internal causes include new discoveries and scientific discoveries, such as the discovery of an impact crater in mass extinction research or the discovery of a supermassive black hole in astronomy. External can provoke scientists to study a subject from new perspectives. (CHENG, 2006).

As a way of showing its use, a search with the term “machine learning” was performed in the Web of Science and the search returned 753 articles on the subject.

One of the great capabilities of the tool is to elucidate the “classic” articles to the reader and to present the new trends chronologically and with the intensity (citations) of each article (see Fig. 2)

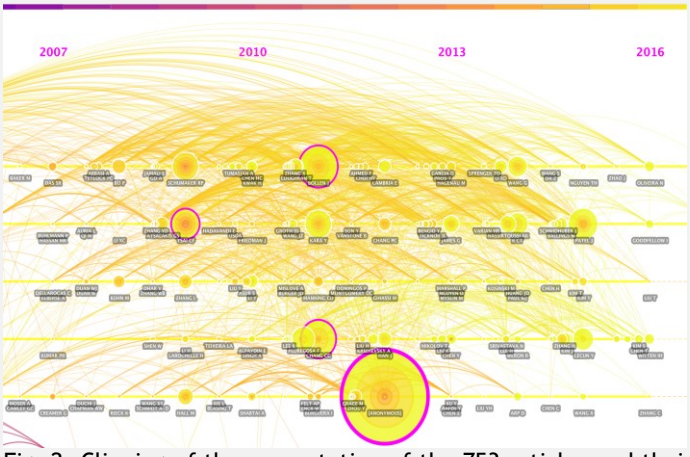

Fig. 2. Clipping of the presentation of the 753 articles and their references. Source: CiteSpace.

The visualization presented is one of several possible and help the researcher to have a more global view on the searched subject. By knowing the most used terms, one can define the line of research to be adopted, either a continuation of a topic already addressed or the introduction of a new approach to the topic.

5 Case Study: *COVID-19*

The current information stress caused by the COVID-19 pandemic demands the use of powerful information gathering and analysis tools.

The Johns Hopkins Coronavirus Resource Center provides accurate, real-time data on the pandemic via the COVID-19 Dashboard by the Center for Systems Science and Engineering CSSE - https://coronavirus.jhu.edu/map.html

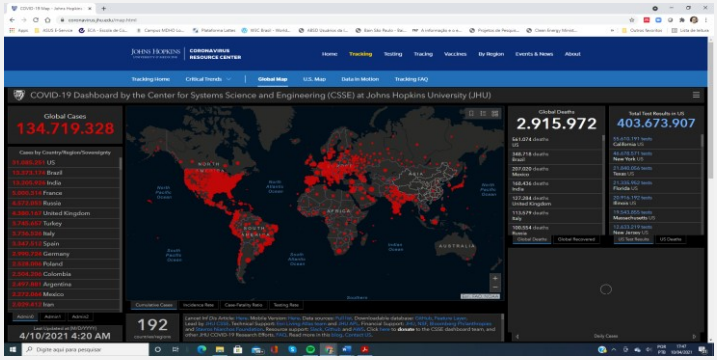

Fig. 3. COVID-19 Dashboard by Johns Hopkins Coronavirus Resource Center

In this context, Poynter Institute - presents itself as a reliable source of information for researchers and professionals working in the management and organization of information about Coronaviruses.

> Led by the International Fact-Checking Network (IFCN) at the Poynter Institute, the #CoronaVirusFacts / #DatosCoronaVirus Alliance unites more than 100 fact-checkers around the world in publishing, sharing, and translating facts surrounding the new coronavirus. The Alliance was launched in January when the spread of the virus was restricted to China but already causing rampant misinformation globally. The World Health Organization now classifies this issue as an **infodemic** — and the Alliance is on the front lines in the fight against it. https://www.poynter.org/

The following graphic shows the categories of fact-checks, which are helping our alliance identify successive waves of misinformation as they travel across the globe. When the new coronavirus pandemic started, many hoaxes were about the origin of the virus. Then the alliance detected falsehoods on how the disease spreads, and cures and preventions. Now we are finding hoaxes about religious groups, politicians, and the impact of COVID-19 on a country's health system. (Poynter, 2021)

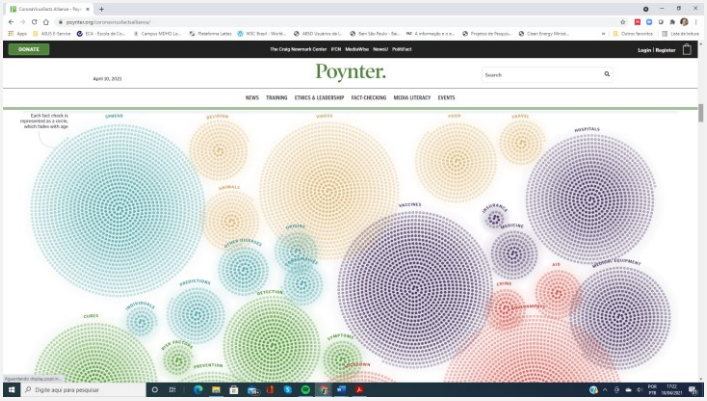

Fig. 4. Waves of Hoaxes - Source: https://www.poynter.org/coronavirusfactsalliance/

WHO Coronavirus (COVID-19) Dashboard is another reliable source of information that can assist the researcher in mining relevant and up-to-date data on the topic.

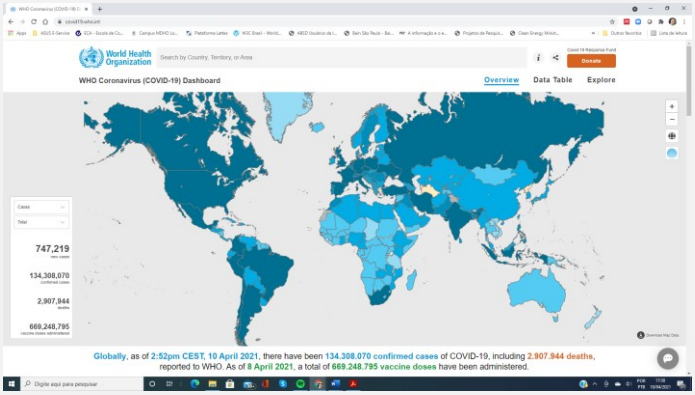

Fig. 5. WHO Coronavirus (COVID-19) Dashboard - Source: https://covid19.who.int/

The process of data mining is intrinsically linked to the quality of the sources of information and will have a significant impact on the quality of research. It is up to the researcher to develop skills and competencies in the process of searching, accessing, retrieving, appropriating, and using information in the production of innovation and new knowledge.

6 Discussion

This paper aims to analyze the complexities associated with data mining in research activities and to reflect on the search, access, retrieval and use of information tools in the Data Web and present a case study on research data present a case study in the use of Covid-19 information

sources.

Researchers need to develop digital skills in the use of information systems and computational resources in the process of organizing information and knowledge. With the help of technology, they must cross-reference information of the materials found, broadening the search within your area and in other areas where the subject matter may have been studied.

Thus, the excess of information is now considered as another point of concern, where researchers are increasingly required in evaluation tasks and choice of material to be used. The technology assists in the information retrieval process, with machine learning methods capable of classifying and determining the main subjects addressed in the processed texts.

Collaborative work with various authors in different academic environments can help enrich the discussion and identify other views on the topics. In this respect, it opens the potential for further studies on virtual collaborative environments.

> The current and future impact of the Internet on our society is one of the most complex and participatory discussions of the moment. An emerging issue of this discussion has been the redefinition of the role of libraries, raising the question of what is (or will be) a "digital library" in a networked world (BORBINHA & DELGADO, 1996).

Other studies can be conducted to elucidate how machine learning methods are able to identify similar or complementary studies across distinct areas of knowledge, broadening the understanding of knowledge more broadly.

The future of research is increasingly intertwined with computer systems, complex networks, and the use of artificial intelligence to streamline information search and retrieval processes and the production of knowledge in the digital age.

References

Bates, M. J. (2002, July). Speculations on browsing, directed searching, and linking in relation to the Bradford distribution. In Emerging frameworks and methods: Proceedings of the Fourth International Conference on Conceptions of Library and Information Science (CoLIS 4) (pp. 137-150). Greenwood Village, CO: Libraries Unlimited.

Borbinha, J., & Delgado, J. (1996). Internet and the New Library. In Proceedings of NIT'96: The 9th International Conference on New Information Technology (pp. 11-14).

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. Journal of the American Society for information Science and Technology, 57(3), 359-377.

Cragin, M. H., & Shankar, K. (2006). Scientific data collections and distributed collective practice. Computer Supported Cooperative Work (CSCW), 15(2-3), 185-204.

COVID-19 map. 2021. Retrieved April 10, 2021, from https://coronavirus.jhu.edu/map.html

Detlor, B., & Lewis, V. (2006). Academic library web sites: current practice and future directions. The Journal of Academic Librarianship, 32(3), 251-258.

Ellis, D. (1989). A behavioural approach to information retrieval system design. Journal of documentation, 45(3), 171-212.

Head, A. J., & Eisenberg, M. B. (2009). Lessons Learned: How College Students Seek Information in the Digital Age. Project Information Literacy Progress Report. Project Information Literacy.

Leiner, B. M., Cerf, V. G., Clark, D. D., Kahn, R. E., Kleinrock, L., Lynch, D. C., ... & Wolff, S. (2009). A brief history of the Internet. ACM

SIGCOMM Computer Communication Review, 39(5), 22-31.

Liu, L. G. (1995). The Internet and Library and Information Services: A Review, Analysis, and Annotated Bibliography. Occasional Papers No. 202.

Mantas, H., & Tardáguila, C. (2021, March 11). Coronavirusfacts alliance. Retrieved April 10, 2021, from https://www.poynter.org/coronavirusfactsalliance/

Meho, L. I., & Tibbo, H. R. (2003). Modeling the information-seeking behavior of social scientists: Ellis's study revisited. Journal of the American society for Information Science and Technology, 54(6), 570-587.

Palmer, C. L., & Cragin, M. H. (2008). Scholarship and disciplinary practices. Annual review of information science and technology, 42(1), 163-212.

Palmer, C. L., Teffeau, L. C., & Pirmann, C. M. (2009). Scholarly information practices in the online environment. Report commissioned by OCLC Research. Published online at: www. oclc. org/programs/publications/reports/2009-02. pdf.

Manning, C., Raghavan, P., & Schütze, H. (2010). Introduction to information retrieval. Natural Language Engineering, 16(1), 100-103.

Zhang, J. (2007). Visualization for information retrieval (Vol. 23). Springer Science & Business Media.

WHO coronavirus (COVID-19) Dashboard. (2021). Retrieved April 10, 2021, from https://covid19.who.int/

Webgrafia

https://www.jstor.org/
https://www.elsevier.com/
https://www.scopus.com
https://www.sciencedirect.com/
https://www.mendeley.com/
https://registro.br/
http://cluster.ischool.drexel.edu/~cchen/citespace/
https://coronavirus.jhu.edu/map.html
https://www.poynter.org/coronavirusfactsalliance/
https://covid19.who.int/

***Acknowledgment*:** FAPESP – Processo 19/01128-7