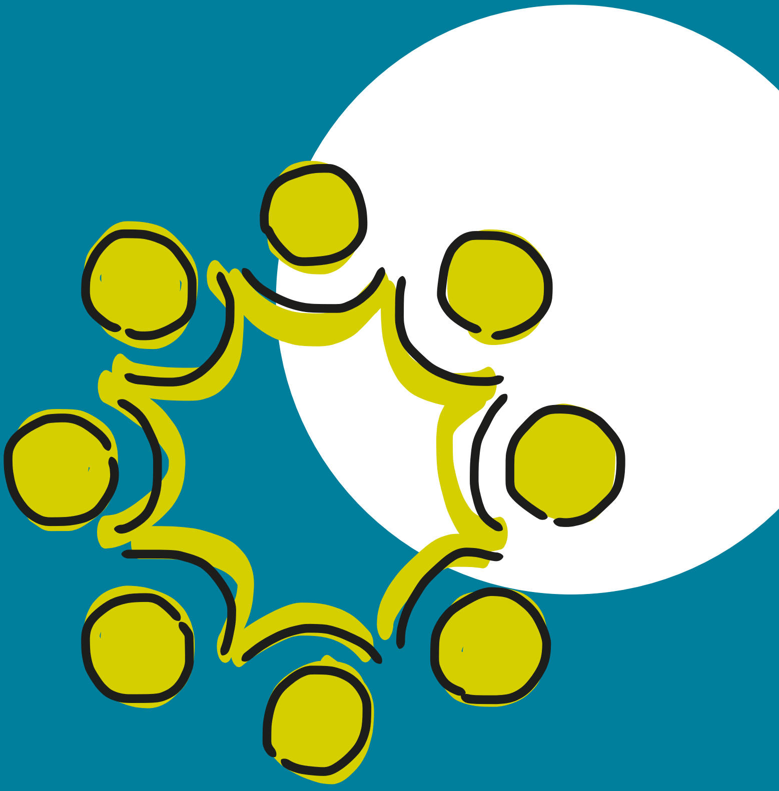
Editors
Andreas Wagner
Michael Granitzer
Christian Guetl
Christine Plote
Stefan Voigt

Proceedings

4th International Open Search Symposium

#ossym2022

10-12 October 2022
CERN, Geneva, Switzerland



open search foundation

CERN

ISSN: 2957-4935
ISBN: 978-92-9083-646-9

Impressum

Editors:

Michael Granitzer, University Passau, Germany
Christian Güetl , Graz University of Technology, Austria
Christine Plote, Open Search Foundation, Germany
Stefan Voigt, Open Search Foundation, Germany
Andreas Wagner, CERN, Geneva, Switzerland

ISSN: 2957-4935
ISBN: 978-92-9083-646-9
DOI: 10.5281/zenodo.8399978

This report should be cited as:

Proceedings of 4th International Open Search Symposium #ossym2022, CERN, Geneva Switzerland, 10-12 October 2022, M. Granitzer, C. Gütl, C. Plote, S. Voigt, A. Wagner (eds). http://doi.org/10.5281/zenodo.8399978

More information

- Conference Website on CERN Indico
  https://indico.cern.ch/e/OSSYM-2022
- Open Search Community at Zenodo
  https://zenodo.org/communities/opensearch/
- Event information at the Open Search Foundation
  https://opensearchfoundation.org/en/events-osf/4th-international-open-search-symposium-ossym22/

Foreword

The 4th International Open Search - #ossym – was held at CERN, Geneva, Switzerland from 10-12 October 2022. It was a great experience to meet everyone in person again after two years of online meetings due to the pandemic. The participants on site enjoyed the personal exchanges, discussions, coffee-breaks, and social events to exchange and advance scientific and personal topics. The symposium saw a series of inspiring keynotes as well as many interesting talks in the research tracks and in the industry and application-oriented sessions. Furthermore, very interesting considerations were shared in the more interactive sessions on search quality and ethics as well as in search literacy and education parts.

The community flagship project OpenWebSearch.eu had just begun a few weeks prior to the symposium, thus many talks and presentation gave a perspective on what research and piloting activities are expected to happen during the three years of this project. All in all, it was a very lively and inspiring meeting, fostering the exchange of many different disciplines involved in the Open Search Initiative: from computer science, information retrieval, law, ethics, economy, industry, education, media-science and many more.

We are very happy to present herewith the 2022 volume of #ossym proceedings, which documents the symposium in 2022 and its many innovative and interdisciplinary contributions, advancing Open Search. We kindly want to thank all the authors, presenters, panellists, and keynote speakers for sharing their science work, ideas, expertise and thoughts at #ossym and within these proceedings.

Finally, we want to extend a warm thanks to CERN as local host and co-organiser of #ossym2023! CERN in its structure, scope and organisational set-up demonstrates impressively that some large-scale science tasks can only be achieved in systematic and humble cooperation of sharp minds, working in synergy with large scale technological facilities. Furthermore, it reminds us, that with impactful and fundamental research findings - as is the case with nuclear physics - large responsibilities emerge for scientists, technicians, decision-makers, and society as a whole: to mindfully and responsibly handle the capabilities that science and technology provide us with! Although very different in nature, Open Search and open, unbiased access to data and information via a public open Web index, similarly, require cutting edge science, large-scale and distributed computing-facilities, clear ethical and legal guard-rails as well as a sound governance to be operated as a public and open infrastructure in the future. We hope that #ossym2022 marked another tiny, and we are hopefully worthwhile, stepping-stone on our joint venture towards open and unbiased information access in the Web and Open Search in general.

Enjoy reading these proceedings and we are looking forward to the next instances of #ossym in the years to come.

Together, for a better net.

Andreas Wagner, Michael Granitzer, Christian Gütl, Christine Plote and Stefan Voigt

Symposium Organisation

Programme Committee

Wolf-Tilo Balke, L3S Research Center, Braunschweig, Germany
Alexander Decker, Technische Hochschule Ingolstadt, Germany
Kai Erenli, University of Applied Sciences BFI Vienna, Austria
Arjen P. de Vries, Radboud University, Netherlands
Stefan Dietze, Heinrich Heine University Düsseldorf and GESIS, Germany
Maria Dimou, CERN, Geneva, Switzerland
Christian Geminn, University Kassel, Germany
Michael Granitzer, University Passau, Germany
Christian Guetl, Graz University of Technology, Austria
Andreas Henrich, University Bamberg, Germany
Nils Jensen, Ostfalia University of Applied Science, Wolfenbüttel, Germany
Mohammed Kaicer, Faculty of Sciences Kenitra, Morocco
Dennis-Kenji Kipker, Riga Graduate School of Law, Latvia
Dieter Kranzlmüller, Leibniz Supercomputing Centre and LMU, Munich, Germany
Dirk Lewandowski, University of Applied Science, Hamburg, Germany
Jelena Mitrovic, University of Passau, Germany
Wolfgang Nejdl, L3S Research Center / Univ. Hannover, Germany
Engelbert Niehaus, University Koblenz-Landau, Landau, Germany
Alexander Nussbaumer, Graz University of Technology, Austria
Philipp Mayer-Schlegel, GESIS, Germany
Monica Palmirani, Università di Bologna, Italy
Martin Potthast, Leipzig University, Germany
Tobias Schreck, Graz University of Technology, Austria
Christin Seifert, University of Duisburg-Essen, Germany
Gianmaria Silvello, University of Padova, Italy
Tim Smith, CERN, Geneva, Switzerland
Benno Stein, Bauhaus-Universität Weimar, Germany
Olivia Tambou, Paris Dauphine University, France
Francesca Tomasi, Università di Bologna, Italy
Marco Verile, European Commission – Joint Research Centre (JRC), Ispra, Italy
Stefan Voigt, Open Search Foundation, Germany
Andreas Wagner, CERN, Geneva, Switzerland

Symposium Organising Committee

Michael Granitzer, University Passau, Germany
Christian Gütl , Graz University of Technology, Austria
Christine Plote, Open Search Foundation, Germany
Stefan Voigt, Open Search Foundation, Germany
Andreas Wagner, CERN, Geneva, Switzerland

Symposium Local Support

Aleksandar Bobic, CERN & Graz University of Technology, Austria
Igor Jakovljevic, CERN & Graz University of Technology, Austria

Contents

Research Track Information

ASE	-	Alternative Search Engines
FPC	-	Future Open Search Paradigms and Concepts
ISA	-	Innovative Search Applications, Open Search Use Cases and Challenges
LAS	-	Legal Aspects of Open Search
MLM	-	Machine Learning, Web Mining, Content Retrieval and Web Analysis
SQE	-	Search Quality and Search Ethics

https://doi.org/10.5281/zenodo.7799190

NGI SEARCH: THE NEED FOR TRUST AND PRIVACY IN SEARCH, DISCOVERY AND INDEXING

Aurora González-Vidal*[1], Antonio F. Skarmeta[1], Mirko Presser[2], Marie Claire Tonna[3], Manuel Noya[4], Pierre-Yves Gibello [5]

[1] Dep. of Information and Communication Engineering, University of Murcia, Murcia, 30100, Spain

[2] Dep. of Business Development and Technology, Aarhus University, Herning, 7400, Denmark

[3] FundingBox Accelerator Sp. z o.o., Warszawa, 02-305, Poland

[4] Linknovate Science S.L., Santiago de Compostela, A Coruña, 15896, Spain

[5] OW2, 7 rue de Phalsbourg, 75017 Paris, France

Abstract

Internet-based data sources and resources continue to grow exponentially, making the mechanisms for searching and discovering insights, and making sense of data, a crucial field of research. The objective of Next Generation Internet (NGI) Search is to support innovative projects to develop trustworthy solutions towards the development of new ways of searching data by addressing the challenges of power cognitive search, natural language processing and social computing amongst other cutting-edge fields. The projects will be compliant with open, collaborative and unbiased values. NGI Search will offer five Open Calls to find and select projects proposed by talented researchers, innovators and activists (NGI Surveyors) working in search and discovery within a human-centric context (meaning privacy-aware and trust-oriented) as well as vertical use-cases developed jointly with the industry. This will lead to more transparency and choice with a focus on privacy and trust, contributing to the overall vision of a more human-centric Internet.

INTRODUCTION

Since the inception of the internet, applications and services using the internet have evolved significantly. In addition, the number of devices connected to the internet, the creation of vast data lakes as well as distributed data cooperatives has made searching and discovering data and generally resources, a difficult, yet very important field of research, development and innovation.

Search is one of the most intimate and uncovering mechanisms on the internet. To efficiently search we need to know what is discoverable looking at all possible sources. In addition we also cannot hide what we are searching for when we make the inquiry. Both sides to search are immensely powerful pieces of knowledge that have been monetized through rigorous and systematic analysis.

Today it is even more important to be able to find, represent and ultimately make sense of internet-based data sources and resources without the need of relying on non-transparent organisations offering such services, potentially violating privacy and trust.

The Next Generation Internet (NGI) [1] is a European Commission initiative that aims to shape the development and evolution of the Internet into an Internet of Humans. The initiative has already supported around 1,000 Internet researchers and innovators involved in many hundreds of projects. Support goes beyond financing, to mentoring and the journey from an idea to a real business.

The Vision of the NGI Search project, under the umbrella of the NGI initiative is to change the way we use and experience, search and discover data and resources in general, on the internet and web. This includes new user interfaces for searching and representing data (e.g. voice and image based), heterogeneous data sources (e.g. IoT, semantic data, multimedia, social media as well as traditional websites), new methodologies (e.g. machine learning and natural language processing) in generic context and specific use cases (e.g. industry 4.0, health, social media).

The third party prototypical projects that NGI search will support will address the problems related to sharing information by offering privacy and trust with respect to search. This could manifest in projects addressing new user interfaces on how people search - masking who and how people search; it could result in new decentralised discovery mechanisms; or novel ways of describing data quality and validity considering rapidly growing contributors to data such as social media, Internet of Things (IoT) and multimedia.

The NGI Search project Mission is to help develop technologies and solutions enabling new and trustworthy ways of searching and discovering information on the internet across a variety of resources such as personal, scientific, industrial and environmental data, connected devices and smart objects, services, multimedia content, intranets and other ICT resources, both public and private. It also aims to empower end-users, including through agents acting on their behalf, to share and discover more data and reliable information sources, while preserving their privacy and increasing public trust in search results. This mission will be achieved as follows:

- Intelligently scout and run 5 Open Calls to find and select talented researchers, innovators and activists (NGI Surveyors) working in search and discovery within a human-centric context (meaning privacy-aware and

* aurora.gonzalez2@um.es

[1] https://www.ngi.eu/

https://doi.org/10.5281/zenodo.7799190

trust-oriented) as well as vertical use-cases developed jointly with industry.

- Support and mentor the selected candidates over a 12-month custom program using 10 value-added services.
- Integrate them into the NGI community, in particular the sister RIA on infrastructure, as well as communicate and disseminate their results to help the uptake and use of their developed technologies and solutions.
- Synergise with national, regional and international initiatives to establish a peer-based system of quality and dissemination at the global level.

The NGI Search project's core values are

1. Open Source (minimum open core) development and real working code that is in a deployable state;
2. Contributions to standards and larger communities that are already working on solutions or have a solid track record in the community;
3. Collaboration between researchers, innovators and activists on deep technology to provide a foundation towards entering the market as a standard, open source project and/or commercial service;
4. Adhere to the Open Science principle;
5. Address transversal challenges, specifically gender and sustainability challenges.

STATE-OF-THE-ART AND PROJECT CHALLENGES

NGI Search looks for proposals addressing deep-tech development and research-based solutions that address the NGI initiative at the core of their developments whilst supporting the outcomes of the abovementioned vision and mission. The following list of topics is a set of problems that the consortium has identified upfront under the topic of Search and Discovery. These topics will stimulate project submissions on these challenges, and additional challenges will also be considered. This Work Programme will be updated during the project, to match the key challenges defined for each Call with the rapid developments in this area of research, as well as to reflect the outcomes of previous Open Calls, achieving a comprehensive set of sub-projects.

Project topics and challenges

The next generation of intelligent voice-based assistants. Voice assistants are used for routine tasks, such as asking for the weather or making a phone call. They can have positive social influences, since they can decrease depression and simulate interest in physical activity [1]. However, as familiarity increases, more complex tasks will require the addition of intelligence to the assistants [2]. Traditional written query formulations are simpler than voice-based searches and analysing the word choices and interactions provides more context about the intent of the user. Some challenges relate to situationally induced impairments and security, since the search can be done while performing other activities and in public spaces (the assistant demands private information) and to mixed modal interactions, where questions and answers not only have voice content but also images, text etc.

Power cognitive search by reinforcement learning. Cognitive search uses Artificial Intelligence (AI) to improve users' search queries and extract relevant information from diverse data sets while providing automated tagging and personalization. While cognitive computing is widely known [3], the term cognitive search is linked to companies such as Microsoft with the Azure Cognitive Search. Clustering and classification can help limit searches to specific groups, and building similarities can synthesize the interactions between data. We encourage the development of mechanisms that contribute to a reinforcement learning system able to learn from the interactions how to choose the data and algorithms to make a search more relevant.

Natural language processing. Natural Language Processing (NLP) methods are widely in use in the area of machine translation. The majority of search engines use the huge amounts of previously accumulated user requests for predicting the search output without taking into account the user's intention [4]. Model complexity of the current state-of-the-art models is increasing and implies the use of great amounts of energy for computation. At the same time, they assume that each device would have full access to powerful processors, generous memory, and, generally, cloud connectivity. This may not be possible for many edge devices. We want to initiate the paradigm of tinyNLP by searching ways to adapt NLP methods to edge and fog computing, studying ways to apply transfer learning in NLP and improve the energy efficiency of the current NLP approaches towards simpler and more sustainable NLP research and practices.

Machine-based data (IoT). With the rapid increase in the observation and measurement data emerging from IoT deployments in open networks such as open urban data portals or intranet sensors, finding and accessing the data is becoming a challenge [5,6]. Most of the current IoT systems rely on meta-data descriptions and have limited means to search for the patterns within the IoT data stream. In this sense, it is necessary to enable the search and discovery of information based on historical data and pattern extraction by means of algorithms that can adapt to the characteristics of the IoT sources: geospatial information, events and time series.

Semantic analysis. Semantic data integration generates a common representation of concepts and their relations using domain knowledge formalisms in the form of ontologies and reasoning capabilities and therefore, can aid in the integration of heterogeneous data [7]. Information about a subject or topic might be spread across different data sources so there exists the need for the integration of the knowledge. Question answering and data analytics can make use of such

https://doi.org/10.5281/zenodo.7799190

knowledge which in turn can be applied for decision making, that should be many times based on near real time data. The next steps on the semantic analysis field could be to search dynamic relations between concepts using "fresh" data and to minimize query execution while maximizing answer completeness based on federated principles. Federated query processing techniques integrate data from autonomous, distributed, and heterogeneous sources in a uniform way [8].

AI-based taxonomies. Taxonomies consist of machine-interpretable semantics and provide valuable knowledge for many applications such as product recommendations and enhancing query understanding. With the fast-growing volume of web content, existing taxonomies will become outdated and fail to capture emerging knowledge [9]. At the same time, these generic taxonomies cannot satisfy user's specific interests. Moreover, the nature of instance taxonomy treats each node as a single word, which has low semantic coverage [10]. We encourage research about the automatic creation and expansion of taxonomies by means of AI techniques that model inter-dependency among new concepts.

Network analysis. Complex network analysis, including time series network analysis [11], can be linked to semantic modeling in the sense that the outcome of such processes is an interlinked network of distributed resources which can be queried. These structures are known as knowledge graphs and they can be leveraged for computing centrality, clustering, etc. to gain insights about the domain being described [12]. Formal semantics for property graphs to derive conclusions based on taking into account the meaning of labels and property-value pairs on node and edges, similarity-based query relaxation (approximate answers to exact queries), shape induction, expressive graph neural networks and rule and axiom mining should be addressed. Scalability, quality of the induced models, diversity on the managed data and dynamicity (use of streaming data) are also general challenges of knowledge graphs to account for.

Social computing. The development of technologies that require interaction with humans imposes an interesting challenge since they have to succeed in improving motivation, encouraging participation and enhancing the learning process for their success. The interaction between social behaviour and technologies needs to be addressed in order to reach substantial changes in the behaviour of the adopters [13]. Human-related data presents big data characteristics and therefore, edge social computing should be considered in these scenarios in order to process and filter data of the network to reduce bandwidth costs, storage and energy consumption [14]. The implementation of edge social computing by means of context-aware learning, collaborative learning and other proposals in this direction are encouraged.

Data visualisation. Data visualization has attracted much attention recently, calling for joint actions in different research fields such as information visualization, human-computer interaction, machine learning, data management and mining, and computer graphics [15]. We seek interactive tools and mechanisms that allow visualizations for machine learning results that can provide user recommendations and support user-driven actions. This includes new applications of visually-driven analysis of spatio-temporal, textual and other kinds of data, progressive visualizations (in batches) and other kinds of scalable and efficient solutions [16, 17].

Enabling new ways of discovering and accessing information. Due to the rapid development of the IoT and the variability and volume of data sources, mechanisms for searching and integrating data are essential to leverage all relevant knowledge for improving processes and services [18]. New ways of discovering information need to be created in the form of platforms and products that deal algorithmically with data. The integration of data-driven machine learning with human knowledge can effectively lead to explainable AI [19] that would provide us ways to discover and access information where only raw data is present. The Challenge is to develop new algorithms and methodologies to discover and access information by combining Big Data technologies.

Addressing verticals - "Service discovery and composition in smart environments". The convergence of AI and IoT is changing the way systems are operating, from manual towards a more intelligent, efficient and automatic way. However, this is not an easy step, and architectures and methodologies are needed in order to discover the services that can be implemented depending on the installed sensors [20, 21]. For service discovery and composition, three principal functionalities are identified (i) a semantic functional description of the environment's objects, (ii) a distributed service directory that embodies available services for service lookup and discovery, (iii) planning tools for selecting and chaining basic services to compose new complex services. The challenge consists of proposing solutions to discover and compose services implemented across at least 2 verticals. Some examples of verticals are the following, but the scope is not restricted to them: smart cities, smart buildings, smart homes, industrial IoT, transportation logistics, smart oil & gas, smart agriculture, e-health.

Other topics. This may include Federated Search, enabling a user to search several different data sources at once by making a single query [22], Federated Learning for data sharing in decentralised machine learning applications [23–25], Transfer Learning that transfers the knowledge obtained using AI between domains with different characteristics [26] and data segmentation and representation methods, that reduce the dimensionality of data while maintaining the information that it contains, easing search and discovery [6]. Conversational search, that refers to the use of complete sentences and verbal units in search queries, zero-query search, that are systems that push information to the users based on their context and not on a specific query, and reproducibility of search are other very hot topics in information retrieval [27] that are within the scope of the project.

Transversal challenges - Gender dimension

AI simulates human behaviour, including voices, patterns, personalities, and appearances. Models exhibit gender-biased in multiple parts such as the training data, resources,

https://doi.org/10.5281/zenodo.7799190

pre-trained models and algorithms themselves. Various customer-facing service robots around the world feature gendered appearances that contribute to gender-task associations. In that sense, the prominence of female-sounding voice assistants encourages stereotypes of women as submissive and compliant. This is known as representation bias, which is when associations between gender with certain concepts are captured in word embedding and model parameters. As a domain of AI, NLP models may also propagate and even amplify gender bias found in text corpora. This can be seen when models often behave better on data associated with majority gender (allocation bias) and has been observed when automatic resume filtering systems give preference to male subjects. We will make sure that the developed solutions depict gender characteristics in a fair way, respond to gender-based harassment, and improve diversity within the solutions so that it makes our society advance in gender equality terms. We will encourage the use of debiasing mechanisms such as data augmentation and resampling [28], promote gender tagging , and bias fine-tuning [29], that incorporates transfer learning from an unbiased data set to ensure that a model contains minimal bias before fine-tuning so that we grow our solutions towards Responsible and Explainable AI in gender terms.

METHODOLOGY

During a 12-month support program, successful applicants will receive dedicated support according to their needs. This support comes in the form of 10 added-value services which are split on three levels: technical, business and finally innovation, which acts as more of a transversal service.

Technical support

1. Technology mentoring and advice on technologies for storing, managing and accessing data, advice on tools, infrastructures, platforms and software according to the size and goals of the project, suggestions on the algorithms that need to be programmed as baseline and on the strategies to achieve novel results. The advice will also include support for standardization and collaboration with different stakeholders.

2. Beta testing — the project will leverage ReachOut beta-testing platform[2] to support NGI-Search beneficiaries with its expertise and will support them during the whole process of preparing and implementing their beta-testing campaigns.

3. Efforts to link to Standards and Foundations will help identify potential standards and other outlets for the projects, and provide general advice on approaching these bodies and make introductions.

Business support

4. Market Readiness Level[3] provides a Market Readiness Programme that facilitates adoption of open source by mainstream decision-makers.

5. Pitch training will involve the running of training workshops offered to each group of Open Call winners, to learn how to better pitch their solution to potential end-users and investors.

6. Business modeling and coaching involves the development of business models for NGI type of projects focussing on open source, trust and privacy as core values. Projects will be offered tailored business model advice and coaching sessions.

Innovation management

7. Open source licensing — will provide guidance in the selection and management of open source licenses.

8. Market landscaping and research will provide online services based on Linknovate.com for innovation scouting and monitoring.

9. Open science advice will provide appropriate open science practices mentoring to the projects, best practices on reproducible research and open research data philosophy in general (transparency, sharing, collaboration), including the promotion of inclusion and exchange of knowledge within diverse and traditionally underrepresented groups.

10. Content creation support for marketing materials will assist NGI Surveyors to jointly produce content with the consortium in order to showcase their project results.

CONCLUSIONS

For the next 3 years, 5 open calls will offer talented researchers, innovators and activists the opportunity to perform equity-free research and development on search, indexing and discovery for the Next Generation Internet Horizon Europe initiative. The core values are open source, contributions to the wider internet community, collaboration between deep tech and industry, innovation and standardisation, open science principles as well as transversal challenges including gender and sustainability. NGI search addresses a very wide field of research and innovation and will be looking for the projects fostering strong synergies with the NGI mission on a more human centric internet with focus on privacy and trust as key concepts.

ACKNOWLEDGEMENT

The NGI Search project is funded by the European Union under the Horizon Europe Programme, grant agreement 101069364.

[2] https://www.reachout-project.eu/view/Main/

[3] https://www.ow2.org/view/MRL/

https://doi.org/10.5281/zenodo.7799190

REFERENCES

[1] R. Kachouie, S. Sedighadeli, R. Khosla, and M.-T. Chu, "Socially assistive robots in elderly care: a mixed-method systematic literature review," *International Journal of Human-Computer Interaction*, vol. 30, no. 5, pp. 369–393, 2014.

[2] X. Ma and A. Liu, "Challenges in supporting exploratory search through voice assistants," in *Proceedings of the 2nd Conference on Conversational User Interfaces*, 2020, pp. 1–3.

[3] S. Gupta, A. K. Kar, A. Baabdullah, and W. A. Al-Khowaiter, "Big data with cognitive computing: A review for the future," *International Journal of Information Management*, vol. 42, pp. 78–89, 2018.

[4] A. Chernyshov, A. Balandina, and V. Klimov, "Intelligent processing of natural language search queries using semantic mapping for user intention extracting," in *Biologically Inspired Cognitive Architectures Meeting*. Springer, 2018, pp. 56–61.

[5] V. Janeiko, R. Rezvani, N. Pourshahrokhi, S. Enshaeifar, M. Krogbæk, S. H. Christophersen, T. Elsaleh, and P. Barnaghi, "Enabling context-aware search using extracted insights from iot data streams," in *2020 Global Internet of Things Summit (GIoTS)*. IEEE, 2020, pp. 1–6.

[6] A. Gonzalez-Vidal, P. Barnaghi, and A. F. Skarmeta, "Beats: Blocks of eigenvalues algorithm for time series segmentation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 11, pp. 2051–2064, 2018.

[7] A. Al-Lahham, *Ontology-based context-aware model for event processing in an IoT environment*. University of Salford (United Kingdom), 2020.

[8] D. C. Vargas, *Strategies and Techniques for Federated Semantic Knowledge Retrieval and Integration.* IOS Press, 2019, vol. 44.

[9] J. Shen, Z. Shen, C. Xiong, C. Wang, K. Wang, and J. Han, "Taxoexpan: Self-supervised taxonomy expansion with position-enhanced graph neural network," in *Proceedings of The Web Conference 2020*, 2020, pp. 486–497.

[10] J. Huang, Y. Xie, Y. Meng, Y. Zhang, and J. Han, "Corel: Seed-guided topical taxonomy construction by concept learning and relation transferring," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 1928–1936.

[11] Z.-K. Gao, M. Small, and J. Kurths, "Complex network analysis of time series," *EPL (Europhysics Letters)*, vol. 116, no. 5, p. 50001, 2017.

[12] C. Gutiérrez and J. F. Sequeda, "Knowledge graphs," *Communications of the ACM*, vol. 64, no. 3, pp. 96–104, 2021.

[13] Ó. García, R. S. Alonso, J. Prieto, and J. M. Corchado, "Energy efficiency in public buildings through context-aware social computing," *Sensors*, vol. 17, no. 4, p. 826, 2017.

[14] I. Sittón-Candanedo, R. S. Alonso, Ó. García, L. Muñoz, and S. Rodríguez-González, "Edge computing, iot and social computing in smart energy scenarios," *Sensors*, vol. 19, no. 15, p. 3353, 2019.

[15] G. Andrienko, N. Andrienko, S. Drucker, J.-D. Fekete, D. Fisher, S. Idreos, T. Kraska, G. Li, K.-L. Ma, J. Mackinlay *et al.*, "Big data visualization and analytics: Future research challenges and emerging applications," in *BigVis 2020-3rd International Workshop on Big Data Visual Exploration and Analytics*, 2020.

[16] N. Silva, T. Blascheck, R. Jianu, N. Rodrigues, D. Weiskopf, M. Raubal, and T. Schreck, "Eye tracking support for visual analytics systems: foundations, current applications, and research challenges," in *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, 2019, pp. 1–10.

[17] P. Caillou, J. Renault, J.-D. Fekete, A.-C. Letournel, and M. Sebag, "Cartolabe: A web-based scalable visualization of large document collections," *IEEE Computer Graphics and Applications*, vol. 41, no. 2, pp. 76–88, 2020.

[18] T. Iggena, E. Bin Ilyas, M. Fischer, R. Tönjes, T. Elsaleh, R. Rezvani, N. Pourshahrokhi, S. Bischof, A. Fernbach, J. Xavier Parreira *et al.*, "Iotcrawler: Challenges and solutions for searching the internet of things," *Sensors*, vol. 21, no. 5, p. 1559, 2021.

[19] Y.-t. Zhuang, F. Wu, C. Chen, and Y.-h. Pan, "Challenges and opportunities: from big data to knowledge in ai 2.0," *Frontiers of Information Technology & Electronic Engineering*, vol. 18, no. 1, pp. 3–14, 2017.

[20] M. V. Moreno, F. Terroso-Sáenz, A. González-Vidal, M. Valdés-Vela, A. F. Skarmeta, M. A. Zamora, and V. Chang, "Applicability of big data techniques to smart cities deployments," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 2, pp. 800–809, 2016.

[21] F. Sivrikaya, N. Ben-Sassi, X.-T. Dang, O. C. Görür, and C. Kuster, "Internet of smart city objects: A distributed framework for service discovery and composition," *IEEE Access*, vol. 7, pp. 14 434–14 454, 2019.

[22] K. A. Mohamed and A. Hassan, "Evaluating federated search tools: usability and retrievability framework," *The Electronic Library*, 2015.

[23] E. M. Campos, P. F. Saura, A. González-Vidal, J. L. Hernández-Ramos, J. B. Bernabe, G. Baldini, and A. Skarmeta, "Evaluating federated learning for intrusion detection in internet of things: Review and challenges," *Computer Networks*, p. 108661, 2021.

[24] P. Ruzafa-Alcazar, P. Fernandez-Saura, E. Marmol-Campos, A. Gonzalez-Vidal, J. L. H. Ramos, J. Bernal, and A. F. Skarmeta, "Intrusion detection based on privacy-preserving federated learning for the industrial iot," *IEEE Transactions on Industrial Informatics*, 2021.

[25] Y. Zhao, P. Barnaghi, and H. Haddadi, "Multimodal federated learning on iot data."

[26] A. Gonzalez-Vidal, J. Mendoza-Bernal, S. Niu, A. F. Skarmeta, and H. Song, "A transfer learning framework for predictive energy-related scenarios in smart buildings," *IEEE Transactions on Industry Applications*, 2022.

[27] J. S. Culpepper, F. Diaz, and M. D. Smucker, "Research frontiers in information retrieval: Report from the third strategic workshop on information retrieval in lorne (swirl 2018)," in *ACM SIGIR Forum*, vol. 52, no. 1. ACM New York, NY, USA, 2018, pp. 34–90.

[28] Y. Li and N. Vasconcelos, "Repair: Removing representation bias by dataset resampling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9572–9581.

https://doi.org/10.5281/zenodo.7799190

[29] X. Jin, F. Barbieri, A. M. Davani, B. Kennedy, L. Neves, and X. Ren, "Efficiently mitigating classification bias via transfer learning," *arXiv preprint arXiv:2010.12864*, 2020.

https://doi.org/10.5281/zenodo.8075794

OPEN SCIENCE PLATFORMS AS DATA REPOSITORIES FOR AUTOMATED SUMMARIZATION

S. Frank*[1], CERN, 1211 Meyrin, Switzerland
[1]also at ISDS, Graz University of Technology, 8010 Graz, Austria
A. Wagner, CERN, 1211 Meyrin, Switzerland
C. Gütl, ISDS, Graz University of Technology, 8010 Graz, Austria

Abstract

Although open science platforms play a significant role in making research findings freely available, the search for relevant information can take a significant amount of time. This paper outlines a concept for automatic summarization that is intended to speed up and simplify this selection process by creating a summary from recent publications. The idea is to use automatic summarization to give users an overview over recently uploaded content relating to their field(s), coupled with intelligent selection of content that is relevant to the user's interest. In this, it takes into account user preferences regarding domain, timeframe, and summary length. The hope is that by utilizing notification systems and/or newsletters with these summaries, users will find staying up to date on research more time-efficient and easily accessible.

INTRODUCTION

Open science platforms and the content they provide play a considerable role in making research accessible to other scientists, as well as to the interested public. Depending on the platform, its content can range from published papers and presentations to data sets or tables. However, with the amount of material that is continually added, often of miscellaneous types and from a variety of scientific disciplines, keeping track of new research can be time-consuming and troublesome, and searching for content of interest can be challenging.

While the focus of automated summarization is frequently on singular texts, the automatic creation of multi-document summaries is a possible way of creating a more effective way of finding relevant research among a wide array of publications. In the concept described here, the user can choose to receive a summary that includes the core findings of a user-specified subgroup of articles rather than being required to actively search for potential articles of interest.

To this point, the concept places a strong emphasis on personalisation of the automated summary. The paper outlines two different ways of selecting the text sources of interest, either by predefined topic(s) or by user selection.

It is necessary to state that this concept for automated summaries is intended to help with pre-selection and gaining an overview only, it is not supposed to replace reading the papers in their entirety. Depending on their level of interest, the user will likely still have to go to read specific content to get in-depth knowledge. However, the summary allows them to choose papers more easily than by title, and it should give a sufficient overview to those who simply want to get a rough idea of recent developments in a specific field of research.

RELATED WORK

Automated text summarization has been a topic of interest for years, with entire conferences and/or conference tracks dedicated to its research; some of the most used data sets, for example, originated from Document Understanding Conferences (DUC) in the years 2001 to 2007 [1], and the Text Analysis Conferences (TAC) summarization track from 2008 to 2011, as well as 2014 [2]. In the years since, there has been much progress in the field, particularly after the first creation of a pre-trained, transformer-based architecture - BERT [3] - which has led to an increasing number of modified and extended versions thereof. However, larger data sets often prioritise news articles, such as Multi-News [4] and Newsroom [5]. On the other hand, the creation of summaries for scientific texts tends to present an expensive problem in terms of time expended, since the so-called "golden" or reference summaries are usually still manually written by humans.

Because of this, finding large training datasets which provide scientific texts as well as a summarization thereof can be challenging. Although the DUC and TAC data sets consist of papers and their summaries, they are generally limited in size, which may present problems for machine learning tasks. One recent paper by DeYoung et al. [10] introduced a data set for automatic summarization of medical studies which contains both scientific documents and summaries created from systematic literature reviews.

Furthermore, a substantial amount of research focuses on the summarization of single, shorter texts, as opposed to creating a single summary for multiple inputs. In particular, scientific texts, such as journal articles, are usually longer and thus require a different approach from inputs such as news articles. The further the input differs from that, the harder it becomes to obtain a satisfactory result, as was found when Dima et al. researched the application of NLP for technical text [6]. Even with scientific papers, which do face quite the same difficulties as technical text, problems such as redundancy may present challenges to create optimal summaries. Special considerations do not only have to be made for the used algorithms and models, either, but begins at the choice of format for the summary, for which purely textual forms may not always be the best choice.

[*] sarah.frank@cern.ch

https://doi.org/10.5281/zenodo.8075794

Open science platforms such as Zenodo[1] and INSPIRE[2] provide access to a wide range of content that can sometimes be overwhelming. Depending on the accepted file types on the platform, the available content may take many forms. From presentations over research papers to images or entire data sets, users can freely access a variety of material. In addition to this, Open Science Platforms often do not limit the provided content to only one field of science, which can present further difficulty in finding relevant research. Some platforms show keywords that can be searched with, allowing for the selection of a subset of connected articles. However, even then, it is usually necessary to access every record to be able to read their abstracts.

Due to this, and the amount of content available, it can become cumbersome to keep track of recent findings in one's topic(s) of interest and get a proper overview of new material. To get a proper idea, a user would need to first search for their topic of interest, potentially in a way that depends on the platform's search system, and then access a large number of records to reach all abstracts. For many people, this is simply inaccessible due to time constraints. However, at the same time, this wealth of resources presents a chance for research to thrive and progress rapidly by more easily building on each other's work.

Due to the usefulness of automated summarization, the creation of such summaries is a highly researched field in a variety of domains. Although much of the research has been done using news articles for data sets [8], [7], in recent years there has been an increasing number of publications that specifically focus more on scientific problem cases. While Xiao et al. showed results using a variety of data sets, some of them including scientific papers [9], another study focused on medicine in particular, where the literature review of scientific studies can take up a significant amount of time [10]. While a summary of a large number of papers cannot replace reading the papers themselves, it is possible to get an overview over the key points and overlaps, as well as distinctions between the papers, which can help with further selection of material to be read in-depth, or to get a first insight into the current state of a field that may be new to the reader. Depending on the application of the system, this can be more focused on single-document summarization or multi-document summarization, as well as abstractive or extractive summarization. Different approaches show different results depending on the application domain. In this case, focus will lie on multi-document summarization, where the objective will likely require abstractive summarization to form a coherent text from the input.

Another possibility for the automated creation of summaries is the use of knowledge graphs, where the problem can generally be split into two phases, information extraction (into the graph) and text creation (from the graph). In recent research, a phenom-based approach has for example been used to create news articles from legislative proceedings [13]. Pasunuru et al. created a method that allowed for the inclusion of graph information into the encoder/decoder-based model used to summarise multiple documents [15].

USER-FOCUSED AUTOMATED SUMMARIZATION

The goal for the concept introduced in this paper is to mitigate the problem of information overflow and attempt to make the utilisation of open science platforms more attractive to both the scientific community and the wider public. The idea is to use automatic summarization to give users an overview over recently uploaded content relating to their field(s) of interest, coupled with intelligent selection of content that is relevant to the user's interest. Ideally, the functionally would be introduced as an additional part of a notification/newsletter infrastructure. This will eliminate both the need to search for fitting content and reduce the time necessary to read through them in search of their key findings.

As such, when it comes to the utilisation of open science platforms as data repositories for an automatic summarization system, the following main requirements can be defined:

- Selection of relevant text sources
- Extraction of key information from each source
- Summarization for a specified distribution technology
- Personalisation through user specification

Consequently, the user would be able to get a firm idea of which papers are of further interest from reading the summary and, if they want to gain deeper understanding, then go on to read only those. In this context, it should be mentioned that the intention of this system is not to make any sort of decision on whether the information in the papers is considered to be scientifically sound or correct, but rather to accurately represent the content of the papers themselves.

With the specified high-level requirements the components and interactions can be roughly illustrated as in Figure 1. The two points of particular interest for the proposed system are the summarization system and the user specification, which can both be approached in a variety of ways, briefly outlined in the following sections.

Concept Proposal

To implement the pre-selection, the user would first specify their area(s) of interest, after which an algorithm would use this information to determine appropriate material on the open science platform. It may be useful to specify how far back the algorithm should look to avoid out-of-date information.

After this, the inputs - much of which usually comes in PDF form, as is the norm for published papers - go through a preparation stage, during which any text is extracted and prepared for the summarization process.

[1] `https://zenodo.org/`

[2] `https://inspirehep.net/`

https://doi.org/10.5281/zenodo.8075794

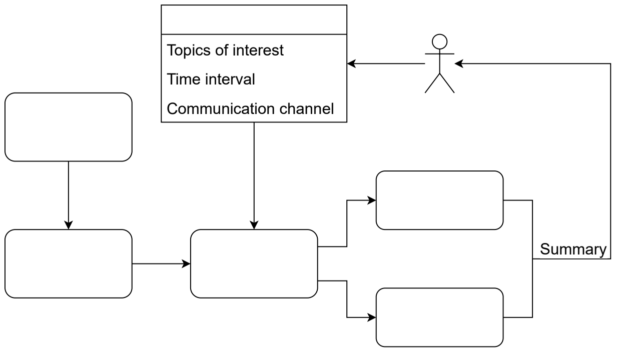


Figure 1: Diagram of an open science platform-based automatic summarization system

The extracted text then needs to be summarized, with the particular goal of stressing similarities and differences between the papers, if possible. Reading the single summary should give enough information for the user to be able to decide which papers, if any, are of further use to their research or just in general of interest. Depending on the user, the ideal summary length may be different, as well as the time between two newsletters.

For short summaries, it may be of interest to summarize the papers in one sentence, similar to the work presented by Cachola et al., in which they created TLDR versions of scientific texts [12]. An ultrashort summary of multiple papers may only need to be put together from one-sentence summaries of each paper.

The idea, however, is to find a medium between ultrashort and full-text versions of papers. In this, the summary is intended to also briefly outline where the summarised papers overlap - for example in methods used, algorithms utilised, etc. This could be where alternative modes of output, such as tables, may be of use.

Finally, this summary is communicated to the user through a previously specified way, depending on the available communication channels, such as email or app notification.

Automated Summarization

Though it is only shown as one component in Figure 1, the summarization system itself can once again be split into two main tasks: key information extraction and text creation.

Personalisation

As partially and briefly mentioned in the section before, this system is intended to provide a number of possibilities for personalisation. In the first place, the topic selection for the papers should be according to the user's selected interests. This presents the problem of how granular this selection should be; specifying singular keywords can present difficulties both due to the subjectivity of a paper's keywords and the possibility of typos, etc. Additionally, some sources may not have keywords at all, particularly in cases where other material than the papers is supposed to be part of the summarized content.

However, choosing an entire field such as Physics encompasses a wide array of subtopics, of which not all may be relevant. The solution may be platform-dependent. If there are mandatory categories when uploading content, this may later be useful to group the papers. Or, it may be necessary to categorize the papers according to topic before the actual summarization task can take place if no simpler solution for the problem can be found.

Secondly, as mentioned before, the length of the summary should be according to the user's preference. While some may prefer very short and concise summaries used only to keep vaguely up to date on evolving research directions in a field, others may wish to use it as a tool for pre-selection of reading material. Depending on the use case, the ideal summary may look different.

CONCLUSION

Open science repositories present a chance to build on each other's work in a way that can only further scientific progress. Furthermore, it allows anyone interested insight into what is currently happening in research. However, due to the wealth of information available, it can be difficult to keep track of new material of interest for the user. The proposed solution aims to automatically select content that is relevant to the user's interest, as well as to use automatic summarization to make the selected research more accessible by cutting down the time necessary to get an overview over relevant recent findings and communicating the resulting summary to the user directly.

ACKNOWLEDGMENTS

The authors thank the EC project OWS [3] for their support while working on this topic.

REFERENCES

[1] Document Understanding Conferences - Past Data, `https://www-nlpir.nist.gov/projects/duc/data.html`

[2] Text Analysis Conference (TAC) Tracks, `https://tac.nist.gov/tracks/index.html`

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding", in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, USA, June 2019. pp. 4171–4186. `doi:10.18653/v1/n19-1423`

[4] A.R. Fabbri, I. Li, T. She, S. Li, and D.R. Radev, "Multi-News: a Large-Scale Multi-Document Summarization Dataset and Abstractive Hierarchical Model", in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, July/August 2019. pp. 1074–1084. `doi:10.48550/arXiv.1906.01749`

[5] M. Grusky, M. Naaman, and Y. Artzi, "Newsroom: A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies.",

[3] `https://openwebsearch.eu/`

https://doi.org/10.5281/zenodo.8075794

in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana, USA, June 2018. pp. 708–719. doi:10.18653/v1/N18-1065

[6] A. Dima, S. Lukens, M. Hodkiewicz, T. Sexton, and M.P. Brundage, "Adapting natural language processing for technical text", in *Applied AI Letters*, volume 2, issue 3, 2021. doi:10.1002/ail2.33

[7] L. Perez-Beltrachini, and M. Lapata, "Multi-Document Summarization with Determinantal Point Process Attention", in *Journal of Artificial Intelligence Research*, 21, 2021. pp. 371-399

[8] Ghadimi, A., and Beigy, H. (2022). Hybrid multi-document summarization using pre-trained language models. Expert Systems with Applications, 192, 116292.

[9] Xiao, W., Beltagy, I., Carenini, G., and Cohan, A. (2021). Primer: Pyramid-based masked sentence pre-training for multi-document summarization. arXiv preprint arXiv:2110.08499.

[10] J. DeYoung, I. Beltagy, M. van Zuylen, B. Kuehl, and L.L. Wang. "Ms2: Multi-document summarization of medical studies", in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic, November 2021. pp. 7494–7513. doi:10.18653/v1/2021.emnlp-main.594

[11] S. Erera, M. Shmueli-Scheuer, G. Feigenblat, O.P. Nakash, O. Boni, H. Roitman, D. Cohen, B. Weiner, Y. Mass, O. Rivlin, and G. Lev, 2019. "A summarization system for scientific documents." arXiv preprint arXiv:1908.11152.

[12] I. Cachola, K. Lo, A. Cohan, D.S. Weld, "TLDR: Extreme Summarization of Scientific Documents", arXiv preprint, 2020. doi:10.48550/arXiv.2004.15011

[13] Klimashevskaia, A., Gadgil, R., Gerrity, T., Khosmood, F., Gütl, C., and Howe, P. (2021, November). Automatic News Article Generation from Legislative Proceedings: A Phenom-Based Approach. In International Conference on Statistical Language and Speech Processing (pp. 15-26). Springer, Cham.

[14] Li, W., Xiao, X., Liu, J., Wu, H., Wang, H., Du, J. (2020). Leveraging graph to improve abstractive multi-document summarization. arXiv preprint. doi:10.48550/arXiv.2005.10043

[15] Pasunuru, R., Liu, M., Bansal, M., Ravi, S., Dreyer, M. (2021, June). Efficiently summarizing text and graph encodings of multi-document clusters. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 4768-4779).

[16] Bobic, A., Le Goff, J. M., Gütl, C. (2021). Collaboration Spotting X-A Visual Network Exploration Tool. In in Proceedings of the The Eighth International Conference on Social Networks Analysis, Management and Security: SNAMS 2021.

https://doi.org/10.5281/zenodo.7799458

SEARCHING AND STRUCTURING THE TWITTER STREAM FOR CRISIS RESPONSE: A FLEXIBLE CONCEPT TO SUPPORT RESEARCH AND PRACTICE

J. H. Bongard*, J. Kersten, F. Klan, German Aerospace Center, [07745] Jena, Germany

Abstract

In the context of crisis situations, the high information value of Twitter has already been demonstrated in several studies. However, there are various issues that prevent users from deploying information from social media in practice. This paper outlines the experiences and results of the Data4Human project. In collaboration with the World Food Programme, the German Red Cross and the Humanitarian Open Street Map Team, requirements were defined that have to be fulfilled for a practical use of the data. On this basis, a modular processing system was developed which, in combination with a dashboard, enables the practitioners to evaluate the data and analysis results in a structured and interactive manner. Pre-trained and adjustable machine learning methods combined with search- and unsupervised aggregation-capabilities allow to yield different thematic views on the data. By using two use cases, Cyclone Idai and Kenneth in 2019 in Southern Africa and the current war in Ukraine, it becomes apparent, that interactive and customizable methods are of great importance, as well as an effective data representation that enables the identification of relevant content. Based on the results obtained and the user feedback received, future research guidelines are defined in order to close the evident gap between research and practice.

INTRODUCTION

Crisis informatics research from the past two decades impressively demonstrates the potential of social media data as a valuable information source for various crisis management applications [1]. Studies on social media usage and communication patterns [2], public information and warning [3], event/topic detection and tracking [4], situational awareness [5], disaster preparedness [6] and resilience [7], and decision-making [8] cover the complete disaster cycle. However, issues like data retrieval, information overload, limited representativity of data, location uncertainties, and unknown trustworthiness along with privacy and ethical concerns, still hamper the exploitation of social media data in practice.

With special emphasis on crisis and emergency response, the German Aerospace Center (DLR) project Data4Human (D4H) [9], among others, addresses the question of how social media can be utilized in practice. The goal is to enable our partners World Food Programme (WFP), German Red Cross (GRC), and the Humanitarian Open Street Map Team (HOTOSM) to gain additional local insights and information, for example regarding (sub) events as well as impacts on population and infrastructure. Twitter data analysis for cyclones Idai and Kenneth (Mozambique, 2019) as well as for the war in Ukraine immediately made clear that potential thematic questions can change rapidly over time and are often not even known in advance, as they may arise from the content of the data itself. Purposeful yet flexible search mechanisms to robustly identify and aggregate relevant information, along with sophisticated information management, can help to address the aforementioned shortcomings and therefore, play an extraordinary role for application-oriented crisis informatics research.

In this paper, the evident gap between research and emergency management practice is addressed. Based on our findings and experiences within the D4H project, we outline general guidelines that we are convinced will help to tailor future social media research for practical applications. Our proposed prototype for Twitter data acquisition, analysis and interactive information access is designed according to these findings and is iteratively shared with and evaluated by our partners during the development. The use case of bi-directional validation of information candidates from multiple sources appears to be our first step towards a contribution of Twitter information to existing emergency response workflows.

In the following section, the scientific literature is introduced. Then, different humanitarian frameworks as well as the requirements defined by the practitioners involved in D4H are presented. Thereafter, a prototype of a modular workflow and data visualization is presented and finally discussed based on the two use cases and with respect to the specified requirements.

RELATED WORK

A primary function of social media during a crisis response is to improve situational awareness [3, 8]. According to Endsley's model [10], situational awareness is defined as: "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future". Based on this perception, decisions are taken and actions are executed.

Within the process of gaining situational awareness through local information, the actual question at hand and the definition of relevance appears to be role dependent. Therefore, interactivity and the customization of social media filtering and analysis algorithms turn out to be essential [11, 12]. In contrast, the identification or classification of informative [13], crisis-related [14] or actionable [15] tweets is commonly approached with pre-trained machine

[*] jan.bongard@dlr.de

https://doi.org/10.5281/zenodo.7799458

learning models [16], often focusing on specific pre-defined event types, like landslides [17] or floods [18, 19], or on specific topics, like injured or dead people [5] and eyewitness reports [20].

With emphasis on rather generic applications, several comprehensive monitoring systems comprising a collection of analysis methods and functionalities have been proposed. Examples are EAIMS [21] for real-time detection of emergency events, related information finding and credibility analysis, and Event Tracker [22] providing a unified view of an event, integrating information from news sources, emergency response officers, social media, and volunteers. However, only few studies are available, in which Twitter analysis methods proposed by researchers have been evaluated by practitioners. Promising results obtained with a visualization framework for situational awareness [23] and a real-time system for detecting landslide reports on social media [17] underpin the need for flexible systems comprising different methods of data analysis and interactive information access.

Finally, only a few studies examine the integration of social media data into emergency management processes. They emphasize the importance of geolocation as well as search and filtering functionalities [24]. Nevertheless, interactive "dashboards have received trivial attention in terms of the usable representation of such information" [25]. Therefore, little is known about the role social media can play in the decision-making processes. It’s a causality dilemma as, on the one hand, crisis management does not know what to expect from social media data, and on the other hand, scientists do not know existing workflows and issues that come up during a mission. For these reasons, it is important to work closely with the practitioners and provide them with flexible tools that allow them to access and explore the data based on varying information needs and to evaluate the potential for their purposes.

USER REQUIREMENTS

In the course of a crisis, each practitioner is confronted with a variety of tasks and questions that may change rapidly depending on the type of crisis or its development. For example, the analysis framework of the International Federation of Red Cross and Red Crescent Societies (IFRC) focuses on three analytical pillars covering crisis impact, severity of conditions and gaps in crisis response. For this purpose, information is needed covering, for example, the potential driver of the crisis, damages & losses, physical & mental well-being, or availability of supplies & services [26]. Beyond this "big picture", local relief forces need information on “implicit and explicit requests related to emergency needs that should be fulfilled or serviced as soon as possible” [27]. In comparison, WFP focuses on aspects of food security and analyzes, for example, agro-economical conditions, as well as political and socio-economic factors to quantify their impact on food intake and health status of individuals [28]. Hence, interactivity and the customization of social media filtering and analysis algorithms are essential for improving situational awareness [11, 12].

More specifically, the D4H practitioners highlighted six requirements addressing the data, information extraction, the processing pipeline and the visualization of the results.

(1) Quick & continuous data availability: As additional data source, social media unfolds its potential when relief workers cannot be on site and therefore no first-hand information is available. In most cases, this is especially true at the beginning of a crisis or in cut-off places to which relief workers cannot gain access. If data meets this requirement, it is possible to bridge existing information gaps, for example, between the acquisition of satellite images.

(2) Location is key: In order to capture the extent and severity of affected areas, geospatial information - including uncertainty information - is crucial. Again, especially at the beginning of a crisis, such data can help to organize relief workers and mapping efforts.

(3) Globally applicable workflow: To be able to operate in numerous countries, the workflow should function independently of the respective language and should be scalable to highly different amounts of data volumes.

(4) Flexible methods: Since each organizations has its own thematic focus and each situation can change, the question at hand can vary quickly. Hence, the tools should ideally be applicable to different scenarios such like hurricanes, droughts, armed conflicts, or social domino effects resulting from a crisis. The flexibility and the social component in particular, offers the possibility to proactively identify threatening situations in advance in order to avert the crisis or lessen its effects.

(5) Data bias & trustworthiness: Only a small fraction of affected people post online. It must be made clear if certain areas and certain population groups are not well represented in the data. Moreover, information obtained from a single tweet is not trusted.

(6) Visualization: On the one hand, the visualization should capture the big picture so that context is given and a summary is provided for a quick understanding of the situation. On the other hand, detailed information such as precise coordinates, media, or urgent requests should provide a foundation for immediate actions. In addition, the platform should be user-friendly and operable without technical and methodological background in Geographic Information Systems (GIS) or Natural Language Processing (NLP).

MODULAR PROCESSING SYSTEM

The concept of the resulting modular processing system includes a framework of loosely assembled modules ranging from data acquisition to pre-processing, as well as information extraction, classification and verification. The concept allows to add modules, to adapt them individually to the respective user, or to update modules to the current state of research. In combination with data representation via a dashboard, the concept offers the possibility to involve the practitioners through feedback loops. This enables our users to gain own insights and learn what types of information is hidden in the data. According to their evaluations, further requirements and questions for the scientific work unfold. The initial prototype (Fig. 1) includes:

Data Acquisition The respective user can trigger the processing pipeline by defining an area of interest. Thus, location specific queries can be directed to Twitter's live endpoint or full archive endpoint. In turn, georeferenced tweets and its media can be retrieved continuously, contributing to fulfill requirement (1) and (2).

Translation Translators like Google Translate offer a large portfolio of more than 100 languages [29]. Semantic subtleties may be lost, but by translating into English, various modules can be added to the workflow that are primarily designed for English language, such as for named entity recognition (NER) or sentiment analysis. Therefore, modules can be used flexibly and almost worldwide, addressing requirement (3). Moreover, a translation allows consistent interpretation of the tweet content by the user.

Text pre-processing & Text Embedding A usual pre-processing is performed, including the replacement of URLs, hashtags, mentions, emojis or character repetitions, as well as removing spaces or line breaks. After this cleanup, only tweets with at least three words are forwarded and transformed into an 512-dimensional Universal Sentence Encoder [30] embedding, which is performing well for supervised and unsupervised classification tasks [31, 32].

Supervised Classification To identify crisis-related tweets, a deep neural network was trained and tested (F1-Score 0.90) based on the HumAid data set [33]. Based on this pre-trained model, a separation into relevant and non-relevant content can be made. Furthermore, the same data set was used to train a second deep neural network to classify tweets into 10 different humanitarian categories like *infrastructure & utility damage*, *injured or dead people* or *requests or urgent needs*. In order to meet requirement (4), this data set was selected in agreement with the practitioners, as it is specifically designed to interpret humanitarian aid-related information from different types of disasters [33] and comes closest to the practitioners' own framework.

Named Entity Recognition As a last processing step, entities are extracted from each tweet text. Persons, organizations, and miscellaneous tokens are obtained by using Flair [34]. In addition, GazPNE2 [35] is used for a reliable identification of place names. In comparison to other approaches, GazPNE2 achieves state-of-the-art results (F1-Score: 0.8). So far, identified entities serve as a basis for summaries as demanded in requirement (6).

Database Along with the initial tweet information retrieved from the Twitter API, its translation, stopwords, embedding, the classification results and named entities get stored in a geospatial database which can be updated on a regular basis.

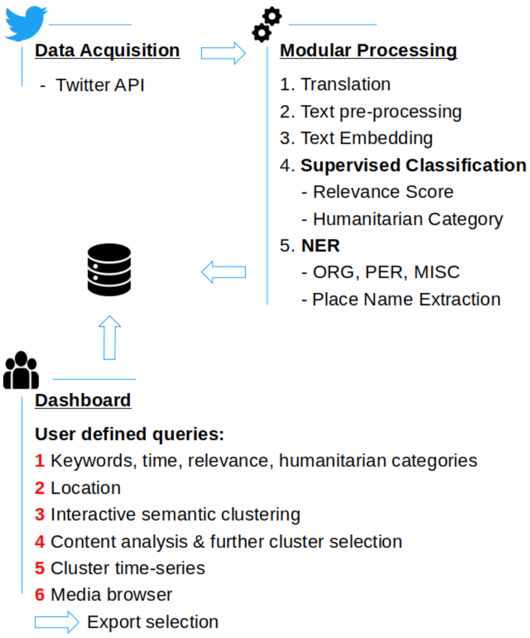


Figure 1: Modular processing system

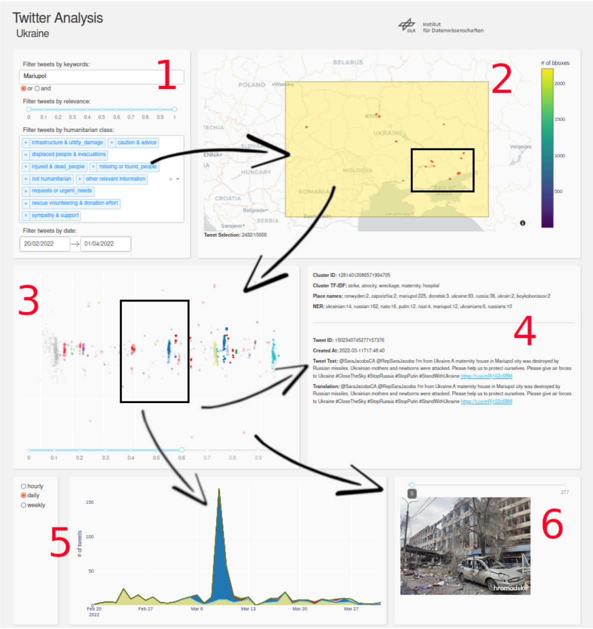


Figure 2: Dashboard design and structure

DATA VISUALIZATION

The user is given access to the database with the help of a dashboard. Its functionality supports the user in browsing through the data and directs him to potentially relevant topics. Furthermore, it offers the possibility to specify self-defined queries to find out whether and which information is available under consideration of a particular topic (requirement 4). Therefore, the dashboard offers a total of six different

https://doi.org/10.5281/zenodo.7799458

containers. Container 1 can be used to filter the database based on keywords and time, but also based on various features resulting from the processing pipeline, like relevance score and humanitarian categories. This selection is then forwarded to container 2, where the geographical distribution of the tweets is displayed as a heat map including point and bounding boxes features. Based on this geographical overview, the user can specify a region of interest.

Tweets originating from a specific area are then forwarded for detailed content analysis to container 3. Based on the sentence embedding, tweets get clustered by a Chinese Restaurant Process clustering approach [36] and then visualized by using UMAP [37] as dimension reduction technique. The user can interactively specify the cosine similarity threshold, which decides about the cluster assignments. In the background of this interactive clustering process, place names and other named entities are summarized, and the term frequency-inverse document frequency (tf-idf) is calculated for each cluster. Those cluster summaries as well as individual tweet information such as text, its translation and time stamp are shown to the user in container 4 when hovering over the data points. In combination with the summary, the clustering structures the feature space and provides a quick overview of the data. In addition, it is possible to refine the query by the *request or urgent needs* category to identify circumstances that require a rapid response (requirement 6). In addition, semantic clustering of tweets can be used to place individual information in context. Together with the visualization of the temporal development of a cluster and its media in container 5 and 6, additional indicators for the verification are provided to increase trustworthiness (requirement 5). Finally, a user can export a selection of tweets for further analysis.

PRACTICAL USE CASES

The case of Cyclone Idai and Kenneth was chosen because it is considered as one of the most severe natural disasters in South Africa, with which the practitioners are already familiar. The database obtained from Twitter's full archive endpoint contains 82,482 tweets, that were posted within Malawi, Zimbabwe or Mozambique between 15/03/2019 and 26/04/2019. Although these countries do not produce a high tweet volume and the number of relevant tweets is low (~150/day), it is possible to get an impression of the crisis over time.

For example, tweets of the *request and urgent need* category indicate the difficult supply situation of the affected population, especially of children and elderly, a cholera outbreak, as well as warnings about gender-based violence or human trafficking. The category *rescue volunteering & donation efforts* is holding not only calls for donations, but also a small discussion about Zimbabwe's political party in charge, allegations of corruption and rumors about the theft of donated relief supplies. Based on this thematic overview, more precise keyword-based queries provide more insight into topics, such as efforts undertaken by organizations to contain the cholera outbreak, including vaccinations and provision of drinking water. Nevertheless, compared to the scale of the crisis, which affected millions of people, the density of information is quite low.

Table 1: Selection of tweets reporting about hospitals

'Russian invasion: hospital attacked in Balakleya: "Russian troops fired on a hospital in Balakleya, Kharkiv region, which they occupied. 70 medics and patients have to be evacuated.'
'Mariupol. Direct strike of Russian troops at the maternity hospital. Atrocity! How much longer will the world be an accomplice ignoring terror? People and children are under the wreckage. #SafeAirliftUkraine'
'Here I am with two of the five medical doctors who remain in Zolochiv, which is located in the far north of #Ukraine, (about 10 miles from Russia), after being given the okay to come up from the hospital's bomb shelter after today's artillery attack on us.#StandWithUkraine'
'#RussianOccupants completely destroyed city center and hospital in #Izium, #Kharkiv region #Ukraine'

With the beginning of the Russian invasion in Ukraine, the question arose whether the tool could unveil information from Twitter to support a practitioner's project collecting and reviewing secondary data. In the course of the Ukraine war, the database has been regularly updated since the beginning of the invasion on 20/02/2022 until 18/05/2022 and includes 301,792 tweets.

Especially the categories *infrastructure & utility damage*, *displaced people & evacuations* or *injured dead people* offer a wide range of local information. Particularly noticeable are detailed reports about the medical infrastructure. After being guided to that topic, a "hospital" - keyword query reveals reports as exemplified in Table 1.

According to the practitioners, this information is of value because, due to the complex course of the war, only sparse and sometimes contradictory information on the state of hospitals is available. In this regard, tweets can serve as an indicator revealing valuable and timely insights. Using the dashboard, hospital related information can be identified and exported for further analysis, enabling a comparison with other secondary data sources as well as internal reports from the practitioners.

Since location is key, it is noteworthy that the majority of tweets (~95%) are geolocated by a bounding box, which in the case of the Ukraine, spans around the entire country. Therefore, 526 hospital related tweets were extracted containing an extracted place name, which could be allocated to 48 different cities belonging to 17 different oblasts. Regardless of the course of individual events, the database contains information on 25 different cities where attacks on hospitals were reported. Furthermore, 33 cities were reported to have functioning or partially operating hospitals, and nine cities were reported where hospitals are completely destroyed. As often only city names are mentioned in the tweets and hospi-

https://doi.org/10.5281/zenodo.7799458

tal names are missing, exact identification is not possible in most cases. Moreover, nearly 50% of the tweets report on the same event over a longer period of time, without introducing additional information.

As this practical demonstration shows, the initial prototype provides a first flexible toolkit that allow the practitioners to access the data, examine it according to different perspectives, and cross-check these findings with further data sources. Nevertheless, the potential of Twitter is highly dependent on the country and the virality of certain topics.

DISCUSSION

From the promising results of the current prototype, further guidelines are discussed in the following in order to direct future research to better meet the user requirements.

To flexibly adapt the modular system to a respective practitioner (requirement 4), a close cooperation is required. Several organizations have their own innovation teams and even databases containing documents and reports organized according to their respective framework. This provides a foundation to adapt especially the pre-trained supervised modules intended for overload reduction and categorization, to the practitioner's framework. However, to keep the customization effort low, it is important to evaluate the application of methods optimizing the labeling and training process, such as few shot models, one-class models, or model-assisted labeling [38] [39].

Subsequent to information discovery, information representation plays an important role (requirement 6), that is scarcely investigated so far. To further improve the usability, standards addressing data structure and its representation should be defined together with the practitioners. According to the practitioners' feedback, grouping and contextualization of selected content by interactive clustering already provides a solid overview. Moreover, it should be iterated how meaningful summaries can be extracted from a data selection, capturing most important and relevant information that are further optimized for the practitioner's internal communication.

Since location is key (requirement 2), the coarse resolution of Twitter's geolocation is noteworthy. A large part of the investigated tweets refers to the country level. Therefore, further implementations and future research on place name extraction and disambiguation are fundamentally important to correctly identify the place and thus to capture the extent and development of a crisis as precisely as possible. The Ukrainian use case shows that localization at least on city level is feasible for Twitter data.

The ability to extract and locate place names also offers the advantage of synergistically integrating multiple data sources. Text-based data, such as documents, reports, news or blogs can be used to derive further information while also enabling cross-validation of information, increasing trust (requirement 5). In addition, the synergistic use of GIS data such as satellite imagery, administrative statistics or Open Street Map can be used to identify areas and populations that may be underrepresented by web data (requirement 5) or to link critical infrastructure. Therefore, social media has to be considered as an additional data source of information, which stands out due to its fast and easy data availability.

By integrating multiple data sources, overload reduction and information management in particular become more important. Furthermore, challenges arise with regard to the thematic linking of different data sources, the spatial linking due to different resolutions and with regard to a consistent data acquisition for global applicability (requirement 3). Hence, more collaborative research on how to tailor and systematically fuse social media analysis and other data sources is required.

CONCLUSION

This paper introduces a modular processing pipeline and interactive data visualization tool which was iteratively developed together with DRK, WFP and HOTOSM enabling the practitioners to explore the potential of Twitter data to enhance situational awareness under consideration of varying information needs and requirements. The two use cases demonstrated the potential of the prototype and provided further guidance for future implementations and research.

REFERENCES

[1] C. Reuter, S. Stieglitz, and M. Imran, "Social media in conflicts and crises," *Behaviour & Information Technology*, vol. 39, no. 3, pp. 241–251, 2020.

[2] C. Reuter and M.-A. Kaufhold, "Fifteen years of social media in emergencies: A retrospective review and future directions for crisis informatics," *Journal of contingencies and crisis management*, vol. 26, no. 1, pp. 41–57, 2018.

[3] C. Zhang, C. Fan, W. Yao, X. Hu, and A. Mostafavi, "Social media for intelligent public information and warning in disasters: An interdisciplinary review," *Int. Journal of Information Management*, 2019. DOI: 10.1016/j.ijinfomgt.2019.04.004.

[4] L. Chen, H. Zhang, J. M. Jose, H. Yu, Y. Moshfeghi, and P. Triantafillou, "Topic detection and tracking on heterogeneous information," *J. Intell. Inf. Syst.*, vol. 51, no. 1, pp. 115–137, Aug. 2018, ISSN: 0925-9902.

[5] H. Omar, A. Sinha, and P. Kumar, "System for situational awareness using geospatial twitter data," in *International Conference on Innovative Computing and Communications*, Springer, 2022, pp. 731–738.

[6] S. Anson, H. Watson, K. Wadhwa, and K. Metz, "Analysing social media data for disaster preparedness: Understanding the opportunities and barriers faced by humanitarian actors," *International Journal of Disaster Risk Reduction*, vol. 21, pp. 131–139, 2017, ISSN: 2212-4209.

[7] M. Jurgens and I. Helsloot, "The effect of social media on the dynamics of (self) resilience during disasters: A literature review," *Journal of Contingencies and Crisis Management*, vol. 26, no. 1, pp. 79–88, 2018.

[8] K. Eismann, O. Posegga, and K. Fischbach, "Decision making in emergency management: The role of social media," in *ECIS*, 2018.

[9] Reliefweb, *Data4Human: Aid Organisations Get a Helping Hand From the German Aerospace Center*, https://reliefweb.int/report/world/data4human-aid-organisations-get-helping-hand-german-aerospace-center, 2020.

[10] M. R. Endsley, "Design and evaluation for situation awareness enhancement," in *Proceedings of the Human Factors Society annual meeting*, Sage Publications Sage CA: Los Angeles, CA, vol. 32, 1988, pp. 97–101.

[11] S. Stieglitz, M. Mirbabaie, J. Fromm, and S. Melzer, "The adoption of social media analytics for crisis management–challenges and opportunities," 2018.

[12] R. Fathi, D. Thom, S. Koch, T. Ertl, and F. Fiedrich, "Vost: A case study in voluntary digital participation for collaborative emergency management," *Information Processing & Management*, vol. 57, no. 4, p. 102 174, 2020, issn: 0306-4573.

[13] S. S. M. Win and T. N. Aung, "Target oriented tweets monitoring system during natural disasters," in *Proceedings of the IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)*, May 2017, pp. 143–148.

[14] M. Wiegmann, J. Kersten, F. Klan, M. Potthast, and B. Stein, "Analysis of Detection Models for Disaster-Related Tweets," in *17th ISCRAM Conference*, A. Hughes, F. McNeill, and C. Zobel, Eds., ISCRAM, May 2020.

[15] H. Zade, K. Shah, V. Rangarajan, P. Kshirsagar, M. Imran, and K. Starbird, "From situational awareness to actionability: Towards improving the utility of social media data for crisis response," *Proc. ACM Hum.-Comput. Interact.*, vol. 2, no. CSCW, Nov. 2018.

[16] A. Kruspe, J. Kersten, and F. Klan, "Review article: Detection of actionable tweets in crisis events," *Natural Hazards and Earth System Sciences*, vol. 21, no. 6, pp. 1825–1845, 2021. doi: 10.5194/nhess-21-1825-2021. https://nhess.copernicus.org/articles/21/1825/2021/

[17] F. Ofli *et al.*, "A real-time system for detecting landslide reports on social media using artificial intelligence," *arXiv preprint arXiv:2202.07475*, 2022.

[18] Q. Khan, E. Kalbus, N. Zaki, and M. M. Mohamed, "Utilization of social media in floods assessment using data mining techniques," *Plos one*, vol. 17, no. 4, e0267079, 2022.

[19] C. Bono, B. Pernici, J. L. Fernandez-Marquez, A. R. Shankar, M. O. Mülâyim, and E. Nemni, "Triggercit: Early flood alerting using twitter and geolocation–a comparison with alternative sources," *arXiv preprint arXiv:2202.12014*, 2022.

[20] K. Zahra, M. Imran, and F. O. Ostermann, "Automatic identification of eyewitness messages on twitter during disasters," *Information Processing & Management*, vol. 57, no. 1, p. 102 107, 2020, issn: 0306-4573. doi: https://doi.org/10.1016/j.ipm.2019.102107. https://www.sciencedirect.com/science/article/pii/S0306457319303590

[21] R. McCreadie, C. Macdonald, and I. Ounis, "EAIMS: emergency analysis identification and management system," in *SIGIR*, ACM, 2016, pp. 1101–1104.

[22] C. Thomas, R. McCreadie, and I. Ounis, "Event tracker: A text analytics platform for use during disasters," in *SIGIR*, ACM, 2019, pp. 1341–1344.

[23] D. Thom *et al.*, "Can twitter really save your life? a case study of visual social media analytics for situation awareness," in *2015 IEEE Pacific Visualization Symposium (PacificVis)*, 2015, pp. 183–190. doi: 10.1109/PACIFICVIS.2015.7156376.

[24] T. Onorati, P. Dıaz, and B. Carrion, "From social networks to emergency operation centers: A semantic visualization approach," *Future Generation Computer Systems*, vol. 95, pp. 829–840, 2019.

[25] A. AlAbdulaali, A. Asif, S. Khatoon, and M. Alshamari, "Designing multimodal interactive dashboard of disaster management systems," *Sensors*, vol. 22, no. 11, p. 4292, 2022.

[26] IFRC. "The ifrc revised approach for assessment and planning in emergencies." (), https://assessments.hpc.tools/attachments/24483290-97e5-4423-a03e-e4eb86e3507b/IFRC%5C%2520Approach%5C%2520for%5C%2520Assessment%5C%2520and%5C%2520Planning%5C%2520in%5C%2520Emergencies_HIAC.pdf

[27] M. Imran, C. Castillo, F. Diaz, and S. Vieweg, "Processing social media messages in mass emergency: Survey summary," in *Companion Proceedings of the The Web Conference 2018*, ser. WWW '18, Lyon, France: International World Wide Web Conferences Steering Committee, 2018, pp. 507–511.

[28] P. McKinney, "Comprehensive food security vulnerability analysis," *VAM*, pp. 57–80, 2009.

[29] G. LLC. "Translate explore the world in over 100 languages." (), https://translate.google.com/intl/en-GB/about/languages/

[30] D. Cer *et al.*, "Universal sentence encoder," *arXiv preprint arXiv:1803.11175*, 2018.

[31] M. Wiegmann, J. Kersten, F. Klan, M. Potthast, and B. Stein, "Analysis of detection models for disaster-related tweets," *Analysis of Detection Models for Disaster-Related Tweets*, pp. 872–880, 2020.

[32] L. Wang, C. Gao, J. Wei, W. Ma, R. Liu, and S. Vosoughi, "An empirical survey of unsupervised text representation methods on twitter data," *arXiv preprint arXiv:2012.03468*, 2020.

[33] F. Alam, U. Qazi, M. Imran, and F. Ofli, "Humaid: Human-annotated disaster incidents data from twitter with deep learning benchmarks," in *Proceedings of the Fifteenth International AAAI Conference on Web and Social Media, ICWSM. AAAI Press*, 2021, pp. 933–942.

[34] A. Akbik, D. Blythe, and R. Vollgraf, "Contextual string embeddings for sequence labeling," in *COLING 2018, 27th International Conference on Computational Linguistics*, 2018, pp. 1638–1649.

[35] X. Hu *et al.*, "Gazpne2: A general place name extractor for microblogs fusing gazetteers and pretrained transformer models," *IEEE Internet of Things Journal*, 2022.

[36] D. M. Blei and P. I. Frazier, "Distance dependent chinese restaurant processes.," *Journal of Machine Learning Research*, vol. 12, no. 8, 2011.

[37] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[38] A. Kruspe, J. Kersten, and F. Klan, "Detecting event-related tweets by example using few-shot models," in *16th International Conference on Information Systems for Crisis Response and Management*, 2019.

[39] J. Kersten and J. Bongard, "Gaussian processes for binary and one-class classification of crisis-related tweets," in *Accepted at: 19th International Conference on Information Systems for Crisis Response and Management*, 2022.

https://doi.org/10.5281/zenodo.8061100

TOWARDS A PRIVACY-AWARE REPRODUCIBLE MACHINE LEARNING PIPELINE FOR OPEN DATA

Igor Jakovljevic, ISDS, Graz University of Technology, Graz, Austria
also at CERN, Geneva, Switzerland
Andreas Wagner, CERN, Geneva, Switzerland
Christian Gütl, ISDS, Graz University of Technology, Graz, Austria

Abstract

Different definitions, structures, guidelines, and architectures for ML pipelines exist. However, most of these guidelines do not require a level of privacy and anonymization for sensitive data and also do not account for reproducibility. To address the mentioned problems of providing a platform for open sourcing, deployment, and evaluation of reproducible machine learning models, steps necessary to reproduce these models, and privacy-aware data storage and distribution, we propose a Privacy-Aware Machine Learning Pipeline solution. The first part of this research focuses on the analysis of existing machine learning pipeline structures, the necessary steps to create these structures to accommodate machine learning processes, and their connection to open data and privacy. Finally, the steps necessary for the creation of an ML Pipeline for an Open Search System with Kubeflow have been described and outlined, together with future ideas and possible weaknesses.

INTRODUCTION

Machine learning (ML) is an impactful topic that has been integrated into diverse areas such as medicine, decision support, navigation, and more. Building an ML application is a difficult task that one person cannot usually do. To succeed in building ML applications and to integrate ML into organizations, projects, and ideas, it is necessary to collaborate with a group of individuals or between multiple groups. For example, one group acquires raw data and labels the data from a different group to construct training data. Multiple groups need to collaborate to build an ML model. A different group can then use this model to solve business problems [1, 2].

Machine learning applications are pipelines combining different processes and different groups of experts [3]. As ML becomes more competitive, many organizations and researchers feel driven to rush in on the implementation without following procedures and not documenting methods, all for faster results. ML is facing a replication crisis, such as the ones that have been seen in psychology, medicine, and other fields over the past decades [4]. Researchers have found it difficult to reproduce many key results, raising doubts about research methods and publication protocols. Without the ability to replicate previous results, it is difficult to compare results to determine if newly developed algorithms are better than the previous ones [1, 2].

To create novel and reproducible ML algorithms, multiple steps need to be taken; in some cases, they are done by one team or by multiple teams. Example steps include data collection, data cleaning, model training, model delivery, model evaluation, and others. Different tools can be used in any of the previously mentioned steps, for example, cloud providers (Amazon S3[1], OneDrive[2], Gdrive[3], etc.) can be used for data storage, git (e.g. gitlab[4], github[5], etc.) can be used for keeping track of the code, Python libraries for model creation, docker for containerization, and more. The combination of the previously mentioned steps and tools with documentation and instructions form an ML pipeline. However, only a few solutions combine all the individual tools used for specific steps of an ML pipeline together [2,3].

Personalization through ML offers powerful tools for enhancing the user experience in diverse systems. At the same time, it raises new privacy concerns that may discourage the adoption of personalization in ML [5]. Despite privacy challenges, user data records have widely been used to study various behavioral patterns, interactions, problems, and others. Furthermore, many ML models have been trained on this type of data to perform various tasks, such as creating predictive models, recommender systems, detecting user groups, and more [6]. Usage of private data for ML has not been without abuse, the Cambridge Analytica Facebook scandal led to the revelation that an estimated 87 million Facebook user profiles had been harvested without the users' consent and knowledge. These data were used to influence the opinion of individuals, by recommending specific news articles and displaying specific content to users [7].

When distributing sensitive information, it is important to follow guidelines, which can be given by countries, public entities, or even organizations themselves. Most of these guidelines require a level of privacy and anonymization for sensitive data [8]. With a higher degree of anonymization, the less useful the information is [9]. Open Data is the term used to describe data available freely for anyone to use for analysis and research [8]. with the goal to increase accountability and transparency.

To address the mentioned problems of providing a platform for open sourcing, deployment, and evaluation of reproducible machine learning models, steps necessary to reproduce these models, and privacy-aware data storage and distribution, we propose a privacy-aware machine learning pipeline solution.

[1] https://aws.amazon.com/aws/s3
[2] https://onedrive.live.com
[3] http://drive.google.com/
[4] http://gitlab.com
[5] http://github.com

https://doi.org/10.5281/zenodo.8061100

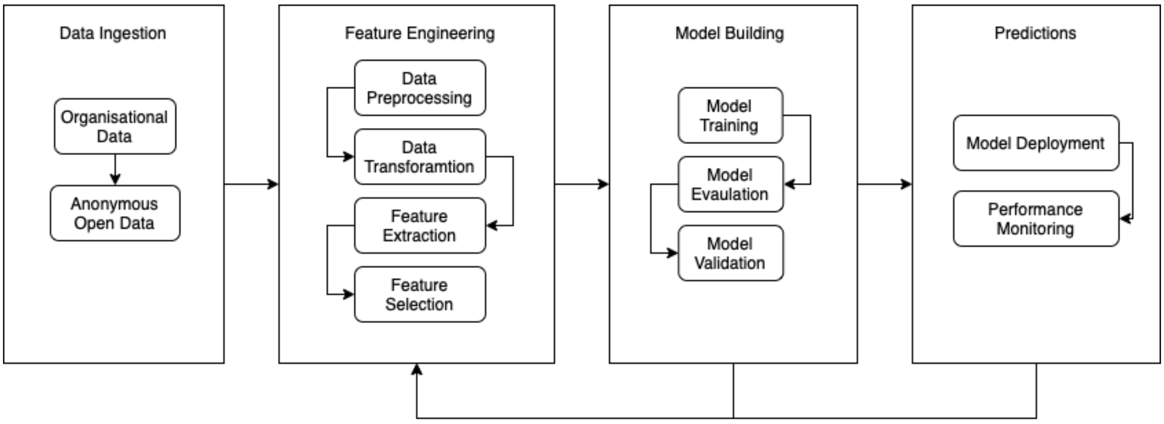


Figure 1: Open ML Pipeline Steps based on AutoML [10]

Based on the statements expressed above, the main research questions are:

- **RQ1:** How can Open Data facilitate the creation of privacy-respecting ML Pipelines?
- **RQ2:** How can Open Data-based ML pipelines be used for the implementation of ML algorithms while ensuring reproducibility using search as the application domain for ML algorithms?

The first part of this research focuses on the analysis of existing machine learning pipeline structures, the necessary steps to create these structures to accommodate machine learning processes, and their connection to open data, privacy, and data storage. In the second part of this research, we describe the necessary implementation steps for constructing such a pipeline for large interconnected organizations on the example of CERN infrastructure. Next we focus on the use case of implementing ML algoritms for search using Open Data and how can reproducibility be ensured. The paper concludes with an explanation of the drawbacks, lessons learned, and future steps to facilitate such machine learning pipelines.

BACKGROUND AND RELATED WORK

Reproducibility and Open Science

Reproducibility can be defined as the ability to replicate a model that produces the same result as the original model given the same input data [2]. It is also essential to promote open and accessible research, the use of robust experimental workflows, and to allow researchers to quickly convert ideas into practice while reducing unintentional errors [11]. Similar to the principles of reproducibility, a movement to conduct science transparently by making code, data, scientific communications, and any other research artifact publicly available and easily accessible over the long-term is called Open Science [8].

Machine Learning and Privacy

ML refers to algorithms and statistical models used by computer systems to efficiently perform specific tasks without the use of explicit instructions [12]. The need to provide personalized and evolving artificial intelligence (AI) services such as voice assistant, word suggestion, facial recognition, and smart video feeds has increased the amount of data generated. In most of these applications, machine learning models are refined by continuously feeding in new user data (as features) and their feedback (as labels). However, these data, such as type history, web access logs, and frequently visited locations, often include sensitive and private information [13]. Because of these risks, initial work on mitigating privacy risks during the machine learning process focuses on privacy challenges and risks associated with the ML models. Commonly used methods for protecting sensitive data in ML processes are Local Differential Privacy and Federated Machine Learning [13, 14].

Machine Learning Pipelines

The objective of an ML pipeline is to outline the ML model creation process with a series of steps that take a model from initial development to deployment and usage. It is an iterative process, as each step continuously improves and builds a more useful model [15]. Microsoft, Amazon, IBM, and other cloud providers have launched ML as a service that reduces costs, time, and risk of building ML infrastructure by offering premade generic ML tools. These tools recreate specific steps of ML pipelines, allowing fast implementation of ML algorithms [16].

Based on previous research, the basic steps and substeps of the ML pipelines have been identified and synthesized [2,3,6,15]. These steps shown in Figure 1 are Data Ingestion, Feature Engineering, Model Building, and Predictions.

Data Ingestion is the process of moving data from one or more sources to a destination where they can be stored and further analyzed. In addition to the collection and aggregation process, the data ingestion process can include anonymization steps. Where private attributes are removed

https://doi.org/10.5281/zenodo.8061100

or anonymized to protect user privacy. The results of this step can be also used to publish collected data (e.g. organizational data) as open data [10, 17].

Feature Engineering is the process of selecting, manipulating, and transforming raw data into features that can be used in ML. Feature engineering consists of four main steps. First, the data are pre-processed to remove outliers and noise. Then the preprocessed data is transformed to fit a data type and/or format. From the transformed data, the features are extracted using various feature extraction methods. After extracting features from the data, the last step is to select the most appropriate features given the data and the task [2, 18].

Model Building An ML model is built by learning and generalizing the training data and then applying that acquired knowledge to new data. The three main stages of model building are model training, model evaluation, and model validation [10, 18, 19].

1. Model Training - It is the initial step of the model-building process, in this step an ML algorithm/s are selected for model building. Training data is fed to the selected algorithms to help identify and learn adequate values for all attributes involved [10, 19].
2. Model Evaluation - This step focuses on finding the most suitable model representing the input data and determining how the chosen model will perform in the future. Model performance evaluation is not done with training data because it can generate overoptimistic and overfitted models. To avoid these issues evaluation methods use data from the initial dataset not used for model training [19].
3. Model Validation - It refers to the process of confirming that the model achieves its intended objective. This involves the confirmation that the model is predictive under the requirements of its intended use [18, 19].

If the models created in this phase do not perform as expected, it is possible to go back to Feature Engineering and use the knowledge gathered to improve and select better features.

Predictions After the creation, evaluation, and validation of the model, the next step includes making the implemented ML model available to end users. It requires the model to be deployed in applications and/or to an endpoint for usage. Additionally, the model's performance needs monitoring to extract information for future developments and improvements. If the results of the monitoring phase show that the created model does not perform in an expected manner, it is possible to go back to either the Model Building or Feature Engineering phase and improve the ML model. Data collected from the monitoring phase can be used as additional information to create a better ML model or to select features from the old model [10, 18, 19].

TOWARDS A PRIVACY AWARE MACHINE LEARNING PIPELINE

Taking into account the previously defined steps for ML pipelines, multiple software packages provide the necessary tools to reproduce the mentioned steps. Due to the positive impact of open source software and principles on reproducibility, it was decided to focus on open source solutions. Based on GitHub reviews, usage, frequency of updates, and community interaction the most popular open source projects are MLflow[6], Flyte[7], ML Run[8], and KubeFlow[9]. Table 1 provides a summarization of the main functionalities necessary for the creation of an ML pipeline. Based on it, MLFLow and Flyte do not provide the options to orchestrate pipelines and deploy models, which makes them not a good option for the creation of reproducible ML pipelines. Since it is not possible to define the steps necessary to create pipelines. It is also not possible to create a deployment version of an ML model created by these software packages. On the other hand, ML Run provides the necessary elements for the creation of the ML pipeline, but KubeFlow provides better documentation and an interactive dashboard. Taking this into account, organizational constraints and the level of features provided, KubeFlow was selected.

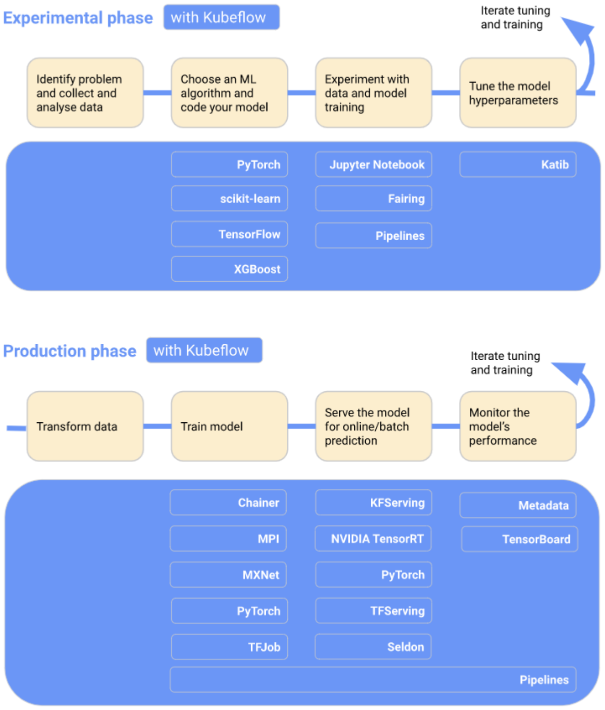

Experimental phase
with Kubeflow
Iterate tuning and training
Identify problem and collect and analyse data
Choose an ML algorithm and code your model
Experiment with data and model training
Tune the model hyperparameters
PyTorch
scikit-learn
TensorFlow
XGBoost
Jupyter Notebook
Fairing
Pipelines
Katib
Production phase
with Kubeflow
Iterate tuning and training
Transform data
Train model
Serve the model for online/batch prediction
Monitor the model's performance
Chainer
MPI
MXNet
PyTorch
TFJob
KFServing
NVIDIA TensorRT
PyTorch
TFServing
Seldon
Metadata
TensorBoard
Pipelines

Figure 2: Kubeflow components in the ML workflow [10]

Figure 2 describes the flow diagram provided by the Kubeflow documentation for the creation of ML pipelines with

[6] https://mlflow.org/

[7] https://flyte.org/

[8] https://www.mlrun.org/

[9] https://www.kubeflow.org/

[10] https://www.kubeflow.org/docs/started/architecture/

https://doi.org/10.5281/zenodo.8061100

	Kubeflow	MLFlow	Flyte	ML Run
Open Source	Yes	Yes	Yes	Yes
Language	Python	Python/R/Java	Python	Python
Documentation	Very Good	Good	Poor	Good
Tracking and Versioning	Yes	Yes	Yes	Yes
Pipeline Orchestration	Yes	No	No	Yes
Model Deployment	Yes	No	No	Yes
Scheduler	Yes	No	Yes	No
Dashboard	Yes	Yes	Yes	Limited Functionality

Table 1: Software Packages for Building ML Applications

tools that can be used in each step. The experimental and production phases are the two main phases for developing and deploying an ML system that have been identified by the Kubeflow community.

In the experimental phase, the model is developed, tested, and updated based on initial assumptions to produce favorable results. This phase is compromised of four steps. First, it is necessary to identify the ML problem. The data are then collected and analyzed for training. Afterward, an ML framework and algorithm are selected, and the initial version of the model is produced. The last step includes experimenting with the data and with model training and tuning the model's hyperparameters to ensure the most efficient processing and the most accurate results possible.

The ML model is reproduced, trained, and deployed in the production phase. To ensure that the model behaves consistently during training and prediction, the transformation of the data into a format for model training from the experimental phase must also be reproduced in the production phases. After the data transformation step, the ML model is trained and served for online predictions. In addition, the performance of the model is monitored and the prediction results are evaluated and processed for model tuning and retraining.

Based on the steps defined in Figure 1 and the previously defined Kubeflow workflow, it is noticeable that steps of the workflow can be directly mapped to Open ML Pipeline steps. The Kubeflow workflows will produce an ML model that aims to solve the identified problem, but it does not ensure reproducibility. To ensure reproducibility, it is necessary to expand the steps outlined in Kubeflow ML workflow based on Figure 1. Additional steps are needed to ensure that the data and algorithms used in the ML pipeline are available without restrictions. These steps aim to produce open data by removing sensitive information from the original dataset. Following the DataLift framework for open source organization data, it is necessary to publish data with the definition of the purpose and scope of the data, the classification of attributes and the risks of the attributes [17]. Additionally, the configuration of the ML pipeline with the used algorithms needs to be published. The data and configurations can be published to a public repository (e.g. Zenodo, Github, Gitlab, etc.). Expanding the Kubeflow ML workflow with regards to reproducibility, the number of ambiguities related to reproducing ML experiments is minimized while ensuring that reproducing results from the data becomes a straightforward process.

Reproducible Open Data and Search Use Case

The previously defined generic ML pipeline steps and selected ML pipeline framework can be used for a variety of ML applications. To demonstrate the real-world application of Kubeflow, this section focuses on determining the usability and application of Kubeflow for the creation of an ML pipeline for an Open Search System based on Open Data. Many components have to work together to implement a successful search system, from securely storing information to providing an easy-to-use interface for the user. Focusing on the open search application domain, we concentrate on the creation of an ML Pipeline that produces an ML model for relevant ranked information retrieval based on input parameters. Figure 3 represents the open search structure proposed at the 3rd International Open Search Symposium extended with ML Pipelines. The structure integrates concepts of information retrieval in large organizations, information sharing between external and internal organizational users, and the creation of user profiles that respect client privacy.

Data Ingestion

As previously mentioned this step focuses on gathering data for ML model building. With an open search system, it is necessary first to transform organizational and user data into anonymous data. Otherwise, due to privacy issues, these data could not be used for ML and model building. Additionally, users must control how their data is used and processed. This implies that the data ingestion process can only take place when all personal information is removed from the query or if the user has explicitly approved the use of the data for model training. In the case that the user has approved data usage for model training, it still needs to be anonymized.

Feature Engineering

After the data ingestion step, feature engineering starts by removing outliers and reducing the noise from the data. Based on the use case defined for the ML pipeline, the data

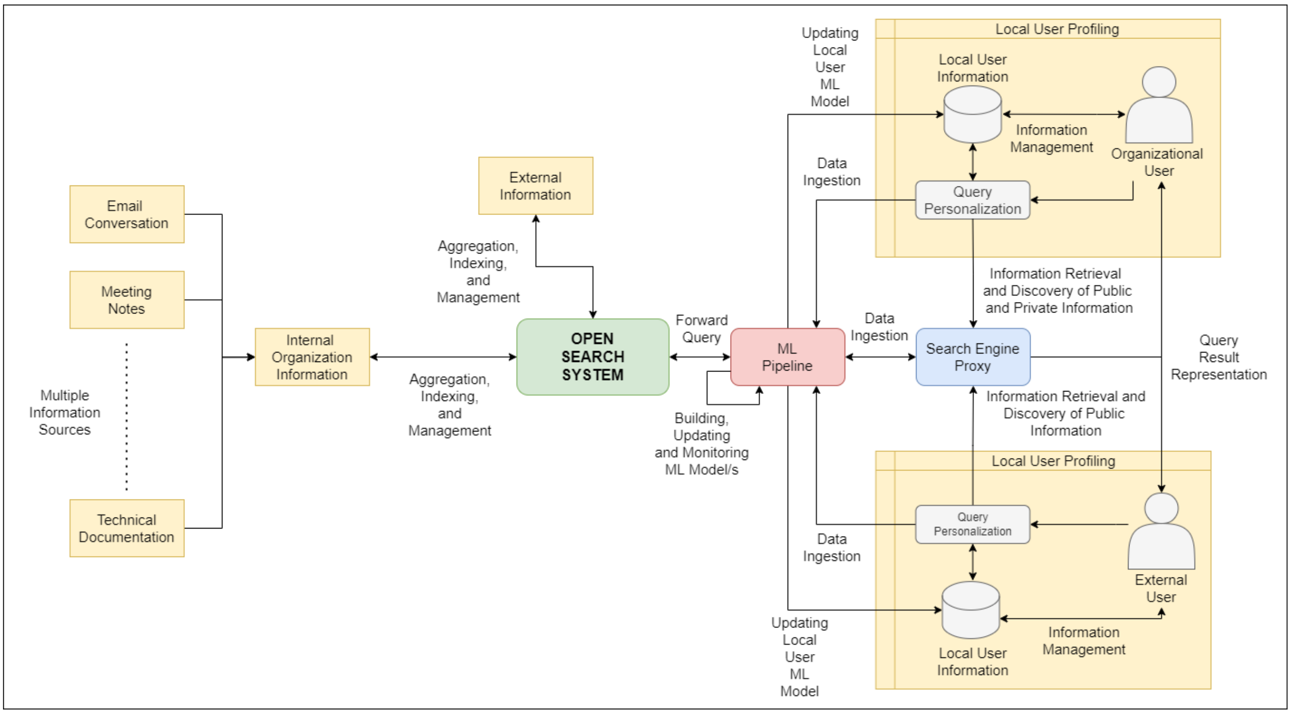

Email Conversation
Meeting Notes
Multiple Information Sources
Technical Documentation
Internal Organization Information
External Information
Aggregation, Indexing, and Management
Aggregation, Indexing, and Management
OPEN SEARCH SYSTEM
Forward Query
ML Pipeline
Building, Updating and Monitoring ML Model/s
Data Ingestion
Search Engine Proxy
Updating Local User ML Model
Data Ingestion
Local User Profiling
Local User Information
Information Management
Organizational User
Query Personalization
Information Retrieval and Discovery of Public and Private Information
Query Result Representation
Information Retrieval and Discovery of Public Information
Local User Profiling
Query Personalization
External User
Data Ingestion
Updating Local User ML Model
Local User Information
Information Management

Figure 3: Conceptual Integration Diagram of the Open Search System for Large and Highly Connected Organizations and ML Pipelines

is transformed to fit the predefined data format. From the transformed data, all possible features are extracted. Based on domain knowledge and analysis, potential features are selected for the model-building step. For an ML pipeline that works with anonymized user data, it is necessary to protect user privacy by avoiding the creation of features that could enable user identification based on anonymized data. Additionally, anonymized ingested data is published to be freely available for other services, ensuring reproducibility.

Model Building

The data created with the selected features are separated into test and validation data sets. Both data sets are compromised of input values (anonymized user queries) and output values (preferred user results). Based on data attributes (e.g., data distribution, sparseness, etc.), defined initial problem, and hardware and software constraints, an ML model is selected and trained on the test data. The built model is evaluated using statistical accuracy metrics and/or decision support accuracy metrics. In the case where the model evaluation yields positive results, it is validated by applying the model to the validation data and comparing if the model predicts the output values correctly and to what degree. To ensure reproducibility, the ML model and the metadata linked to the model must be stored and distributed as open data.

If the evaluation or validation steps do not produce satisfactory results, it is possible to repeat the step with a different model. Additionally, it is possible to go back to feature engineering and extract new features based on findings from this step

Predictions

The last step of the ML pipeline is providing a usable endpoint for the created ML model. A usable endpoint is a standalone endpoint(e.g. REST API endpoint) that can be reached by various applications and generate predictions based on the input data. The input data must be in the same format used to train and validate the model. In contrast to traditional methods where user profiles are stored on a central module, for the proposed Open Search System user profiles must be stored locally.

For local user profiling presented in Figure 3, it is necessary to share the created ML model with different users for local usage. Users would be able to filter results based on a local model without revealing personal information to the ML Pipeline and Open Search System.

FUTURE WORK AND CONCLUSION

An Open Search System needs to address many challenges to create an ML model and/or ML pipeline that is privacy aware. These include efficiently storing, generating, maintaining, and sharing anonymous user information. The drawbacks and benefits of creating a privacy-aware ML model must be analyzed in depth to find new ways to educate users about the data that are being collected and processed.

In conclusion, steps necessary for the creation of an ML Pipeline for an Open Search System with Kubeflow have been described and outlined, as well as ideas for future ideas and possible weaknesses. A generic guideline for ML pipeline creation has been synthesized, highlighting the advantages and disadvantages of each step. Four software packages for building ML applications and pipelines have been analyzed, and their usability has been assessed.

ACKNOWLEDGEMENTS

We greatly thank OpenWebSearch.EU[11] project and the members for their help and support with this publication.

REFERENCES

[1] M. Hutson, "Artificial intelligence faces reproducibility crisis," *Science (New York, N.Y.)*, vol. 359, pp. 725–726, 02 2018.

[2] P. Sugimura and F. Hartl, "Building a reproducible machine learning pipeline," 2018.

[3] Z. Cong, X. Luo, J. Pei, F. Zhu, and Y. Zhang, "Data pricing in machine learning pipelines," *Knowledge and Information Systems*, 05 2022.

[4] S. Sonnenburg, M. Braun, C. S. Ong, S. Bengio, L. Bottou, G. Holmes, Y. Lecun, K.-R. Müller, F. Pereira, C. Rasmussen, G. Rätsch, B. Schölkopf, A. Smola, P. Vincent, J. Weston, and R. Williamson, "The need for open source software in machine learning.," *Journal of Machine Learning Research*, vol. 8, pp. 2443–2466, 10 2007.

[5] E. Toch, Y. Wang, and L. F. Cranor, "Personalization and privacy: a survey of privacy risks and remedies in personalization-based systems," *User Modeling and User-Adapted Interaction*, vol. 22, no. 1-2, pp. 203–220, 2012.

[6] M. Gupta, B. Gallamoza, N. Cutrona, P. Dhakal, R. Poulain, and R. Beheshti, "An extensive data processing pipeline for mimic-iv," 2022.

[7] M. Hu, "Cambridge analytica's black box," *Big Data Society*, vol. 7, p. 205395172093809, 07 2020.

[8] S. Antony and D. Salian, *Usability of Open Data Datasets*, pp. 410–422. 10 2021.

[9] P. Ohm, "Broken promises of privacy: Responding to the surprising failure of anonymization," *UCLA Law Review*, vol. 57, 08 2009.

[10] R. E. Shawi, M. Maher, and S. Sakr, "Automated machine learning: State-of-the-art and open challenges," *CoRR*, vol. abs/1906.02287, 2019.

[11] J. Pineau, P. Vincent-Lamarre, K. Sinha, V. Larivière, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and H. Larochelle, "Improving reproducibility in machine learning research (A report from the neurips 2019 reproducibility program)," *CoRR*, vol. abs/2003.12206, 2020.

[12] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1 ed., 2007.

[13] H. Zheng, H. Hu, and Z. Han, "Preserving user privacy for machine learning: Local differential privacy or federated machine learning?," *IEEE Intelligent Systems*, vol. 35, no. 4, pp. 5–14, 2020.

[14] B. Liu, M. Ding, S. Shaham, W. Rahayu, F. Farokhi, and Z. Lin, "When machine learning meets privacy: A survey and outlook," *ACM Comput. Surv.*, vol. 54, mar 2021.

[15] S. Shaikh, H. Vishwakarma, S. Mehta, K. R. Varshney, K. N. Ramamurthy, and D. Wei, "An end-to-end machine learning pipeline that ensures fairness policies," 2017.

[16] E. De Cristofaro, "A critical overview of privacy in machine learning," *IEEE Security Privacy*, vol. 19, no. 4, pp. 19–27, 2021.

[17] G. C. W. A. Jakovljevic, I. and A. Nussbaumer, "Compiling open datasets in context of large organizations while protecting user privacy and guaranteeing plausible deniability," *In Proceedings of the 11th International Conference on Data Science, Technology and Applications (DATA 2022)*, 2022.

[18] A. Zheng and A. Casari, *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. O'Reilly Media, Inc., 1st ed., 2018.

[19] G. Luo, "Predict-ml: A tool for automating machine learning model building with big clinical data," *Health Information Science and Systems*, vol. 4, 06 2016.

[11] http://www.OpenWebSearch.EU

https://doi.org/10.5281/zenodo.8063922

THE IMPACT OF ONLINE AFFILIATE MARKETING ON WEB SEARCH

Janek Bevendorff* Matti Wiegmann* Martin Potthast Benno Stein
Bauhaus-Universität Weimar Leipzig University
<given>.<last>@uni-weimar.de

Abstract

From small independent blogs to large commercial product review portals, online affiliate marketing has become ubiquitous on the web. According to the BVDW [1], affiliate marketing was responsible for some 14 % of all German online sales in 2019. Often unnoticed by customers, inconspicuous referral links to online retailers result in a commission being paid on conversion to the referring website, making it an easy stream of passive income. With the growing affiliate market, however, we also observe a growing conflict of interest in online content creators, particularly among those who make a living from testing and reviewing products, which manifests itself in the following development: away from providing high-quality, unbiased reviews and towards maximizing conversions. Besides obvious mass affiliate spam, we noticed a growing amount of search-engine-optimized low-quality content, some of which seem hard to detect even for the large search engines today. In our research, we therefore (1) conduct an exploratory study of the landscape of Amazon affiliate links on the web to get a grasp of the issue and (2) identify website genres that search engines need to pay particular attention to in terms of content quality. To the best of our knowledge, we are the first to formulate the quality dilemma from a search engine operator's perspective.

INTRODUCTION

Most online content is free so that it can be found with search engines—a situation that puts for-profit content creators on the web under pressure to find new methods of generating revenue from their work. Many online financing models are based on advertising, premium subscriptions, dedicated crowdfunding platforms, donations, and, increasingly, affiliate marketing. In an (offline) affiliate marketing program, the affiliate partner refers customers to a seller, and if a sale occurs as a result of it, the affiliate earns a commission. In an online scenario, this usually comes in the form of specially crafted and identifiable product links; the commission depends on clicks and click-to-sale paths [2]. Affiliate marketing relies heavily on the trust relationship between customer and affiliate [3], which makes it attractive to influencers and social media marketers [4]. Many websites use affiliate links in addition to donations and advertising.

Unfortunately, the generally impersonal nature of the web makes it easy to abuse this relationship of trust, with the result that the focus shifts from producing high-quality content to maximizing conversions. To counteract this development, search engine operators publish guidelines on what they consider to be high-quality content [5] and trustworthy affiliate referrals [6]. Unsurprisingly, these guidelines, in practice, also serve as recipes for running highly search-engine-optimized affiliate campaigns. Unlike with other financing models (particularly donations, but also ads), revenue does not directly depend on the quality of the content itself but solely on conversions from referrals. Content creators are tempted to optimize against the search engine: instead of creating user-centric content, they will produce search-engine-centric content, i.e., content that serves as a mere vehicle for a page to be indexed and listed. This strategy works as long as the content conveys sufficient trustworthiness and expertise to a casual user at first glance.

Content designed to abuse search engine rankings in order to push low-quality affiliate content is undesired but not necessarily classic "spam". Search engines need to take action against this kind of search engine optimization (SEO) by either *de-indexing* or *de-ranking* pages. However, detecting such deceptive or low-quality affiliate campaigns is difficult even for established big search businesses and requires a deep understanding of the problem.

Previous work in this area focuses primarily on search-engine-optimized web spam in general [7] and its mitigation [8–11]. While affiliate spam was found to be one of the most frequent forms of web spam, it neglects a fundamental analysis of the relationship between affiliate marketing and page quality with the advent of low-quality (pseudo) review websites. Other work on affiliate marketing abuse centers around security-related aspects, such as planting malicious cookies in users' browsers [12]. Research on fake product reviews [13], review spam [14], or review quality and helpfulness [15] deals primarily with user-contributed reviews on retail websites and not with dedicated (comparative) review websites on the web and how these may be shaped by affiliate marketing as a financing model.

The paper in hand contributes in this regard as follows: (1) We extract Amazon affiliate links and several efficiently computable content metrics inspired by Google's web quality and SEO guidelines from four Common Crawls,[1] (2) we aggregate the metrics and conduct an initial exploratory study to identify classes of affiliate websites with potentially abusive design and behavior, and (3) we introduce a categorization of affiliate websites into seven sub-genres, based on the websites' usage of affiliate links and design goals. We were able to find approximate boundaries between some genres based on only a few key metrics and identified other more problematic genres which search engines should pay particular attention to and for which more and focussed research is needed.

[*] equal contribution

[1] https://commoncrawl.org/

Feature	2015	2020	2021	2022
Element Descriptions				
`<img>` FWR	–	–	–	–
`<img>` TTR	↘	↗	↗	↗
`<img>` word (avg.)	–	↗	↗	↗
`<a>` FWR	–	–	↘	–
`<a>` TTR	–	↘	↘	↘
`<a>` words (avg.)	–	↗	↗	–
`<meta>` FWR	–	–	↗	↗
`<meta>` TTR	–	–	–	–
`<meta>` words (w/o 0)	–	–	–	–
`<h1>` FWR	–	–	–	–
`<h1>` TTR	–	–	–	–
`<h1>` words (w/o 0)	–	↗	↗	↗
`<title>` length	–	–	–	–
Page Structure				
`<h1>` count	–	–	–	–
`<h2>` count	–	–	–	–
`<p>+<h[1-6]>` ratio	–	↗	↗	↗
`<img>` count	↗*	–	–	–
`<a>` count	↗*	↘	↘	↘*
Data-element count	–	–	–	–
Anchor-to-content ratio	↗	↗*	↗*	↗*
Main Content				
Content word count	↗	↗	↗	↗
Content FWR	–	↘	↘*	↘*
Content TTR	↘	↘	↘	↘
Content Flesch score	–	↘*	↘*	↘*
URL Structure				
URL path depth	–	↘	↘	–
URL path length	–	↘*	↘	↘
URL number of digits	–	↘	↘	↘
URL hyphen ratio	–	↗	↗*	↗*

Table 1: Website quality features and their relationship with affiliate link counts for affiliate web pages in the four Common Crawls. The relationships are marked uncorrelated (–), increasing (↗), or decreasing (↘) for pages with 1–35 links. (*) indicates that a correlation also holds for 35–100 links.

ANALYTICAL FRAMEWORK

To explore the characteristics of affiliate websites, we derived 28 approximate characteristics of site quality that can be operationalized in an easy-to-scale manner, inspired by Google's SEO [5] and affiliate marketing guidelines [6]. We calculated these features on over 15 million affiliate pages from four Common Crawls and aggregated them as page-level macro and domain-level micro averages.

Operationalizing Page Quality

In the following, we lay out our 5-step process of operationalizing Google's guidelines:

1. Rephrasing free-text recommendations that either encourage (*"do"* or *"use"*) or discourage (*"avoid"*) an action into 61 plain, imperative statements (e.g., *"avoid keyword stuffing anchor texts"*).

CC Name	Total		With Affiliate Links		
	Pages	Domains	Pages	Domains	Suffixes
CC-2022-05	2.9B	35.5M	3.5M	160k	153k
CC-2021-04	3.3B	35.3M	3.8M	167k	159k
CC-2020-10	2.6B	36.1M	3.2M	152k	143k
CC-2015-11	1.7B	14.9M	4.7M	165k	121k

Table 2: Page und domain (suffix) counts of the four crawls before and after affiliate link extraction. Suffixes were calculated using the Public Suffix List.

2. Removal of 24 statements that cannot be measured using surface-level text features from the page HTML source code, e.g., because they require executing JavaScript (e.g., *"avoid interstitial popups"*), rendering CSS, resolving link graphs (e.g., *"avoid broken links"*), or which cannot be reproduced from a Common Crawl (e.g., *"avoid distracting advertisements"*).

3. Removal of 10 statements too expensive to calculate at scale, such as *"avoid grammar and spelling mistakes"* or *"avoid complex navigation patterns"*.

4. Removal of 10 statements that are irrelevant, such as *"use explicit image filenames (will be distorted by most CMS)"*, or included in other statements, such as *"use less than 1,000 links on a page"*.

5. Engineering of 28 numeric or boolean features for the remaining statements.

With this procedure, we compiled a feature set that measures (1) the length and lexical diversity of `<a>` anchor and `<img>` alt texts, `<meta>` descriptions, and `<h1>` headings by extracting word and character counts, type-token ratio (TTR) and function word ratios (FWR); (2) the structuredness of a page by counting `<h1>`, `<h2>`, `<img>`, and `<a>` tags, the ratio of `<p>` and `<h[1-6]>` elements to main content words, and the existence of Open Graph or JSON linked data (JSON-LD); (3) the length, diversity, and readability of the main content, and the ratio of words in affiliate link anchors to main content words as a measure of link spam; (6) length and structure of the page URLs. We further calculated the number of affiliate and non-affiliate links. Table 1 lists all features and their behaviors on the four crawls.

Since product review pages naturally use affiliate marketing, we also computed the review-to-non-review ratio (Figure 1d) depending on the affiliate link count. For this, we naively classified a website as "product review" if its headline elements contained typical review phrases such as "`Best` *n*", "`top picks`", or "`Review`".

We also calculated Bahri et al.'s GPT-2 page quality proxy measure [16] on two of the crawls but were unable to find any correlation with affiliate links or actual page quality, which suggests that the extracted page contents, even though of vastly varying quality, were not (yet) GPT-generated.

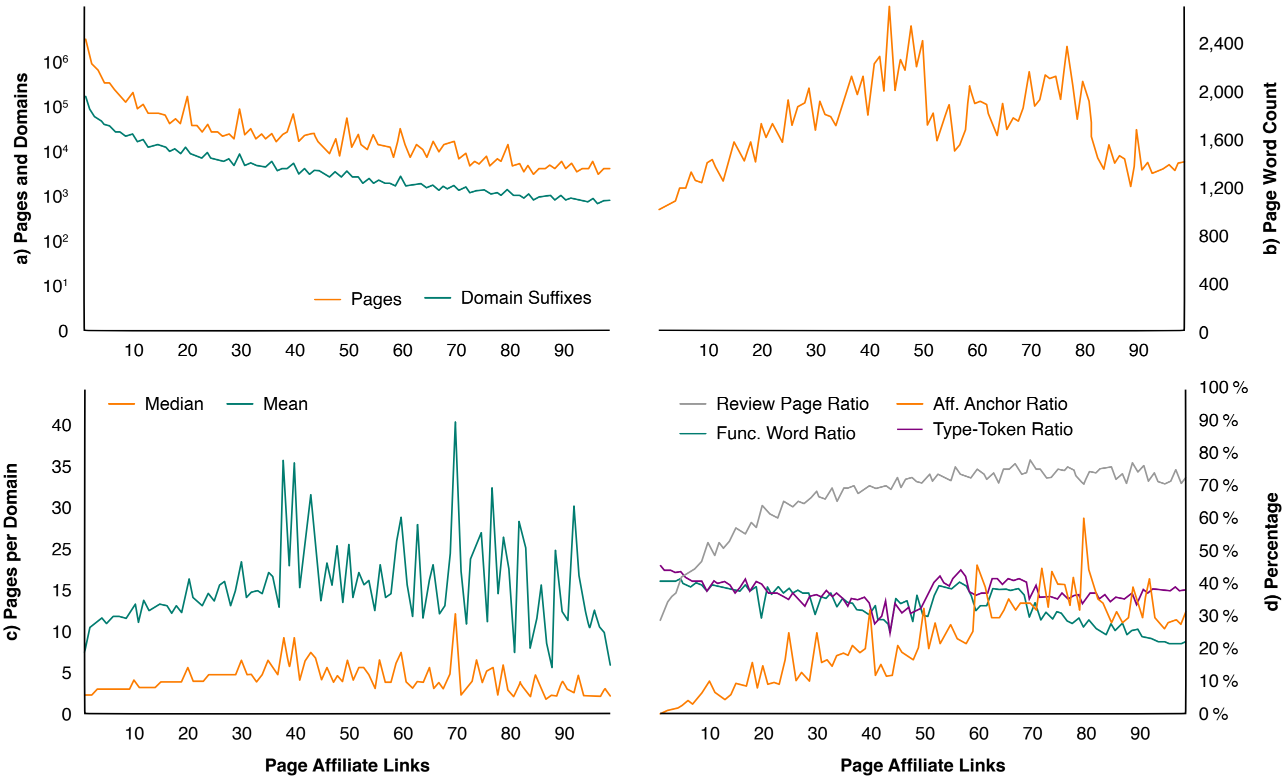


Figure 1: Page quality features averaged over the number of affiliate links on the 2020–22 Common Crawls. In relation to the number of affiliate links, we show **a)** the (log-scaled) number of individual pages and domain suffixes, **b)** the median per-page main content word count, **c)** the mean and median number of pages per domain, and **d)** the percentages of review to non-review pages, main content TTR and FWR, as well as the ratio of affiliate link anchor words to main content words.

Web Data and Extraction Pipeline

Our explorative study is based on four Common Crawls from 2022, 2021, 2020, and 2015 with 10.5 billion pages total, from which we extracted all pages containing affiliate links. For simplicity, we considered only *Amazon Associates*, as it is the largest affiliate network, and left other networks for future work. Table 2 gives an overview of the page and domain counts per crawl before and after affiliate link extraction. To get a more accurate idea of how many affiliate websites there are, we stripped the domain names of their subdomains using Mozilla's Public Suffix List (PSL).[2] Since the popular blogging platform `wordpress.com` is one of the most frequent second-level domains in the crawl but, unlike `blogspot.com`, not listed in the PSL, we added an extra rule for this domain.

In the three most recent crawls, about one in 900 pages and one in 300 domains contain Amazon affiliate links. For the 2015 crawl, these values are much higher, with one in 360 and one in 90, respectively. Overall, we detected affiliate links on 15.1 million web pages but decided to exclude the 2015 Common Crawl from further analysis due to its highly skewed distribution compared to all other crawls. This skew stems mainly from artifacts caused by obvious clusters of spam websites with many individual pages. Removing this crawl left roughly 10.4 million pages from 270,400 unique domain suffixes.

We extracted the main content of all remaining web pages using the Resiliparse library [17] and removed non-English pages and pages with fewer than 500 words. We then calculated all page-level features from the previous section on an Apache Flink cluster using the Apache Beam Python SDK. After the extraction, we also calculated the domain suffix-level mean and median as micro-average statistics for all numeric features and majority votes for boolean features. Finally, we indexed all page- and domain-level feature statistics to an Elasticsearch cluster for further analysis.

EXPLORATIVE ANALYSIS OF AFFILIATE MARKETING WEBSITES

To check our hypothesis that affiliate marketing incentivizes search-engine-centric design over page quality, we investigated page quality feature aggregates and their relationship to the number of affiliate links on a page. We found that all quality features for which we could detect a correlation do indeed indicate an overall trend towards decreasing page quality with more affiliate links on a page.

Table 1 lists the relationships of all quality features with the number of affiliate links on a page. Figure 1 shows plots of a selection of the features. We focus our analysis on pages

[2] https://publicsuffix.org/

https://doi.org/10.5281/zenodo.8063922

with 35 or fewer affiliate links. Up to this threshold, pages on the order of at least 10^5 and domains on the order of at least 10^4 remain (Figure 1a), leading to stable observations across the crawls from 2020, 2021, and 2022. As mentioned previously, we excluded the 2015 Common Crawl, as it appears to contain large amounts of spam. Even though the trends of some feature statistics are reproducible there as well, many were inconclusive and distorted by spam domains with massive amounts of individual pages.

The overall strongest text quality feature across all crawls turned out to be the type-token ratio (TTR), which measures the lexical variety of a text. High amounts of repetition, as expected from repeated keywords and phrases in largely generated or heavily search-engine-optimized content, leads to an overall lower lexical variety and therefore to a lower TTR of the main content and `<a>` link anchors. Surprisingly, the TTR of (usually invisible) `<img>` alternative texts tends to increase, which may indicate keyword stuffing (i.e., the listing of many noun keywords for SEO purposes).

Another effective indicator of a low-quality website is the use of ill-formed English and frequent keyword stuffing. We measured this with the function word ratio (FWR), the Flesch reading ease score, and the word count in `<img>` alt texts, `<meta>` descriptions, `<a>` anchors, and `<h1>` headings. We found that the FWR decreased in the main content, hinting at overall lower text quality. We further observed a reduction in the Flesch reading ease score from 60 to 50, which indicates more complex texts with longer words and sentences. The average number of words in `<img>` alt texts, `<meta>` descriptions, `<a>` anchors, and `<h1>` elements also increases (excluding empty elements), another typical indicator of keyword stuffing or synthetic text.

Additionally, we found that the total number of links on a page negatively correlates with the number of affiliate links on it. This correlation means that affiliate websites tend to use, on average, fewer non-affiliate links, which would indicate an overall simpler page structure. A manual review of the most frequent websites with different numbers of affiliate links confirmed our observations. Pages with many affiliate links frequently listed excessive amounts of product titles and specifications, resulting in a low function word count, high repetitiveness, more complex words, and fewer sentence boundaries.

We found other quality features inconclusive or potentially misleading, such as the overall consistent increase in main content length. One might expect the amount of main content text to correlate with higher page quality. However, we found that websites with a higher number of affiliate links become spammier, more repetitive, and more synthetic. This observation is further supported by the increasing affiliate anchor-to-content ratio, which means that a larger portion of the main content is within affiliate link anchor texts when there are many affiliate links.

As shown in Figure 1, there appears to be a consistent relationship between several quality features and the number of affiliate links up to about 30–40. Above this range, most metrics show a sudden increase in variance and asymmetry, influencing both the median and the arithmetic mean. This is largely explained by the exponential decrease in the number of samples with, at the same time, a sustained increase in the number of pages per domain (Figure 1c) and an overall change in genre. A manual review confirmed that websites beyond this range are almost exclusively affiliate spam or other search engine abuse. Our ad-hoc keyword classification of product review pages provides further evidence: The vast majority (> 75%) of websites are marked as "product review" past the 30–40 link range.

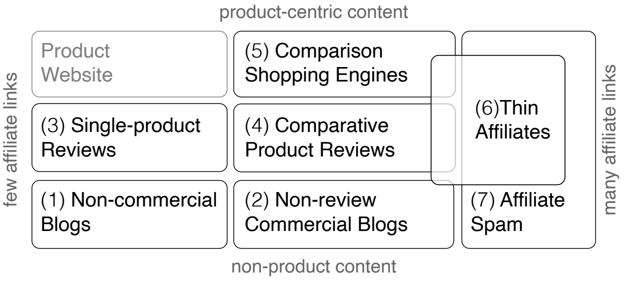


Figure 2: Schematic affiliate website genre categorization. We can characterize affiliate websites by the number of links and by the level of their integration into the main content.

Finally, because the 2015 Common Crawl appears to contain significantly larger spam clusters than the newer crawls, we discovered evidence of a noticeably improved crawling strategy. Yet even if much improved, the newer crawls are not perfect by any means, and the crawler still rather frequently runs into spam domains with many individual pages. This is evident since the mean and median number of pages per domain doubles between one and 35 affiliate links (Figure 1c). Beyond this threshold, there are many extreme domains (some were omitted as outliers). Limiting the number of pages crawled from sites with many affiliate links may potentially increase the quality of future crawls.

CLASSIFICATION OF AFFILIATE SITES

Our explorative analysis indicates that the affiliate count already explains much of the qualitative differences between websites. Affiliate websites with many affiliate links (40+) are often synthetic spam or thin affiliates as per our below definition, and websites with 10–40 links are increasingly low-quality, affiliate-centric review websites. However, qualitative analysis reveals that many low and medium link frequency websites use affiliate links more like ads: Websites place those links beside their content in marked sponsor sections, sidebars, footers, or on about pages. The content is not related to the linked products.

By observing these different usage patterns, we were able to classify most affiliate websites based on their quantitative (how many links) and their qualitative (depth of integration in the content) use of affiliate marketing (AM) into one of seven genres:

1. *Non-commercial blogs* that use AM to pay for expenses,

https://doi.org/10.5281/zenodo.8063922

2. *commercial non-review blogs and news sites* that make at least part of their regular revenue from AM,
3. *single-product reviews* that review individual products,
4. *comparative reviews* that review product ranges,
5. *price comparison* that automatically generates listings of prices and product specifications,
6. *thin affiliate sites* that pretend to publish genuine, high-quality content, though only decorating affiliate referrals with fake or low-effort editorial text that offers no added value, and
7. *affiliate spam*, automatically generated product listings without actual content.

Figure 2 depicts the genres schematically.

Since thin affiliate sites are increasingly found above 20 links per page and spam is increasingly found above 35 links per page and almost universally above 50 links per page, our hypothesis that the quality of content decreases as the number of affiliate links increases is at least partially confirmed. The fact that this trend can already be observed with few affiliate links provides additional support, even though, at this stage, we cannot completely rule out the possibility that undetected spam sites cause part of the effect. Not all genres are, however, inherently harmful. In particular, for pages with only a few affiliate links (which makes up the vast majority of pages), we observed very little spam or other unwanted content looking at the most common sites, and overall a reflection of the diversity of the web as a whole. Pages with more than 50 affiliate links are also unproblematic, as they are almost exclusively spam and can therefore be filtered effectively based on this statistic in conjunction with other traditional web spam detection methods. So the problematic genres we need to examine more closely are the ones in between. We could already confirm empirically with a basic keyword analysis in Section 3 that a fair share of these pages can be classified as comparative review pages. We were able to validate this finding by a manual qualitative review. Since the percentage of (comparative) review pages increases as a function of the number of affiliate links with near-perfect monotonicity, we see a large gray middle area in which actual review websites mix with low-effort, low-quality reviews, and thin affiliates.

In light of our findings, telling genuine reviews apart from low-quality pseudo reviews and thin affiliates in the medium range of affiliate links seems a vital objective for search engines. Our basic page- and domain-level macro statistics are insufficient for drawing a sharp boundary between the genres. However, even highly advanced retrieval systems appear to have difficulties with these genres, as is evident from the fact that we found many of the low-quality reviews and thin affiliates that we discovered during our qualitative analysis also in Google's index and, for specific keywords, more often than not, among the top results.

CONCLUSION

In this work, we scrutinize the prevalence of affiliate marketing on the web, its impact on website quality, and its implications for search engine operations. Based on several page- and domain-level text quality features, which we calculated at scale on four different Common Crawl versions, we find indicators that page quality is (on average) negatively correlated with the number of affiliate links, even in website genres that do not count as spam. Most indicative of this finding is a reduction in the main content type-token and function word ratios and an increase in the Flesch reading ease score. These observations support our hypothesis that a conflict of interest indeed exists between the financial incentives of affiliate marketing and the creation of high-quality content, although further research is needed.

We further find that at more than 20–30 affiliate links, the vast majority of affiliate pages can be classified as "product review" pages, and beyond more than 35–40 links, pages are almost exclusively spam. From this observation, we developed a broad seven-class affiliate website genre classification based on the number of affiliate links and the amount to which a page's contents are centered around the referred-to products. While the "affiliate spam" genre is easy to detect due to its excessive use of affiliate links without any actual content, the "thin affiliate" genre and low-quality examples from the "product review" range are much harder to pin down. From the perspective of a search engine operator, these two genres, therefore, require particular attention and suitable page quality measures for ensuring high-quality search results. We also find that limiting the number of pages crawled from domains that fall into the "affiliate spam" genre might improve the quality of future Common Crawl versions.

REFERENCES

[1] BVDW, "Affiliate Marketing generiert jeden siebten Euro im E-Commerce." https://www.bvdw.org/der-bvdw/news/detail/artikel/affiliate-marketing-generiert-jeden-siebten-euro-im-e-commerce/, 2019. Last accessed: Jun 18, 2022.

[2] R. Olbrich, P. M. Bormann, and M. Hundt, "Analyzing the click path of Affiliate-Marketing campaigns," *J. Advert. Res.*, vol. 59, pp. 342–356, Sept. 2019.

[3] N. Gregori, R. Daniele, and L. Altinay, "Affiliate marketing in tourism: Determinants of consumer trust," *J. Travel Res.*, vol. 53, pp. 196–210, Mar. 2014.

[4] A. Mathur, A. Narayanan, and M. Chetty, "Endorsements on social media: An empirical study of affiliate marketing disclosures on youtube and pinterest," *Proc. ACM Hum.-Comput. Interact.*, vol. 2, nov 2018.

[5] Google Developers, "Write high quality product reviews." https://developers.google.com/search/docs/advanced/ecommerce/write-high-quality-product-reviews, 2022. Last accessed: June 17, 2022.

[6] Google Developers, "Affiliate programs." https://developers.google.com/search/docs/advanced/guidelines/affiliate-programs, 2022. Last accessed: June 17, 2022.

[7] Z. Gyongyi and H. Garcia-Molina, "Spam: it's not just for inboxes anymore," *Computer*, vol. 38, pp. 28–34, Oct. 2005.

https://doi.org/10.5281/zenodo.8063922

[8] C. Castillo, D. Donato, A. Gionis, V. Murdock, and F. Silvestri, "Know your neighbors: web spam detection using the web topology," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, (New York, NY, USA), pp. 423–430, Association for Computing Machinery, July 2007.

[9] A. Chandra, M. Suaib, and R. Beg, "Google search algorithm updates against web spam," *Department of Computer Science & Engineering, Integral University*, vol. 3, no. 1, pp. 1–10, 2015.

[10] F. Asdaghi and A. Soleimani, "An effective feature selection method for web spam detection," *Knowledge-Based Systems*, vol. 166, pp. 198–206, Feb. 2019.

[11] J. Liu, Y. Su, S. Lv, and C. Huang, "Detecting web spam based on novel features from web page source code," *Security and Communication Networks*, vol. 2020, Dec. 2020.

[12] N. Chachra, S. Savage, and G. M. Voelker, "Affiliate crookies: Characterizing affiliate marketing abuse," in *Proceedings of the 2015 Internet Measurement Conference*, IMC '15, (New York, NY, USA), pp. 41–47, Association for Computing Machinery, Oct. 2015.

[13] R. Mohawesh, S. Xu, S. N. Tran, R. Ollington, M. Springer, Y. Jararweh, and S. Maqsood, "Fake reviews detection: A survey," *IEEE Access*, vol. 9, pp. 65771–65802, 2021.

[14] A. Heydari, M. A. Tavakoli, N. Salim, and Z. Heydari, "Detection of review spam: A survey," *Expert Syst. Appl.*, vol. 42, pp. 3634–3642, May 2015.

[15] G. Ocampo Diaz and V. Ng, "Modeling and prediction of online product review helpfulness: A survey," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (Melbourne, Australia), pp. 698–708, Association for Computational Linguistics, July 2018.

[16] D. Bahri, Y. Tay, C. Zheng, C. Brunk, D. Metzler, and A. Tomkins, "Generative models are unsupervised predictors of page quality: A Colossal-Scale study," in *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, WSDM '21, (New York, NY, USA), pp. 301–309, Association for Computing Machinery, Mar. 2021.

[17] J. Bevendorff, M. Potthast, and B. Stein, "FastWARC: Optimizing Large-Scale Web Archive Analytics," in *3nd International Symposium on Open Search Technology (OSSYM 2021)* (A. Wagner, C. Guetl, M. Granitzer, and S. Voigt, eds.), International Open Search Symposium, Oct. 2021.

https://doi.org/10.5281/zenodo.8064587

ETHICS IN SEARCH ENGINES – THE DEVELOPMENT OF THE ETHICS MILL OF INTERNET SEARCH BASED ON A SYSTEMATIC LITERATURE ANALYSIS

Benedikt Hoffmann, Technische Hochschule Ingolstadt, 85049 Ingolstadt, Germany
Alexander Decker, Technische Hochschule Ingolstadt, 85049 Ingolstadt, Germany
Christine Plote, Open Search Foundation, 82319 Starnberg, Germany

Abstract

Ethical discussions have now also gained importance in the digitalised world. The discourse often revolves around areas such as data protection, artificial intelligence or digital ethics in general. However, there is only few research on ethics in internet search so far, although search engines play a central role in access to information and knowledge today. 98% of all access to the internet is via search engines. Ethical problems in internet search are hardly discussed, and if they are, the dialogue is only sporadic or not conducted scientifically. What role ethics play in internet search is still little known.

The aim of this research project was to close this research gap to some extent. A systematic literature review was conducted to find out which ethical fields of action and norms in other areas of the digital world can be applied to internet search. The result is the Ethics Mill of Internet Search. The model summarises the identified fields of action, their interdependencies and meanings in the context of internet search.

ETHICS IN SEARCH – A NEGLECTED AREA IN THE SCIENTIFIC DISCUSSION

Search engines function as a kind of navigator. They show which data sets and information in the World Wide Web could match individual search queries. In a mediatised world full of complex digital structures and dependencies and an enormously confusing variety of data on the internet, they thus take on one of the most important roles for the dissemination of knowledge and information in today's societies [8, p. 19]. Search engine providers have been repeatedly criticised for years [1, p. 207]. As commercially active companies, they are accused of placing economic goals above the needs of users and of violating personal rights through the non-transparent collection of user data, among other purposes for personalised advertisements (micro targeting). It is also criticised that search results are not only individually adapted and changed for users, but also for commercial goals [8, pp. 20-21]. How the relevance of search results is determined is difficult to understand for users as well as for experts [1, pp. 207-208].

But if internet search is not exclusively about users, what other aspects are significant and what role do ethics play in internet search?

Internet search, like all technological innovations of mediatisation and digitalisation, is subject to many social processes and dependencies, including political, economic, technical or social developments. Internet search is developing rapidly. At the same time, its interactions and effects on people and society are becoming increasingly difficult to calculate. Opaque political and social consequences can be the result [6]. A holistic ethical reflection is needed that takes into account all interdependencies and framework conditions as far as possible [4, p. 8].

Ethics in digital contexts have gained attention in recent years, especially with topics such as data protection and privacy, autonomous driving or, at the latest since the 2016 US elections, with the heated fake news debates [2, p. xiii]. However, the topic of internet search has been little explored in this context.

In order to close this research gap to some extent, this paper examines, on the basis of a systematic literature analysis, which fields of action and norms from other areas of digital ethics are transferable to internet search.

METHODOLOGY

The research approach was based on the scientific procedure shown in Fig. 1.

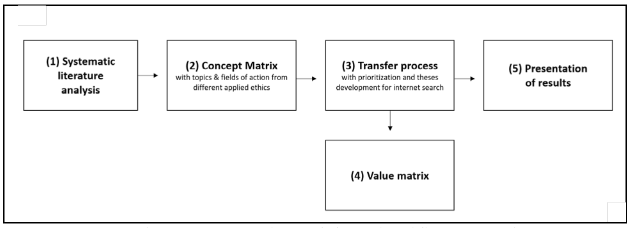

(1) Systematic literature analysis
(2) Concept Matrix
with topics & fields of action from different applied ethics
(3) Transfer process
with prioritization and theses development for internet search
(5) Presentation of results
(4) Value matrix

Figure 1: Overview of the scientific approach

In order to gain insights for the research area "Ethics in Internet Search", a **systematic literature review (1)** was first conducted. This is mainly based on secondary literature, i.e. reports on primary or original science. These are empirical, theoretical, critical/analytical and methodological works from the subject areas of 'digital ethics' and 'digital search'. Possible relevant scientific publications were also integrated. The core of the research work consists of sifting through similar area ethics that are closely related to digitalisation and do not necessarily have a direct connection to digital search. The aim is to identify important ethical fields of action in digitalisation and to determine which questions they can raise in the context of internet search.

A concept-oriented literature analysis is carried out as a basis for the relevance argumentation of different fields of action of Internet search. This means that common concepts, subject areas and similarities, but also contrasts

https://doi.org/10.5281/zenodo.8064587

within the literature are searched for in order to define suitable supercategories. The process is carried out with the help of a **concept matrix (2)** based on the guidance of Webster and Watson [10]. A total of 92 different sources were consulted (see Appendix A for an overview; the details of the individual sources have not been listed etxra in this paper as this would lead too far; however, the sources can be requested from the authors if interested).

Instead of looking specifically for concepts, it makes more sense in this context to draw on the possible ethical fields of action and topics. In the concept matrix, the relevance of the individual topics in a publication was rated from a scale of 0-4. The rating reveals how intensively a work has dealt with the specific topic area:

(4) = stood in the centre
(3) = represented an important part
(2) = was a part
(1) = was briefly mentioned
(0) = did not occur

During the concept-oriented implementation of the literature analysis two question arise: Which aspects of the fields of action/topic areas are relevant for the Internet search? And on which argumentative basis can this be justified? For this purpose, the relevance for the internet search is determined according to several criteria:

1. **Universal argumentation**: an ethical topic can be universally argued for all area ethics equally/across the board and thus also applies to Internet search.
2. **Direct reference**: an ethical problem is already linked to Internet search in the literature of other area ethics.
3. **Indirect reference**: in the general literature on Internet search, topics are considered relevant which have also been prioritised in the literature of other area ethics, but without direct reference to Internet search.
4. **Comparison**: ethical problems arise from the contexts of other ethics, which are partly reflected in Internet Search in a similar or identical way.

If one of the four points mentioned applied, the importance for the internet search was additionally weighted according to how often the fields of action are addressed in the existing literature (quantity). Another criterion was how strongly the individual works focus on the respective field of action (quality) or to what extent a high level of importance can be implied from a correlation of different works of literature.

This **transfer process (3)** is to be understood as thesis development. Corresponding findings are therefore never described as facts, but as considerations or theses, or are addressed as open questions. The purpose of this procedure is to point out possible critical gaps in knowledge. On the one hand, the work provides a good overview of the topic and, on the other hand, highlights the discrepancy between what has already been addressed in the ethical context of internet search and what may still need to be reflected on ethically in the future.

Since ethics is a reflection of morals and values, it is important in the ethical consideration to also identify concrete values. In a further process, it is therefore examined and presented from which central values the fields of action are/can be derived. This was done by creating a **value matrix (4)**. For this purpose, values and 'value demonstrators' (principles, rules or examples that always belong to a certain value and illustrate its meaning in a specific context, such as fields of action in this context) were collected in the literature analysis, which were recorded in connection with the individual topic areas. A complete overview of the values matrix and the related questions can be found in Appendix B.

All of this leads to the development of the Ethics Mill of Internet Search as part of the **presentation of results (5)**.

THE ETHICS MILL OF INTERNET SEARCH

As described under Methodology, different ethical fields of action were identified with varying degrees of importance for internet search, both universal and specific. In order to present their interdependencies and meanings in the context of internet search conclusively and clearly, the most important points were summarised in the concept of the *Ethics Mill of Internet Search*. Fig. 2 provides an overview.

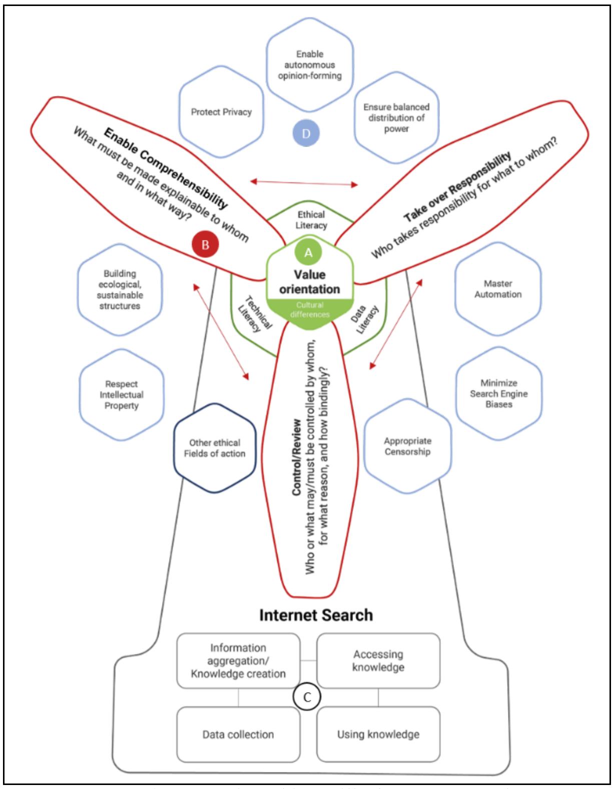

Enable autonomous opinion-forming
Protect Privacy
Ensure balanced distribution of power
D
Enable Comprehensibility
What must be made explainable to whom and in what way?
B
Take over Responsibility
Who takes responsibility for what to whom?
Ethical Literacy
A
Value orientation
Cultural differences
Technical Literacy
Data Literacy
Building ecological, sustainable structures
Respect Intellectual Property
Master Automation
Minimize Search Engine Biases
Other ethical Fields of action
Appropriate Censorship
Control/Review
Who or what may/must be controlled by whom, for what reason, and how bindingly?
Internet Search
Information aggregation/ Knowledge creation
Accessing knowledge
C
Data collection
Using knowledge

Figure 2: The Ethics Mill of Internet Search

Within the framework of the methodological procedure, it was possible to work out that acting ethically always means orienting oneself to certain values negotiated within a society (e.g. human rights). Accordingly, this value orientation (indicated by A in Fig. 2) forms the core of the

https://doi.org/10.5281/zenodo.8064587

ethics mill and must also be the focus for the internet search.

One challenge is that values differ greatly depending on culture and society (**cultural differences**). In an international context, dialogue must therefore take place again and again in order to reach agreement on common basic values despite different points of view. Internationally operating search engines, for example, have to consider to what extent adaptations of internet search to different legal situations are an interpretation of common values.

As a further aspect of value orientation, competences must be disseminated in a society that enable a public ethical dialogue. This means that **ethical literacy**, **technical literacy**, but also **specific data literacy** in the digital context are essential. It is not a question of all people having to acquire all competences equally. Rather, it is about answering the question: How detailed must different actors be trained in the three areas of competence so that they are able to make ethical decisions independently? These fields of action are thus the basis for an ethical orientation of internet search.

In order to ensure this ethical value orientation while taking cultural differences into account, the analysis further revealed that there are three subject areas that must always be considered and in every context of internet search (**see Rotor Blades of the Mill - B**). The first is that processes and systems of internet search **must be made comprehensible** so that they can be understood and ethically reflected upon. Standing up for an ethical value orientation also means **taking responsibility**. In order for processes to be ethically accompanied, it must therefore always be clearly defined who is responsible for what and to whom. And in a third step, it is also necessary to consider how certain processes, actors or even norms must or can be **reviewed and controlled**. This arises from the assumption that people do not always act according to commonly defined values, and ethical norms also need to be reviewed again and again. In this context, it is above all important not only to ask the question "whether, at all", but if so, "to what level of commitment". The level of commitment ranges from an open discursive review in the form of ethical reflection to control in the form of laws with sanctions and must be ethically justified in individual cases.

The issues of comprehensibility, responsibility and control/review generated from this have also recurred due to their generality in all fields of action of ethics in Internet search and have revealed specific challenges to these issues in the individual contexts.

The **millhouse (C)** represents the internet search itself. If one first follows the well-known presentation and argumentation of Meyer [5, p. 6], data are basically observable or measurable symbols (sensor data, metadata, a number, a word, etc.). Only when data are structured or interpreted in the form of symbols, information emerges (for example, a number becomes a date of birth and a sequence of letters becomes a name). Knowledge, on the other hand, only emerges when data is subjected to statistical analysis in order to extract causal models, or circumstantial evidence is developed from it that can support differentiated decision-making.

Since the core of Internet search is the collection and dissemination of information and knowledge, the connections of digital knowledge creation just described can also be seen as fundamental for Internet search. Internet search can therefore be divided into four ethical areas of investigation [adapted from 7, p. 49]:

1. Ethics in data collection – crawling processes, collection of user data and other data collection activities by search engine providers, etc.
2. Ethics in information aggregation and knowledge creation – building of relevance of search results, further processing of user data, etc.
3. Ethics for accessing knowledge – way of presenting results, availability of search engine offers, etc.
4. Ethics of using knowledge – misuse of knowledge by search engine providers, etc.

These four areas of investigation form the processes within the mill (of internet search) that must be reflected upon ethically. Each ethical field of action in internet search must therefore always be derived from at least one of the four areas of investigation. However, each field of action must be examined for relevant ethical aspects not only in one, but in every process area of internet search (data collection, data transfer, access to knowledge, application of knowledge).

Since the mill maps the ethical contexts of internet search, it only runs with "search engine-specific" wind in the form of various ethical topic areas and **fields of action (D)**. These resulted – as described above – from the creation of the concept matrix (2) and the transfer process (3). In the illustration of the Ethics Mill of Internet Search in Fig. 2, the fields of action are described with guiding principles that could be taken from the Value Demonstrators of the Value Matrix (see also Appendix B). They provide a possible target orientation according to which various processes of Internet search should be aligned or reviewed:

- protect privacy,
- enable autonomous opinion-forming,
- ensure balanced distribution of power,
- master automation,
- respect intellectual property,
- minimise search engine biases,
- appropriate censorship,
- building ecologically, sustainable structures.

For this to work, the questions of responsibility, comprehensibility and control (**see rotor blades B**) must be answered in all fields of action. Therefore, the rotor blades fly over the different fields of action again and again in order to make applied ethics in internet search feasible. These are interdependent. Accordingly, the ethics mill can only run with all three rotor blades.

https://doi.org/10.5281/zenodo.8064587

The field "Other ethical fields of action" points to the fact that there is no absolute completeness and correctness in ethics. In this respect, one could, for example, derive further basic principles from the contents of the core of the ethics mill:

- For value orientation: Worry about people
- For comprehensibility: Provide competencies
- For cultural differences: Enable international dialogue on values

Important questions were derived for the individual fields of action in the course of the analysis. Future scientific work or organisations can address these questions in their individual contexts. The list of questions can be found in Appendix C.

CONCLUSION

The aim of this research work was to provide an overview of ethical fields of action in internet search and the associated ethical questions. Answering these questions was not part of the analysis. Also, no ethical recommendations for action were developed and no specific requirements for the topic areas were determined. Rather, the aim was to offer approaches for further work that would deal ethically with the individual aspects of internet search.

The results of this work can be used in two ways.

(1) Scientific continuation:

The catalogue of questions offers many starting points for future scientific research on the ethics of internet search. It would be conceivable to develop answers and possible requirements for an ethical internet search. The fields of action identified could also be further developed. In cooperation with developers and search engine experts, processes (such as crawling, indexing, relevance of search results) could be reflected ethically in detail. Another approach would be to focus on individual process areas of internet search (data collection, data processing, access to knowledge, application of knowledge) or individual fields of action (e.g. according to the guiding principles) as an area of investigation.

(2) Ethical process guidance in internet search organisations:

The results can also be used for organisations in the internet search in order to accompany processes ethically in a holistic way. However, as an overview work that is not specifically designed for process accompaniment of organisations, it makes sense for this to additionally consult ethical guidelines that are specialised in ethical process accompaniment in organisations. These guidelines or frameworks are very universal and not specific to internet search. They thrive on the fact that contextual findings, such as the results of this analysis, can also be used to include as many important ethical aspects as possible in a specific context (here internet search). For example, the ethical data assistant (DEDA) [9] or the IEEE 7000 -2021 [3] could be used for this purpose. Especially the IEEE 7000-2021 standard, newly introduced on 15.09.2021 in the form of a framework, can support organisations or companies to accompany the development of systems from the beginning with ethical considerations and to anchor values in the system. It establishes a set of processes that organisations can incorporate at all stages of concept exploration and development. This standard supports management and engineering in communicating transparently with selected stakeholders to identify and prioritise ethical value

Independently of the approaches described, an ethics of internet search is also heavily dependent on the knowledge of commercial companies. In the future, ethical reflection on internet search will therefore also require cooperation with search engine providers. This can be achieved either top down through control and regulation or, in the authors' view, better through conviction and initiative on the part of the corporations themselves (by assuming independent responsibility).

REFERENCES

[1] Brückerhoff, B. (2018): Orientierung durch Suchmaschinen. Lernende Maschinen erfordern denkende Nutzer - Ein Überblick. 1. edition. Köln: Herbert von Halem Verlag.

[2] Ess, C. (2020): Digital media ethics. Digital Media and Society Book. 3. edition. Cambridge, Medford: Polity.

[3] IEEE 2021 (2021): IEEE 7000-2021 - IEEE Standard Model Process for Addressing Ethical Concerncs during System Design. Edited by the Institute of Electrical and Electronics Engineers. Electronically published via: https://standards.ieee.org/standard/7000-2021.html. Last access on: June, 19th, 2022.

[4] Kalina, A. / Krotz, F. / Rath, M. / Roth-Ebner, C. (2018): Mediatisierte Gesellschaften. Medienkommunikation und Sozialwelten im Wandel. 1. edition. Baden-Baden: Nomos Verlag.

[5] Meyer, B. (2007): The effects of computer-elicited structural and group know-ledge on complex problem solving performance. An application of two computer-based tools for knowledge elicitation. Dissertation. Humboldt-Universität Berlin, Berlin. Mathematisch-Naturwissen-schaftliche Fakultät II. Electronically published via: https://www.academia.edu/2604433/The_effects_of_computer-elicited_structu-ral_and_group_knowledge_on_complex_problem_solving_performance. Last access on: June, 19th, 2022.

[6] Rath, M. / Krotz, F. / Karmasin, M. (2019): Maschinenethik. Normative Grenzen autonomer Systeme. 1. edition. Wiesbaden: Springer Fachmedien.

[7] Spiekermann, S. (2015): Ethical IT innovation. A value-based system design approach. 1. edition. Boca Raton, London, New York: CRC Press Taylor & Francis Group.

[8] Unkel, J. (2019): Informationsselektion mit Suchmaschinen. 1. edition. Baden-Baden: Nomos Verlag.

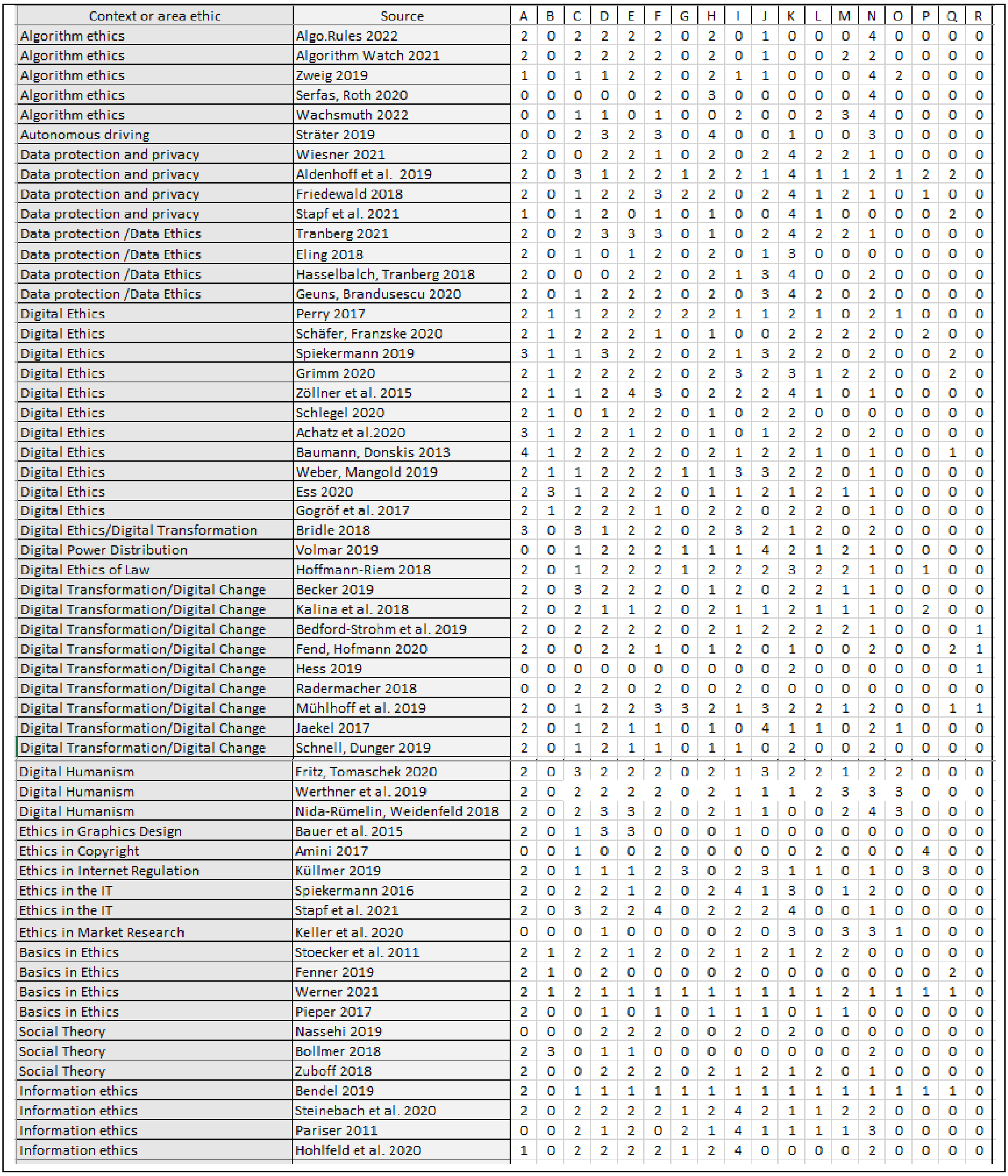
[9] Utrecht Data School (2022): Der Ethische Daten-Assistent. University of Utrecht. Utrecht. Electronically published via: https://dataschool.nl/de/deda/. Last access on: June, 19th, 2022.

[10] Webster, J.T. / Watson, R. (2002): Analyzing the Past to Prepare for the Future: Writing a Literature Review. In: MIS Quarterly (26), p. xiii–xxiii. Electronically published via: http://www.jstor.org/stable/4132319. Last access on: June, 19th, 2022.

APPENDIX A

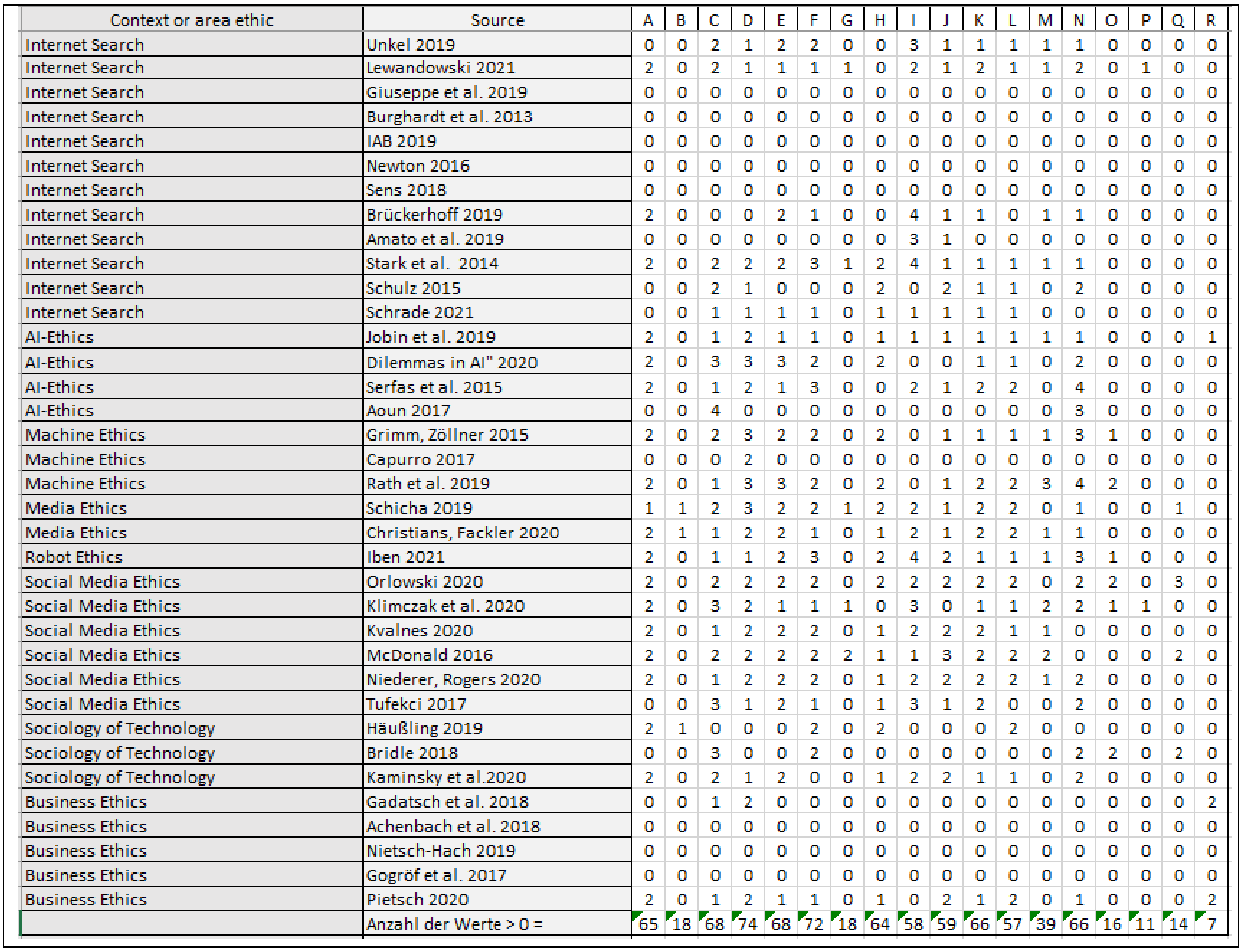
Concept matrix as result from the systematic literature analysis incl. all identified sources

The legend below shows which fields of action can be derived from the literature analysis and how intensively the individual publications have dealt with a field of action (scale from 0 to 4).



Context or area ethic	Source	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
Algorithm ethics	Algo.Rules 2022	2	0	2	2	2	2	0	2	0	1	0	0	0	4	0	0	0	0
Algorithm ethics	Algorithm Watch 2021	2	0	2	2	2	2	0	2	0	1	0	0	2	2	0	0	0	0
Algorithm ethics	Zweig 2019	1	0	1	1	2	2	0	2	1	1	0	0	0	4	2	0	0	0
Algorithm ethics	Serfas, Roth 2020	0	0	0	0	0	2	0	3	0	0	0	0	0	4	0	0	0	0
Algorithm ethics	Wachsmuth 2022	0	0	1	1	0	1	0	0	2	0	0	2	3	4	0	0	0	0
Autonomous driving	Sträter 2019	0	0	2	3	2	3	0	4	0	0	1	0	0	3	0	0	0	0
Data protection and privacy	Wiesner 2021	2	0	0	2	2	1	0	2	0	2	4	2	2	1	0	0	0	0
Data protection and privacy	Aldenhoff et al. 2019	2	0	3	1	2	2	1	2	2	1	4	1	1	2	1	2	2	0
Data protection and privacy	Friedewald 2018	2	0	1	2	2	3	2	2	0	2	4	1	2	1	0	1	0	0
Data protection and privacy	Stapf et al. 2021	1	0	1	2	0	1	0	1	0	0	4	1	0	0	0	0	2	0
Data protection /Data Ethics	Tranberg 2021	2	0	2	3	3	3	0	1	0	2	4	2	2	1	0	0	0	0
Data protection /Data Ethics	Eling 2018	2	0	1	0	1	2	0	2	0	1	3	0	0	0	0	0	0	0
Data protection /Data Ethics	Hasselbalch, Tranberg 2018	2	0	0	0	2	2	0	2	1	3	4	0	0	2	0	0	0	0
Data protection /Data Ethics	Geuns, Brandusescu 2020	2	0	1	2	2	2	0	2	0	3	4	2	0	2	0	0	0	0
Digital Ethics	Perry 2017	2	1	1	2	2	2	2	2	1	1	2	1	0	2	1	0	0	0
Digital Ethics	Schäfer, Franzske 2020	2	1	2	2	2	1	0	1	0	0	2	2	2	2	0	2	0	0
Digital Ethics	Spiekermann 2019	3	1	1	3	2	2	0	2	1	3	2	2	0	2	0	0	2	0
Digital Ethics	Grimm 2020	2	1	2	2	2	2	0	2	3	2	3	1	2	2	0	0	2	0
Digital Ethics	Zöllner et al. 2015	2	1	1	2	4	3	0	2	2	2	4	1	0	1	0	0	0	0
Digital Ethics	Schlegel 2020	2	1	0	1	2	2	0	1	0	2	2	0	0	0	0	0	0	0
Digital Ethics	Achatz et al.2020	3	1	2	2	1	2	0	1	0	1	2	2	0	2	0	0	0	0
Digital Ethics	Baumann, Donskis 2013	4	1	2	2	2	2	0	2	1	2	2	1	0	1	0	0	1	0
Digital Ethics	Weber, Mangold 2019	2	1	1	2	2	2	1	1	3	3	2	2	0	1	0	0	0	0
Digital Ethics	Ess 2020	2	3	1	2	2	2	0	1	1	2	1	2	1	1	0	0	0	0
Digital Ethics	Gogröf et al. 2017	2	1	2	2	2	1	0	2	2	0	2	2	0	1	0	0	0	0
Digital Ethics/Digital Transformation	Bridle 2018	3	0	3	1	2	2	0	2	3	2	1	2	0	2	0	0	0	0
Digital Power Distribution	Volmar 2019	0	0	1	2	2	2	1	1	1	4	2	1	2	1	0	0	0	0
Digital Ethics of Law	Hoffmann-Riem 2018	2	0	1	2	2	2	1	2	2	2	3	2	2	1	0	1	0	0
Digital Transformation/Digital Change	Becker 2019	2	0	3	2	2	2	0	1	2	0	2	2	1	1	0	0	0	0
Digital Transformation/Digital Change	Kalina et al. 2018	2	0	2	1	1	2	0	2	1	1	2	1	1	1	0	2	0	0
Digital Transformation/Digital Change	Bedford-Strohm et al. 2019	2	0	2	2	2	2	0	2	1	2	2	2	2	1	0	0	0	1
Digital Transformation/Digital Change	Fend, Hofmann 2020	2	0	0	2	2	1	0	1	2	0	1	0	0	2	0	0	2	1
Digital Transformation/Digital Change	Hess 2019	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	1
Digital Transformation/Digital Change	Radermacher 2018	0	0	2	2	0	2	0	0	2	0	0	0	0	0	0	0	0	0
Digital Transformation/Digital Change	Mühlhoff et al. 2019	2	0	1	2	2	3	3	2	1	3	2	2	1	2	0	0	1	1
Digital Transformation/Digital Change	Jaekel 2017	2	0	1	2	1	1	0	1	0	4	1	1	0	2	1	0	0	0
Digital Transformation/Digital Change	Schnell, Dunger 2019	2	0	1	2	1	1	0	1	1	0	2	0	0	2	0	0	0	0
Digital Humanism	Fritz, Tomaschek 2020	2	0	3	2	2	2	0	2	1	3	2	2	1	2	2	0	0	0
Digital Humanism	Werthner et al. 2019	2	0	2	2	2	2	0	2	1	1	1	2	3	3	3	0	0	0
Digital Humanism	Nida-Rümelin, Weidenfeld 2018	2	0	2	3	3	2	0	2	1	1	0	0	2	4	3	0	0	0
Ethics in Graphics Design	Bauer et al. 2015	2	0	1	3	3	0	0	0	1	0	0	0	0	0	0	0	0	0
Ethics in Copyright	Amini 2017	0	0	1	0	0	2	0	0	0	0	0	2	0	0	0	4	0	0
Ethics in Internet Regulation	Küllmer 2019	2	0	1	1	1	2	3	0	2	3	1	1	0	1	0	3	0	0
Ethics in the IT	Spiekermann 2016	2	0	2	2	1	2	0	2	4	1	3	0	1	2	0	0	0	0
Ethics in the IT	Stapf et al. 2021	2	0	3	2	2	4	0	2	2	2	4	0	0	1	0	0	0	0
Ethics in Market Research	Keller et al. 2020	0	0	0	1	0	0	0	0	2	0	3	0	3	3	1	0	0	0
Basics in Ethics	Stoecker et al. 2011	2	1	2	2	1	2	0	2	1	2	1	2	2	0	0	0	0	0
Basics in Ethics	Fenner 2019	2	1	0	2	0	0	0	0	2	0	0	0	0	0	0	0	2	0
Basics in Ethics	Werner 2021	2	1	2	1	1	1	1	1	1	1	1	1	2	1	1	1	1	0
Basics in Ethics	Pieper 2017	2	0	0	1	0	1	0	1	1	1	0	1	1	0	0	0	0	0
Social Theory	Nassehi 2019	0	0	0	2	2	2	0	0	2	0	2	0	0	0	0	0	0	0
Social Theory	Bollmer 2018	2	3	0	1	1	0	0	0	0	0	0	0	0	2	0	0	0	0
Social Theory	Zuboff 2018	2	0	0	2	2	2	0	2	1	2	1	2	0	1	0	0	0	0
Information ethics	Bendel 2019	2	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
Information ethics	Steinebach et al. 2020	2	0	2	2	2	2	1	2	4	2	1	1	2	2	0	0	0	0
Information ethics	Pariser 2011	0	0	2	1	2	0	2	1	4	1	1	1	1	3	0	0	0	0
Information ethics	Hohlfeld et al. 2020	1	0	2	2	2	2	1	2	4	0	0	0	0	2	0	0	0	0

https://doi.org/10.5281/zenodo.8064587



Context or area ethic	Source	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
Internet Search	Unkel 2019	0	0	2	1	2	2	0	0	3	1	1	1	1	1	0	0	0	0
Internet Search	Lewandowski 2021	2	0	2	1	1	1	1	0	2	1	2	1	1	2	0	1	0	0
Internet Search	Giuseppe et al. 2019	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Internet Search	Burghardt et al. 2013	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Internet Search	IAB 2019	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Internet Search	Newton 2016	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Internet Search	Sens 2018	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Internet Search	Brückerhoff 2019	2	0	0	0	2	1	0	0	4	1	1	0	1	1	0	0	0	0
Internet Search	Amato et al. 2019	0	0	0	0	0	0	0	0	3	1	0	0	0	0	0	0	0	0
Internet Search	Stark et al. 2014	2	0	2	2	2	3	1	2	4	1	1	1	1	1	0	0	0	0
Internet Search	Schulz 2015	0	0	2	1	0	0	0	2	0	2	1	1	0	2	0	0	0	0
Internet Search	Schrade 2021	0	0	1	1	1	1	0	1	1	1	1	1	0	0	0	0	0	0
AI-Ethics	Jobin et al. 2019	2	0	1	2	1	1	0	1	1	1	1	1	1	1	0	0	0	1
AI-Ethics	Dilemmas in AI" 2020	2	0	3	3	3	2	0	2	0	0	1	1	0	2	0	0	0	0
AI-Ethics	Serfas et al. 2015	2	0	1	2	1	3	0	0	2	1	2	2	0	4	0	0	0	0
AI-Ethics	Aoun 2017	0	0	4	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0
Machine Ethics	Grimm, Zöllner 2015	2	0	2	3	2	2	0	2	0	1	1	1	1	3	1	0	0	0
Machine Ethics	Capurro 2017	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Machine Ethics	Rath et al. 2019	2	0	1	3	3	2	0	2	0	1	2	2	3	4	2	0	0	0
Media Ethics	Schicha 2019	1	1	2	3	2	2	1	2	2	1	2	2	0	1	0	0	1	0
Media Ethics	Christians, Fackler 2020	2	1	1	2	2	1	0	1	2	1	2	2	1	1	0	0	0	0
Robot Ethics	Iben 2021	2	0	1	1	2	3	0	2	4	2	1	1	1	3	1	0	0	0
Social Media Ethics	Orlowski 2020	2	0	2	2	2	2	0	2	2	2	2	2	0	2	2	0	3	0
Social Media Ethics	Klimczak et al. 2020	2	0	3	2	1	1	1	0	3	0	1	1	2	2	1	1	0	0
Social Media Ethics	Kvalnes 2020	2	0	1	2	2	2	0	1	2	2	2	1	1	0	0	0	0	0
Social Media Ethics	McDonald 2016	2	0	2	2	2	2	2	1	1	3	2	2	2	0	0	0	2	0
Social Media Ethics	Niederer, Rogers 2020	2	0	1	2	2	2	0	1	2	2	2	2	1	2	0	0	0	0
Social Media Ethics	Tufekci 2017	0	0	3	1	2	1	0	1	3	1	2	0	0	2	0	0	0	0
Sociology of Technology	Häußling 2019	2	1	0	0	0	2	0	2	0	0	0	2	0	0	0	0	0	0
Sociology of Technology	Bridle 2018	0	0	3	0	0	2	0	0	0	0	0	0	0	2	2	0	2	0
Sociology of Technology	Kaminsky et al.2020	2	0	2	1	2	0	0	1	2	2	1	1	0	2	0	0	0	0
Business Ethics	Gadatsch et al. 2018	0	0	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	2
Business Ethics	Achenbach et al. 2018	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Business Ethics	Nietsch-Hach 2019	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Business Ethics	Gogröf et al. 2017	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Business Ethics	Pietsch 2020	2	0	1	2	1	1	0	1	0	2	1	2	0	1	0	0	0	2
	Anzahl der Werte > 0 =	65	18	68	74	68	72	18	64	58	59	66	57	39	66	16	11	14	7

Legend:

(A) Ethical value orientation:
Publications that fundamentally address the emergence and orientation of values in a digital world.

(B) Cultural aspects:
Publications in which cultural differences in value concepts and international value discourses in digital ethics are addressed.

(C) Public awareness/media literacy & education:
Publications addressing the role of education, the acquisition of certain skills and the importance of public awareness.

(D) Responsibility:
Publications that address the acceptance or attribution of responsibility for ethical aspects.

(E) Transparency and explicability:
Publications that deal with transparency or accountability in digital worlds.

(F) Control:
Publications that ethically reflect on institutions of control and possibilities of control, including discourses on surveillance.

(G) Censorship and information regulation:
Publications in which information regulation and censorship processes are ethically discussed.

(H) Security and safe guarding:
Publications in which the security of human beings is thematised, on the one hand for the protection of life and on the other hand also for the protection of certain rights.

(I) Knowledge creation and opinion formation:
Publications that deal with the creation of digital knowledge or the formation of opinion in digital environments.

(J) Distribution and abuse of power:
Publications in which monopoly positions or unequal power relations are discussed.

(K) Privacy and anonymity:
Publications that deal with privacy, anonymity and data protection.

(L) Equality and Equal Opportunities:
Publications that deal fundamentally with criteria of fair interaction in digital worlds.

(M) Prejudice and discrimination:
Publications that deal with discrimination and prejudices that arise through processes of digitalisation, including "racist" algorithms, cyber-bullying, social biases.

(N) Automation and ethical algorithms:
Publications that deal with fundamental ethical problems and issues in the context of automation processes, including cooperation processes between humans and technology.

(O) Humanisation of technology:
Publications that deal with ethical problems of the humanisation of technology, e.g. chatbots or AIs that are perceived as human.

(P) Copyright and intellectual property rights:
Publications dealing with intellectual property in digital worlds.

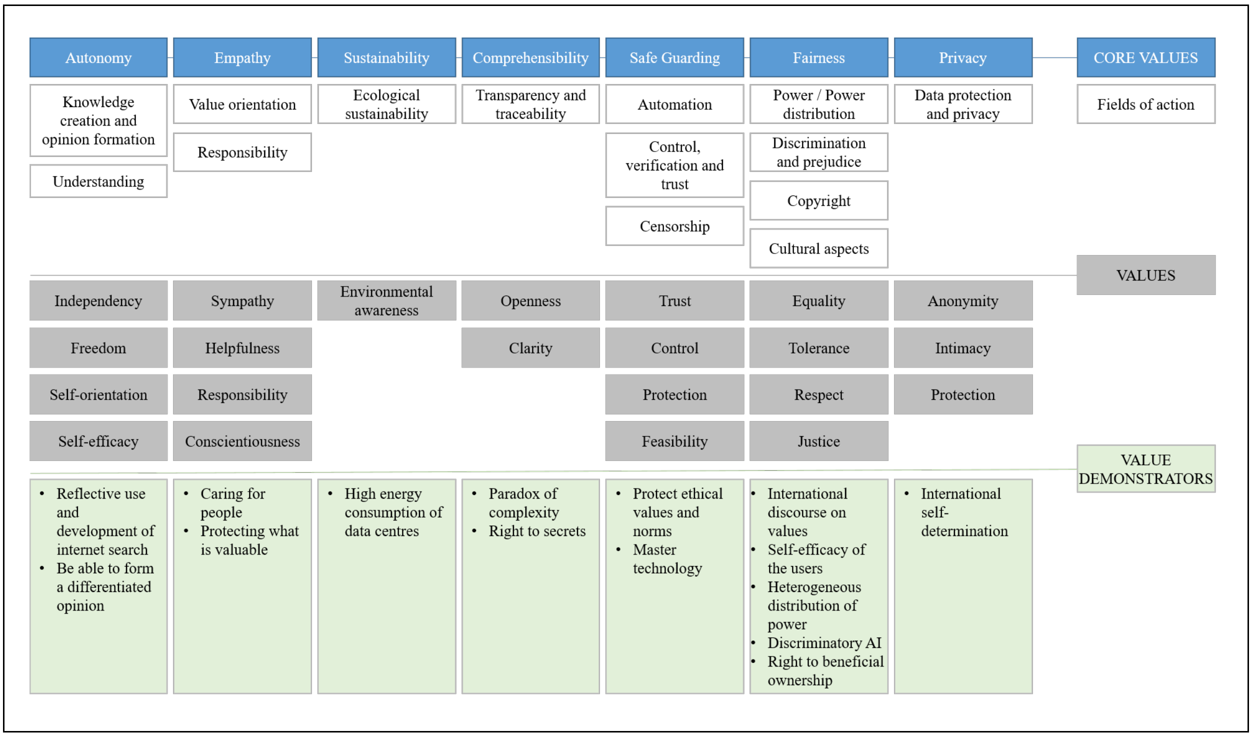
(Q) Addictions and addiction:
Publications dealing with addictions and addictive behaviour in digital worlds, e.g. internet addiction, mobile phone addiction, addiction mechanisms and commercial exploitation.

(R) Ecological sustainability:
Publications that take an ecological perspective on digitalisation.

https://doi.org/10.5281/zenodo.8064587

APPENDIX B

Value matrix incl. core values, values and value demonstrators



Autonomy	Empathy	Sustainability	Comprehensibility	Safe Guarding	Fairness	Privacy	CORE VALUES
Knowledge creation and opinion formation	Value orientation	Ecological sustainability	Transparency and traceability	Automation	Power / Power distribution	Data protection and privacy	Fields of action
Understanding	Responsibility			Control, verification and trust	Discrimination and prejudice		
				Censorship	Copyright		
					Cultural aspects		VALUES
Independency	Sympathy	Environmental awareness	Openness	Trust	Equality	Anonymity	
Freedom	Helpfulness		Clarity	Control	Tolerance	Intimacy	
Self-orientation	Responsibility			Protection	Respect	Protection	
Self-efficacy	Conscientiousness			Feasibility	Justice		
• Reflective use and development of internet search • Be able to form a differentiated opinion	• Caring for people • Protecting what is valuable	• High energy consumption of data centres	• Paradox of complexity • Right to secrets	• Protect ethical values and norms • Master technology	• International discourse on values • Self-efficacy of the users • Heterogeneous distribution of power • Discriminatory AI • Right to beneficial ownership	• International self-determination	VALUE DEMONSTRATORS

https://doi.org/10.5281/zenodo.8064587

APPENDIX C

Questionnaire

A) General fields of actions – Core of the ethic mill of internet search – Value orientation

Value		Questionnaire
1) Value orientation and the ends-means dilemma	**Guiding Principle:** Worry about people	- How can actors (users, search engine providers, governments, etc.) who do not want to act ethically be persuaded to act in the interest of the common good? - How can internet search companies pursue a value strategy and still be profitable? - How can the tension between value strategy and cost pressure be resolved in a capitalist system? - How can it be ensured that internet search technologies are ethically considered from the outset?
	Core Value: Empathy	
2) Public awareness and understanding	**Guiding Principle:** Provide competences	- What education is needed for the different stakeholders of internet search? - What support do people need to feel well informed to search the internet? - How detailed do different stakeholders need to be educated in the three fields of competence in order to be able to make ethical decisions autonomously?
	Core Value: Autonomy	
3) Cultural Differences	**Guiding Principle:** Enable international dialogue on values	- How can ethical standards for internet search be defined in an international context? - What processes and actors are needed for the definition of international ethical standards in internet search?
	Core Value: Fairness	

B) The overarching importance of responsibility, explainability and con-trol

Value		Questionnaire
1) Responsibility	**Core Value:** Empathy	- In internet search, for what must responsibility be assumed by whom towards whom?
2) Transparency and comprehensibility	**Core Value:** Comprehensibility	- What must be made comprehensible to whom and in what form? - What must be made transparent to ensure traceability? - How can good documentation and logical structuring of internet search processes be ensured?
3) Ethical review, control and regulation	**Core Value:** Safe Guarding	- What or who may/must be controlled or checked by whom for what reason and how bindingly? - Which processes/actors of Internet search must be controlled? - Who should be authorised to control Internet search processes and how and, in case of doubt, to regulate them? - How can processes of internet search be designed to enable an ethical exchange and to review applicable ethical norms again and again in the in the specific context?

https://doi.org/10.5281/zenodo.8064587

C) Prioritised ethical fields of action for the internet search

Value		Questionnaire
1) Privacy and data protection	**Guiding Principle:** Protect privacy **Core Value:** Privacy	- How can it be ensured that personal data do not leave the informally protected framework? - How can data protection regulations and laws be adapted to the current data processing methods of Internet search providers? - How can control and traceability over one's own data and its use be guaranteed? - To what extent can the algorithmic analysis of user data in internet searches be made comprehensible for the user? - How can the use of data for personalisation possibly be controlled by the user him/herself? - Does absolute control over one's own data also go hand in hand with absolute anonymity in the use of internet search? - What are the ethical grounds on which a state or an organisation can access personal data in internet search?
2) Knowledge creation and opinion formation	**Guiding Principle:** Enable autonomous opinion-forming **Core Value:** Autonomy	- How and for whom must processes/factors of digital knowledge creation (crawling, indexer, searcher, etc.) by search engine providers be made comprehensible? - What approaches are needed to prevent one-sided influence on opinions? - Should there be a possibility in future to switch off personalisation or to apply it only to certain search results? - Can personalisation also be ensured independently, e.g. by filter systems? - Should advertisements be part of the search result display, and if so, in what way? - What responsibility do search engines have in their role as mediators of disinformation? - Can the spread of misinformation through search engines be restricted?
3) Distribution and abuse of power	**Guiding Principle:** Ensure balanced distribution of power **Core Value:** Autonomy/Fairness	- How can more diversity be ensured in the search engine market? - Which institutions can ethically control the distribution of power in internet search in the most unbiased way possible? - How can data-related economies of scale be minimised in order to prevent monopolies? - For whom must power relations be made comprehensible? - How can end users be empowered in their use of internet search? - How can it be ensured that mechanisms of internet search are not abused by third parties? How transparent can internet search processes be for the public?
4) Automation and the fallibility of algorithms	**Guiding Principle:** Master automation **Core Value:** Safe Guarding	- How can the field of unexpected conditions in Internet search automation processes be minimised? - Which Internet search processes should be automated and which should not? - To what extent can the traceability of automated processes in Internet search be guaranteed? - How can interfaces between different automation processes be controlled and transfer processes be traced? For example, what influence does the archiving of the search index have on the formation of relevance in the search results? - What do the automation processes of Internet search have to look like in order to be controllable? - If the processes of internet search cannot yet be made ethically comprehensible, must the function of internet search be restricted until this is possible?

https://doi.org/10.5281/zenodo.8064587

Value		Questionnaire
5) Discrimination and prejudice	**Guiding Principle:** Minimize search engine biases	- How can data selection procedures be influenced to minimise bias and technically generated discrimination? - How can search algorithms be optimised to avoid erroneous correlations? - To what extent do individualised search results actually systematically lead to disadvantages for different users? - What responsibility do search engine providers have to offer search results equally everywhere in the world?
	Core Value: Fairness	
6) Censorship and the exclusion of content by search engine operators	**Guiding Principle:** Appropriate censorship	- What data may be stored for censorship processes? - How can Internet searches be comprehensibly censored so that freedom of information or opinion are not restricted? - To what extent do search engine providers have to guarantee that third parties and control institutions have insight into their processes? - To what extent should censorship by search engines be dependent on the laws of individual countries?
	Core Value: Safe Guarding	
7) Knowledge Graph results and copyright	**Guiding Principle:** Respect intellectual property	- To what extent may search engines use third party content to make internet search more user-friendly? - How could third parties make their content available for a more efficient internet search without disadvantages for themselves?
	Core Value: Fairness	
8) Ecological sustainability	**Guiding Principle:** Building ecological, sustainable structures	- How can environmentally sustainable data centres for internet search be ensured?
	Core Value: Sustainability	

https://doi.org/10.5281/zenodo.8064759

DESIGNING AN INTEGRATION CONCEPT OF THE PROVENANCE VERIFICATION INDICATOR INTO OPEN WEB SEARCH ENGINES

A. Nussbaumer*, S. M. Ebner, C. Gütl, Graz University of Technology, Graz, Austria
G. Munnelly, B. Spillane, O. Conlan, Trinity College Dublin, Dublin, Ireland
C. Plote, A. Frank, Open Search Foundation, Starnberg, Germany

Abstract

This paper presents a conceptual design on how to combat disinformation by integrating the Verification Indicator into open web search engines. The Verification Indicator is a browser plugin based on a background service that has been developed in the PROVENANCE research project. It provides support for consumers of online news articles to identify problematic content and disinformation by showing warnings on different levels and dimensions. In order to apply this approach in open web search environments, we propose a concept of how it can be transferred and integrated in this field. This concept consists of three levels following the architectural design of search engines. First, documents are analysed in the index generation phase, then in the search and retrieval phase personal settings are taken into account if warnings should be displayed, and finally the warnings are presented in the user interface.

INTRODUCTION

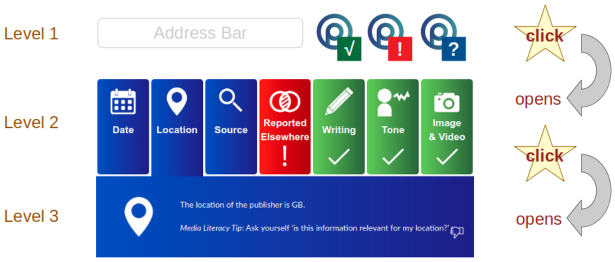
Presently, our society faces an overwhelming amount of fake news and disinformation spread over digital technologies and social media. This situation leads to a general societal challenge that includes the news market, the digital platforms, the social media, and the consumers and their online behaviour [1]. Online news is often distributed via digital platforms and shared in social media. Beside traditional news publishers, an enormous amount of new and often dubious publishers have emerged. Consumers tend to lose overview and control of their media behaviour.

Several approaches have been proposed how to combat this almost ubiquitous problem [2]. Manual fact checking and content verification by non-profit organisations aim at identifying false information on the Web and make them public including the information that is true instead. The automatic detection of fake accounts and related amplification mechanisms of spreading false information should mitigate their spread. Legal initiatives aim at regulating the use of information publication on the national domain, for example by forcing publishers to open up information on themselves. Media literacy training seeks to educate the consumers to identify false information.

This paper presents a conceptual design on how to combat the spread of fake news by integrating the PROVENANCE Verification Indicator in open web search engines. The Verification Indicator is a tool developed in the research project PROVENANCE[1] that informs users of problematic aspects of online news articles while users are visiting them. The conceptual design takes this tool and integrates it in open web search engines on multiple levels based on search engine architectures, including the index generation process, the search and retrieval process, and the user interface. Furthermore, the paper discusses ethical aspects of this approach in relation to the ethical problem that search engines often deliver problematic content.

PROVENANCE VERIFICATION INDICATOR

The Verification Indicator tool has been developed in the context of the PROVENANCE research project with the aim to support users in judging online news. This tool is implemented as a web browser plugin that monitors the articles a user opens and provides warnings in case of problematic content. The tool analyses seven aspects that are relevant for judging online articles, namely the date of an article, the location of the publisher, the publisher of the article, the writing style, the tone of the language, same content reported in (other) reliable news outlets, and the authenticity of the included images. Each of these aspects can result in a warning.



The location of the publisher is GB.

Media Literacy Tip: Ask yourself 'is this information relevant for my location?'

Figure 1: Concept of the PROVENANCE Verification Indicator tool.

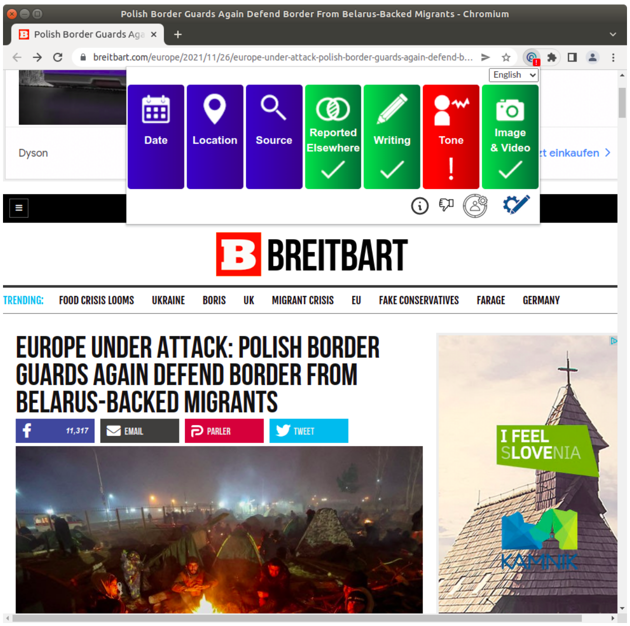
The user interface is organised as a three-warning-level approach (see Figure 1). Level 1 is represented by a single icon in the toolbar of the web browser. It shows a red exclamation mark if one of the seven aspects has a warning, otherwise it shows a green check mark. By clicking on the icon, the level 2 opens where the warnings for the particular aspects are displayed. By clicking on an aspect (with or without warning), further information is given in level 3. An example is shown in Figure 2 where the level 1 warning is displayed in the toolbar and the level 2 warning has been opened.

From a technical point of view, complex analysis is being made in the background, which is not explained in detail

* alexander.nussbaumer@tugraz.at

[1] http://provenanceh2020.eu/

https://doi.org/10.5281/zenodo.8064759


Polish Border Guards Again Defend Border From Belarus-Backed Migrants - Chromium

breitbart.com/europe/2021/11/26/europe-under-attack-polish-border-guards-again-defend-b...

EUROPE UNDER ATTACK: POLISH BORDER GUARDS AGAIN DEFEND BORDER FROM BELARUS-BACKED MIGRANTS

Figure 2: Example of the PROVENANCE Verification Indicator tool.

in this paper. The analysis of the writing style and tone is being made with machine learning algorithms that are trained with respective problematic and non-problematic articles. The analysis of similar content elsewhere is made with semantic analyses of large amounts of new articles. The image analysis is performed with image reverse search techniques and comparison analysis if images have been manipulated. The first three information categories (date, location, and source) are retrieved by analysing the content of the article's webpage and metadata.

It is important to note that the calculations and warnings have to be critically reflected by the user, since the algorithms can fail in their judgements. This matter of fact is exploited by presenting media literacy tips that ask the user to critically reflect the presented tool information.

INTEGRATION CONCEPT

The Verification Indicator in its current implementation supports users to identify disinformation while reading online news articles. Finding and navigating online news articles are often facilitated through bookmarked websites of news publishers and social media, but also through web search. Thus it makes sense to provide warning information already before opening online news articles, which can be achieved by tagging web documents in the search results with warnings.

This section presents a conceptual design of how the Verification Indicator tool can be integrated in an open web search engine [3]. The integration concept is based on a simplified technical architecture of search engines that consists of two key components. First, the index generation component collects web documents and stores them locally in an internal format. Second, the search and retrieval component searches the index and presents the search result to the user. The user interface is a key part of the integration concept, as it presents the warnings of problematic documents to the user. An overview of the integration concept is depicted in Figure 3.

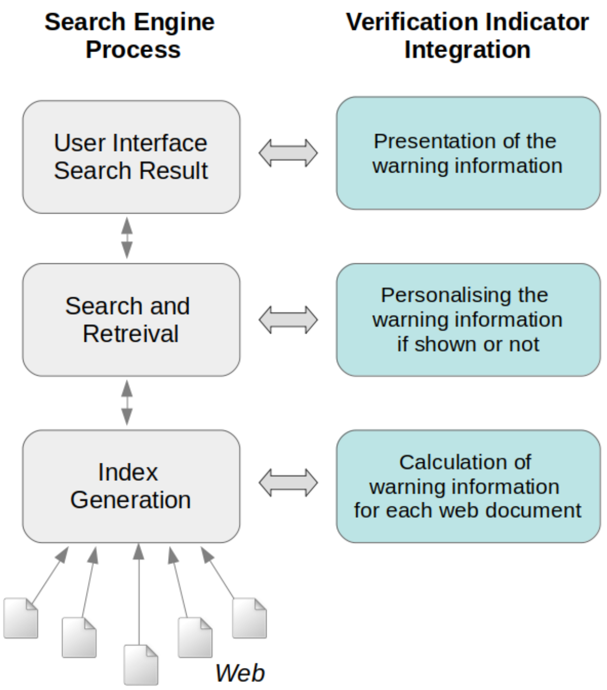

Search Engine Process
Verification Indicator Integration
User Interface Search Result
Presentation of the warning information
Search and Retrieval
Personalising the warning information if shown or not
Index Generation
Calculation of warning information for each web document
Web

Figure 3: Overview of the integration concept.

Index creation component

The index generation component of search engines facilitates the collection and storage of web documents. Websites are systematically crawled and stored to be used for further processing. They are analysed and stored in an internal representation format for enabling efficient search processes. Furthermore, metadata is extracted from the web documents and stored along with them, which is also used by the search and retrieval component. Moreover, content analysis is made to retrieve even further information from the web documents, which can be included in the search and ranking algorithms.

In order to enable disinformation detection, information categories as used by the Verification Indicator tool (see last section) can be derived from each web document and stored as additional metadata. While some information categories are rather easy to extract, others require complex and costly analysis. For example, the *date* and *source* can easily be extracted from the documents, but *reported elsewhere*, *writing quality*, and *tone* need complex calculations. Obviously, the first ones should be calculated during the indexing process, but the latter ones should rather be calculated in a separate and asynchronous process. However, this would lead to a situation where only some of the Verification Indicator aspects are always available, but the others might be missing for some time.

Search and retrieval component

The Verification Indicator is capable of personalising the warnings depending on the media literacy level and preferences of the user. The warning sign on level one is suppressed for users with high media literacy level of if a user makes a respective setting. However, if the user clicks on the icon, the warnings show up on level two and three. Thus the user is prevented from an overload of warnings, but still get warnings in certain situations.

To some extent, this approach can be taken over to open search engines. Presenting warning signs can be omitted in the presentation of the search result (see next subsection), if the media literacy level of the preferences suggests a suppression in the same way as for the PROVENANCE context. The media literacy level is calculated by the frequency a user visits a certain topic. Calculating a media literacy level would require the search engine to trace the user behaviour. However, for ethical reasons the monitoring of the user behaviour is often declined in open search engine context. On the other hand, taking into account user preferences if warning signs should be suppressed for particular topics or information categories is better accepted. In any case the search and retrieval component could take over the decision if a warning is shown on level one or not.

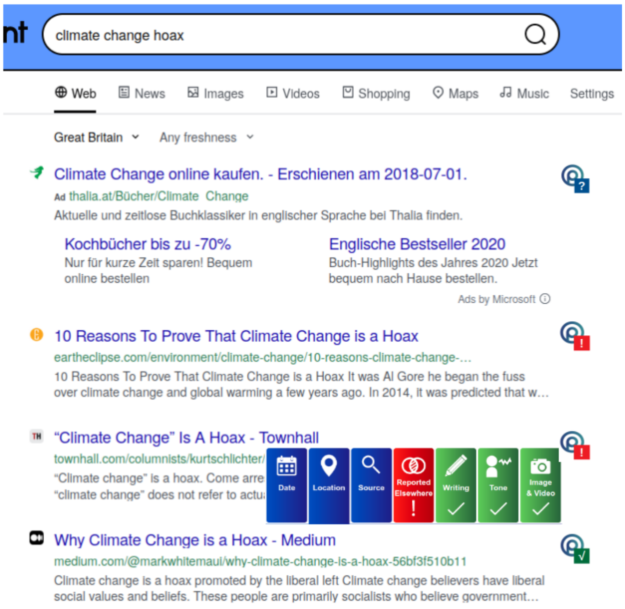
User interface component

The integration of the user interface of the Verification Indicator into web search engines is straight forward. On a result page the warning icon can be displayed next to each found web document. Depending on the document an icon with warning or no warning can be shown. Clicking on an icon will provide level 2 and level 3 information as described above. This concept is outlined in Figure 4. A user might benefit from this approach, as a simple and quick overview on the quality of the documents in the result page is shown. Thus, the user can make an informed decision about which documents to be opened.

ETHICAL IMPLICATIONS

There is an ethical problem inherent to all search engines. If web documents appear on a result page, a user might perceive such information as trustful recommendations. People trust computers for different reasons, such as they think that computers are always right or they do not have enough time to critically reflect the computer-generated results. However, search engines do not automatically detect content that conveys disinformation. This leads to a situation where problematic content is displayed in the same way as non-problematic content. Users who typically search for true information might follow the links to problematic contents, read and share them without understanding its problematic nature. This would conflict with an ethical value of *trustful information* that a user typically expects from digital applications.

The approach presented in this paper addresses this ethical problem. Web documents in the search result would be tagged with warning icons if their content is problematic. In this way information of problematic web documents is provided and problematic documents are distinguished from non-problematic documents. Thus the ethical value *trustful information* is addressed. Even if the computer-based judgement can be incorrect, the users are asked to critically reflect the given information. Beside concrete warnings, this approach might also increase general critical thinking, behaviour, and competence in relation to search results.



climate change hoax

Great Britain ⌄ Any freshness ⌄

Climate Change online kaufen. - Erschienen am 2018-07-01.
Ad thalia.at/Bücher/Climate Change
Aktuelle und zeitlose Buchklassiker in englischer Sprache bei Thalia finden.

Kochbücher bis zu -70%
Nur für kurze Zeit sparen! Bequem online bestellen

Englische Bestseller 2020
Buch-Highlights des Jahres 2020 Jetzt bequem nach Hause bestellen.

Ads by Microsoft ⓘ

10 Reasons To Prove That Climate Change is a Hoax
eartheclipse.com/environment/climate-change/10-reasons-climate-change-...
10 Reasons To Prove That Climate Change is a Hoax It was Al Gore he began the fuss over climate change and global warming a few years ago. In 2014, it was predicted that w...

"Climate Change" Is A Hoax - Townhall
townhall.com/columnists/kurtschlichter/
"Climate change" is a hoax. Come arre
"climate change" does not refer to actu

Why Climate Change is a Hoax - Medium
medium.com/@markwhitemaui/why-climate-change-is-a-hoax-56bf3f510b11
Climate change is a hoax promoted by the liberal left Climate change believers have liberal social values and beliefs. These people are primarily socialists who believe government...

Figure 4: The user interface concept for integration of the Verification Indicator tool. The diagram is a mockup based on a search result from https://www.qwant.com/.

Moreover, the approach also addresses an ethical dilemma. If the search system (content analysis during the indexing) identifies a web document as problematic, a decision has to be made if it should be included in a search result or not. Our approach allows a solution to this dilemma by including the document and tag it with a warning. Thus the system does not withhold the document from the user, but also does not include it as a regular item in the search result.

In this way, the integration of the Verification Indicator constitutes an ethics-by-design approach. Ethics-by-design (or value-by-design) seeks to integrate ethical values in the overall software design process [4][5]. Instead of just asking the user to use the software or tool in an ethical way, ethics-by-design approaches aim to handle ethicals problems by integrating values in the design from the very beginning. Ethical considerations should be made during the whole research and development process and integrating at all phases and components of a technology [6]. So the ethical responsibility is already addressed during the reseaerch and software design process, in order to create a technology that supports users to act in an ethically sound manner.

https://doi.org/10.5281/zenodo.8064759

CONCLUSION

In this paper an integration concept is presented that outlines how the Verification Indicator tool of the PROVENANCE project can be integrated in open web search engines. While the Verification Indicator in its current version is used in combination with opened online articles, the integrated version can be used in combination with search results before opening web documents, which provides an additional type of warning of problematic content. Furthermore, this approach might also be capable of increasing trust in search engines, as it tags problematic content with warning information.

A drawback of the presented approach is the dependency of the Verification Indicator and the required back-end technology that performs the various analyses. There is no guarantee that the PROVENANCE technology will be maintained and available for a long time. However, the integration concept can be modified in a way that such dependency of existing technology can be overcome. During the indexing process web documents are analysed anyway regarding various features. In this processing phase further analysis can be conducted in the same way as in PROVENANCE or at least in a way that has relevance to disinformation. Even if this analysis would be simpler and might lead to a set of different information categories, there is still a chance to support users to identify problematic content. Further research can be undertaken to find out which other information categories can support the identification of disinformation. Since the main goal of the information categories is to warn users and stimulate critical thinking, there is a good chance to discover a range of further categories that have positive influence. In any case. a new evaluation would be needed regarding the effectiveness of such a new set of information categories.

ACKNOWLEDGEMENTS

The work reported has been partially supported by the PROVENANCE project which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 825227, as well as the OpenWebSearch.EU project which has received funding from the European Union's Horizon Europe programme under grant agreement no. 101070014.

REFERENCES

[1] Á. Figueira and L. Oliveira, "The current state of fake news: Challenges and opportunities," *Procedia Computer Science*, vol. 121, pp. 817–825, 2017. DOI: 10.1016/j.procs.2017.11.106.

[2] A. Alaphilippe, A. Gizikis, C. Hanot, and K. Bontcheva, *Automated tackling of disinformation*, 2019. https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624278/EPRS_STU(2019)624278_EN.pdf

[3] R. Baeza-Yates and B. Ribeiro-Neto, *Modern information retrieval*. New York, USA: ACM press, 1999, vol. 463.

[4] M. L. Cummings, "Integrating ethics in design through the value-sensitive design approach," *Science and Engineering Ethics*, pp. 701–715, 2006.

[5] B. Friedman, P. H. Kahn, A. Borning, and A. Huldtgren, "Value sensitive design and information systems," in *Early engagement and new technologies: Opening up the laboratory*, N. Doorn, D. Schuurbiers, I. van de Poel, and M. E. Gorman, Eds. Dordrecht: Springer Netherlands, 2013, pp. 55–95. DOI: 10.1007/978-94-007-7844-3_4.

[6] A. Nussbaumer, A. Pope, and K. Neville, "A framework for applying ethics-by-design to decision support systems for emergency management," *Information Systems Journal*, 2021. DOI: 10.1111/isj.12350.

https://doi.org/10.5281/zenodo.8066392

Developing and promoting search engine literacy in primary education

Melanie Platz, Saarland University, 66123 Saarbrücken, Germany
Friederike Klan, German Aerospace Center (DLR), 07745 Jena, Germany
Alexander Decker, Technische Hochschule Ingolstadt, 85049 Ingolstadt, Germany

Abstract

Students use the internet every day, but the lack of transparency regarding information filtering employed by search engines, user profiling done based on search queries, as well as the design of the user interfaces of popular platforms, lead to increasing immaturity in user behavior and low-risk awareness concerning privacy. This begins with the choice of a search engine. As early as elementary school age, children need the competence, both in the school context and in everyday life, to search for information accurately, evaluate search results, and assess possible risks from disclosing personal data. However, this is not explicitly addressed in previous school concepts. In this paper, we present the first ideas for a teaching concept for promoting search engine literacy (SEL) in primary education.

WHAT IS SEARCH ENGINE LITERACY?

Already in primary school, children are expected to conduct internet searches [1; 2]. Competent searching poses complex challenges, the mastering of which requires a corresponding competence - search engine literacy. To define search engine literacy (SEL), the concepts of information literacy and search literacy are considered [3]:

- **Information literacy** is the ability to understand when information is needed, to seek information efficiently, and evaluate and use information appropriately. It also includes integrating new information with prior knowledge and using it legally, economically, socially, and ethically correct to achieve goals.
- **Search literacy** is a specific aspect of information literacy. It relates directly to the process of obtaining information and refers to the ability to find and access the desired information to satisfy information needs efficiently and effectively.
- Achieving accurate search results requires **search engine literacy**, which is the knowledge of how search engines work and the following: findability, linguistic functions, query language, and ranking [4].

However, search engine literacy can only be achieved inadequately by "learning-by-doing" in the sense of discovery learning - its acquisition must instead be supported in a targeted manner. To implement the topic search engines into the primary school curriculum, fundamental ideas must be identified and linked directly to the existing curriculum [5].

STATE OF THE ART: SEARCH AND INFORMATION LITERACY IN SCHOOL

We address the primary level to contribute to the early education of self-confident and self-determined use of search engines. Primary school children mostly use search engines [6; 7] without knowing or questioning how they work [8]. International studies have investigated the promotion of various aspects of information literacy [9; 10]. Specifically for search literacy, individual concepts exist, including co-creative approaches [11], that aim to help users understand a search process [8; 12; 13]. Due to their complexity, these are less suitable for primary education or only deal with the operation of a search engine [14]. The internal functional principles remain a "black box". Research on the assessment of privacy and disclosure of personal data among children rarely focuses on search engines and risks by deriving user profiles [15].

Surveys with teachers [16] show that promoting information literacy is challenging to implement in the classroom and that integrative, resource-saving concepts are needed. The development and acquisition of search engine literacy cannot be assigned to a single learning area [1].

TEACHING CONCEPT

Consequently, we provide a concept that can be used in elementary school lessons in combination with traditional teaching topics without taking up additional teaching time.

We were able to show that modules to promote SEL can be developed by linking to the traditional content of mathematics instruction [5] and that learning with and about media [17] can be supported by specific elementary school-appropriate digital environments [18; 19], and that co-creative design of digital applications is promising [20]. Observations from the first pilot studies point to gender-specific effects. SEL can be promoted at three levels that address aspects that are interrelated (see figure 1).

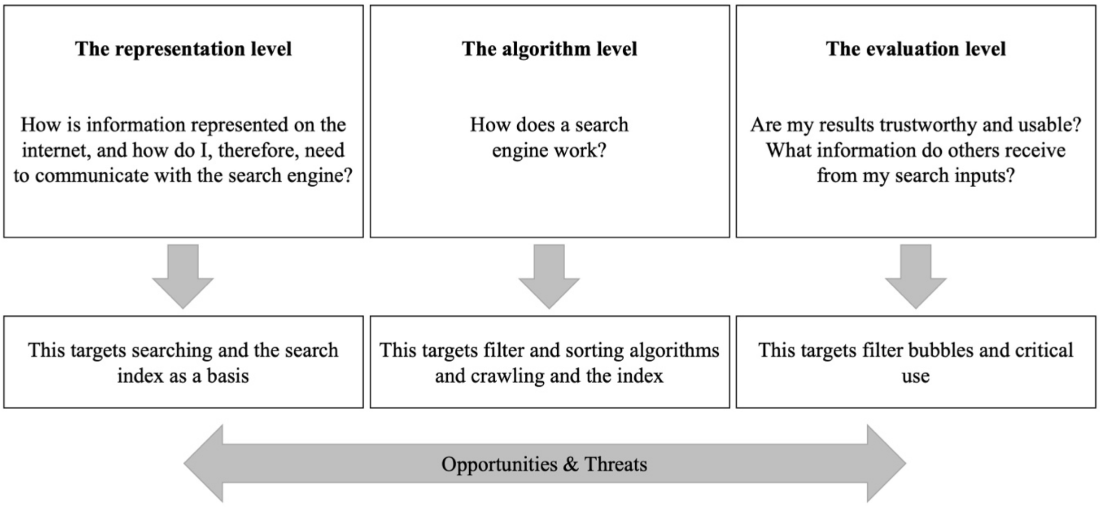
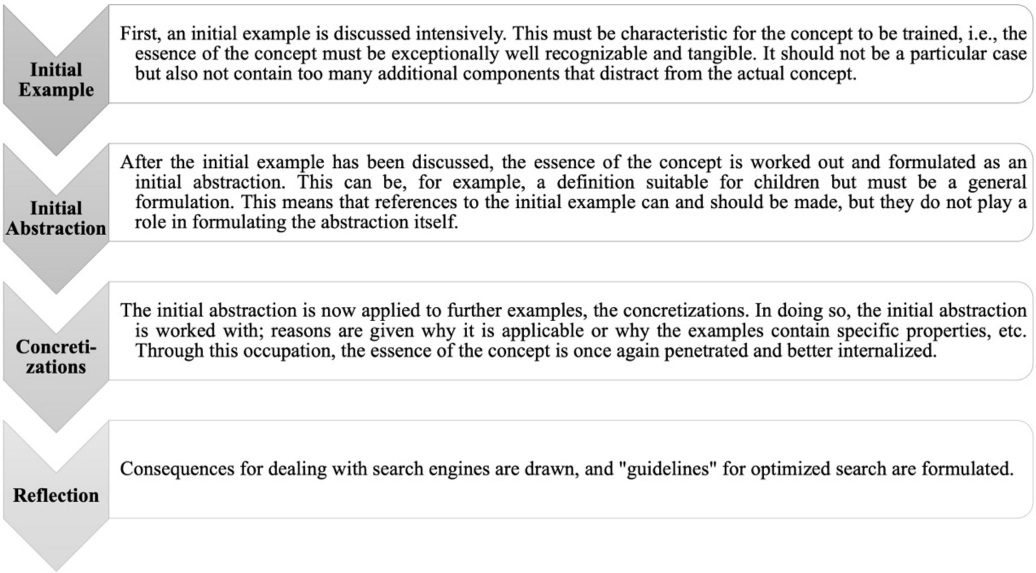

The representation level
How is information represented on the internet, and how do I, therefore, need to communicate with the search engine?
The algorithm level
How does a search engine work?
The evaluation level
Are my results trustworthy and usable? What information do others receive from my search inputs?
This targets searching and the search index as a basis
This targets filter and sorting algorithms and crawling and the index
This targets filter bubbles and critical use
Opportunities & Threats

Figure 1: Three levels of SEL.

For the well-founded penetration of a concept, we transferred the idea of [21] (topic area "algorithms") to "search engines" and extended it by a reflection (see figure 2):



First, an initial example is discussed intensively. This must be characteristic for the concept to be trained, i.e., the essence of the concept must be exceptionally well recognizable and tangible. It should not be a particular case but also not contain too many additional components that distract from the actual concept.

After the initial example has been discussed, the essence of the concept is worked out and formulated as an initial abstraction. This can be, for example, a definition suitable for children but must be a general formulation. This means that references to the initial example can and should be made, but they do not play a role in formulating the abstraction itself.

The initial abstraction is now applied to further examples, the concretizations. In doing so, the initial abstraction is worked with; reasons are given why it is applicable or why the examples contain specific properties, etc. Through this occupation, the essence of the concept is once again penetrated and better internalized.

Consequences for dealing with search engines are drawn, and "guidelines" for optimized search are formulated.

Figure 2: Model for the theoretically founded penetration of a concept; originally on the topic of "algorithms" [21], here transferred to "search engines" and extended by a reflection.

Focusing on the aspect of ranking, an example for the application of this model in primary school in connection with traditional teaching topics in mathematics could be as follows:

Ranking is about putting the search results found in a sensible order, so that the relevant results are displayed first and the less relevant results further down the list. This involves sorting by descending relevance, i.e., the higher a hit is in the list, the more relevant it is. Search engines do not disclose their ranking procedures, yet six areas have emerged that determine how results are ranked: (1) text-specific factors, (2) popularity, (3) recency, (4) locality, (5) personalization, and (6) technical ranking factors [22].

In this example, we focus on (1): Text statistics are used to compare search queries and documents. The first assumption in text statistics ranking is that the entered search term occurs in the document. One assumes that the searcher wants to find the entered term in the document. With this first step, all other documents are excluded, i.e., a first narrowing down of the search result. While the entire index must first be searched, a base set to which the further ranking factors are applied is now determined. All further operations are only performed on this smaller set [22]. First,

https://doi.org/10.5281/zenodo.8066392

one could assume that a document in which the search term occurs particularly frequently is more relevant than one in which the search term occurs less frequently. The document containing the most frequent search term would appear first in the list.

- **Initial Example**: To treat this first idea of ranking based on frequency determination at the primary school level, a search of text sections of the same length for a specific search term can be carried out as a starting example. Initially, texts of the same length are used because absolute frequencies are dealt with first, and longer texts would have a better chance of containing a word frequently.
- The **initial abstraction** to this simple ranking based on frequency determination would be, in this case, for example: A ranking of search results for a search term can be created by first excluding the texts, mathematical objects, etc., that do not contain the search term. Afterward, the order in which the texts, mathematical objects, etc., are to be displayed can be determined using frequency determination. The text, mathematical object, etc., in which the search term is found most often, comes first, the text, mathematical object, etc., in which the search term occurs second most often, comes second, and so on.
- **Concretizations** regarding primary school mathematics could, for example, look as follows:
    - ○ On the set of numbers up to 20, "large, even" is searched for. (Even numbers are divisible by 2 without remainders.) In the first step, all odd numbers are excluded. In the second step, the remaining even numbers are arranged in descending order (as 2 fits or is contained most often in 20, etc.): (20, 18, 16, 14, 12, 8, 6, 4, 2).
    - ○ On the set of geometric solids {cube, sphere, cone, triangular-based pyramid, cylinder} "face" is searched. Each of the geometric solids has faces. They are arranged in descending order according to the number of faces: (cube, triangular-based pyramid, cylinder, cone, sphere).
- For **reflection**, the set we are looking at can be changed; for example, the determination of an exact order and the comparability of the objects can be made more complex to initiate the consideration of structural information in documents, if necessary. In this way, a discussion can be prompted about what "meaningful" search results are that help the searcher: the idea of quality control can be stimulated (see, for example figure 3). For one's search, it can be deduced that the formulation of the search query is essential and that not only the first hits should be considered but also subsequent ones since the ranking "is always only one of many possible algorithmic views of the contents of the World Wide Web" ([22], p. 93).

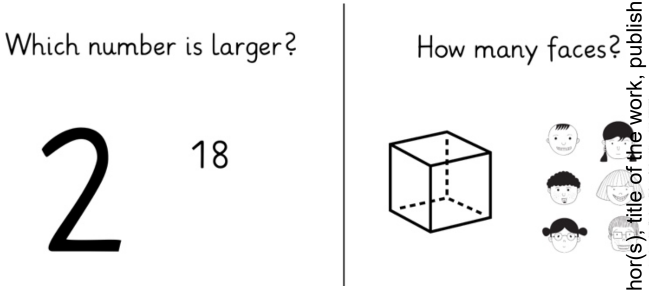
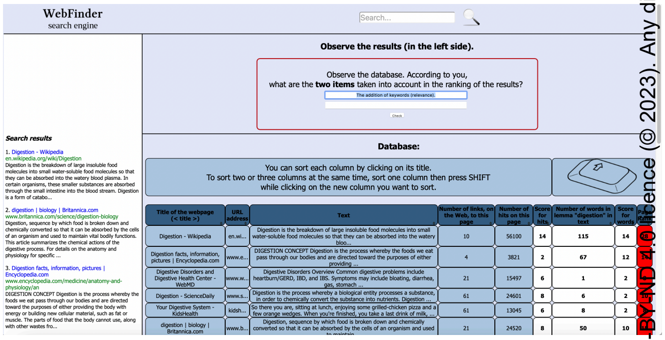


Figure 3: “large number” and “faces”

Based on this idea, existing concepts of SEL can be transferred to the primary school level. For example, it could be shown that an interactive visualization of a filter bubble increases the awareness of the concept and the comprehensibility of the filtering mechanism [23]. The Webfinder tool (https://iddocs.fr/webfinder/index.php; see a screenshot in figure 4) was developed for use in secondary education, and it allows us to understand the steps of classification as well as methods of automatic processing [8]. It additionally enables understanding of the criteria for ranking information according to the principles of relevance (number of occurrences related to the query), popularity (number of visits), and familiarity (number of links). This also helps learners understand an algorithm with a weighting system.



Observe the results (in the left side).

Observe the database. According to you,
what are the **two items** taken into account in the ranking of the results?

The addition of keywords (relevance).

Search results

1. Digestion - Wikipedia
en.wikipedia.org/wiki/Digestion
Digestion is the breakdown of large insoluble food molecules into small water-soluble food molecules so that they can be absorbed into the watery blood plasma. In certain organisms, these smaller substances are absorbed through the small intestine into the blood stream. Digestion is a form of catabo...

2. digestion | biology | Britannica.com
www.britannica.com/science/digestion-biology
Digestion, sequence by which food is broken down and chemically converted so that it can be absorbed by the cells of an organism and used to maintain vital bodily functions. This article summarizes the chemical actions of the digestive process. For details on the anatomy and physiology for specific ...

3. Digestion facts, information, pictures | Encyclopedia.com
www.encyclopedia.com/medicine/anatomy-and-physiology/an
DIGESTION CONCEPT Digestion is the process whereby the foods we eat pass through our bodies and are directed toward the purposes of either providing the body with energy or building new cellular material, such as fat or muscle. The parts of food that the body cannot use, along with other wastes fro...

Database:

You can sort each column by clicking on its title.
To sort two or three columns at the same time, sort one column then press SHIFT
while clicking on the new column you want to sort.

Title of the webpage (< title >)	URL address	Text	Number of links, on the Web, to this page	Number of hits on this page	Score for hits	Number of words in lemma "digestion" in text	Score for words	Page Rank
Digestion - Wikipedia	en.wi...	Digestion is the breakdown of large insoluble food molecules into small water-soluble food molecules so that they can be absorbed into the watery bloo...	10	56100	14	115	14	28
Digestion facts, information, pictures | Encyclopedia.com	www.e...	DIGESTION CONCEPT Digestion is the process whereby the foods we eat pass through our bodies and are directed toward the purposes of either providing ...	4	3821	2	67	12	14
Digestive Disorders and Digestive Health Center - WebMD	www.w...	Digestive Disorders Overview Common digestive problems include heartburn/GERD, IBD, and IBS. Symptoms may include bloating, diarrhea, gas, stomach ...	21	15497	6	1	2	[illegible]
Digestion - ScienceDaily	www.s...	Digestion is the process whereby a biological entity processes a substance, in order to chemically convert the substance into nutrients. Digestion ...	61	24601	8	6	2	10
Your Digestive System - KidsHealth	kidsh...	So there you are, sitting at lunch, enjoying some grilled-chicken pizza and a few orange wedges. When you're finished, you take a last drink of milk, ...	61	13045	6	8	2	8
digestion | biology | Britannica.com	www.b...	Digestion, sequence by which food is broken down and chemically converted so that it can be absorbed by the cells of an organism and used to maintain	21	24520	8	50	10	[illegible]

Figure 4: Screenshot of the Webfinder tool.

Among others, mathematical concepts can describe and explain how search engines work. A key element in promoting SEL is reflecting on one's actions due to engagement with algorithmic selection and sorting functions [24]. Co-creative approaches can contribute to this in particular [11; 25] and develop great potential in the context of SEL if school challenges (time constraints, class sizes, teachers' concerns; [26]) are overcome.

The underlying concepts or fundamental ideas of the topic "search engines" must be identified to address all facets of SEL. Several teaching units are developed in connection to traditional contents treated in primary school on two levels:

1. The question "How does a search engine work?" is focused on the first level. Through relation, reflection, and argumentation, "guidelines" for optimized search are derived: "What do I need to consider when using search engines?; Why is it important?"

https://doi.org/10.5281/zenodo.8066392

2. Through problem-solving and reasoning, search strategies are developed (2nd Big6-Stage, [27]) "How do I get the best search results?; Which strategy fits when?"

Teaching units addressing the first level were developed and tested with pupils as a basis for the second level by primary school student teachers at Saarland University. The teaching units are published as Open Educational Resources at Wikiversity (currently only available in the German language):

https://de.wikiversity.org/wiki/Open-Source4School/Lernumgebungen_zur_Informatischen_Bildung_im_Mathematikunterricht_der_Primarstufe#Search_Engine_Literacy

Preliminary results of the teaching experiments show that the primary school children are very knowledgeable and understand a lot. But there are also some challenges: e.g., "meaningful" input into the search engine was not easy for the children: Instead of only entering the relevant search terms, the children tried to form complete sentences (e.g., instead of “gift”, “I want a gift” was entered).

The teaching concept for promoting SEL is (further) developed by utilizing Design Science Research (DSR). The Dortmund Model (subject didactic development research on diagnosis-guided teaching-learning processes for research and further development of instruction) [28] is combined with the DSR Methodology Process (research and further development of Information Systems Research) [29] to be able to productively use synergy effects of both approaches for the development of technology-supported learning environments [30]. By linking and transferring this idea to other primary school subjects like science and ethics and reflecting on the effect of overcoming the black box effect (in the sense of world-encompassing learning with and about search engines), critical basic education in information and communication technology is promoted. Through such networking, the broader social dimensions of SEL leading to questions in the context of democracy and social participation can be addressed.

CONCLUSION

A teaching concept for the promotion of search engine literacy (SEL) in primary schools is developed in close cooperation with the educational practitioners. Its effectiveness will first be investigated in selected German student laboratories, then evaluated nationwide and finally transferred to schools (see figure 5).

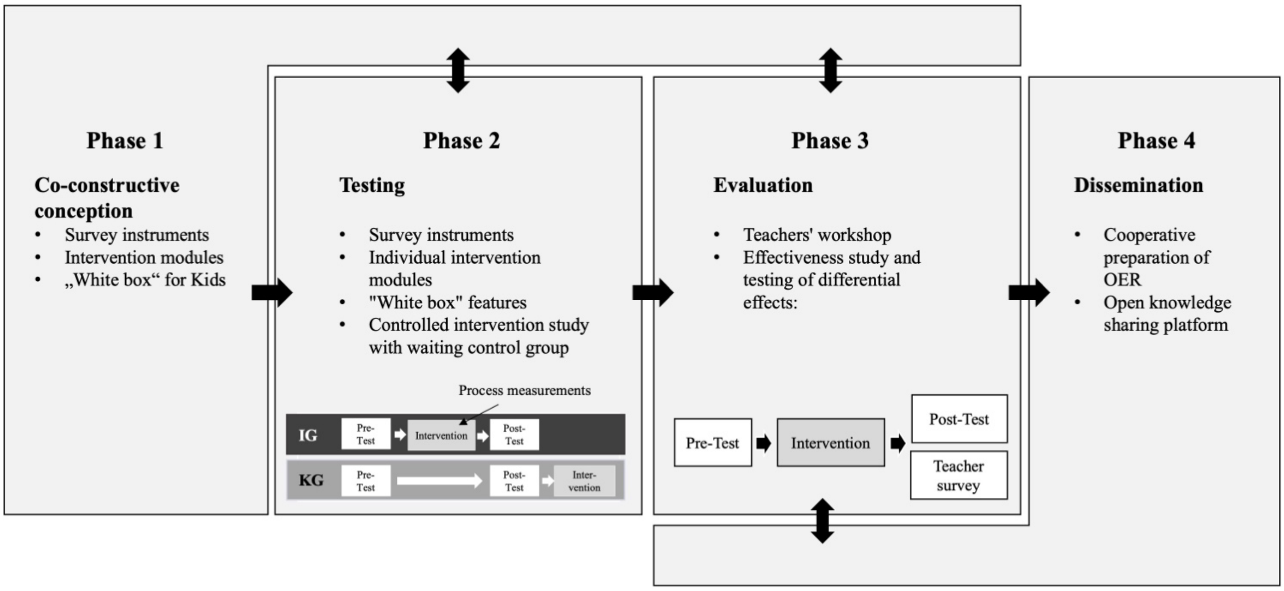


Figure 5: Project phases.

https://doi.org/10.5281/zenodo.8066392

REFERENCES

[1] KMK. *Strategie der Kultusministerkonferenz „Bildung in der digitalen Welt“* (Beschluss vom 08.12.2016). Berlin: Eigendruck; 2016. Abrufbar unter: https://www.kmk.org/fileadmin/pdf/PresseUndAktuelles/2018/Digitalstrategie_2017_mit_Weiterbildung.pdf.

[2] GDSU. *Positionspapier Sachunterricht und Digitalisierung. Erarbeitet von der AG Medien & Digitalisierung der Gesellschaft für Didaktik des Sachunterrichts*; 2021. Abrufbar unter: https://gdsu.de/sites/default/files/PDF/GDSU_2021_Positionspapier_Sachunterricht_und_Digitalisierung_deutsch_de.pdf.

[3] Karatassis, I. A gamification framework for enhancing search literacy. *Proceedings of the 6th BCS-IRSG Symposium on Future Directions in Information Access* 2015; 6: 3–6.

[4] Fuhr, N. *Internet search engines – Lecture script for the course in SS 2014*, 2014. Available online at http://www.is.inf.uni-due.de/courses/ir_ss14/ISMs_1-7.pdf

[5] Platz M, Müller L, Niehaus E, Müller S. Modules for Open Search in Mathematics Teaching. In *Proceedings of the 3rd OSSYM, CERN*. 2021. Abrufbar unter: https://doi.org/10.5281/zenodo.5772576.

[6] Medienpädagogischer Forschungsverbund Südwest. Kinder, Internet, Medien (KIM). *Basisuntersuchung zum Medienumgang 6- bis 13-Jähriger in Deutschland. Mpfs*; 2020. Abrufbar unter https://www.mpfs.de/fileadmin/files/Studien/KIM/2020/KIM-Studie2020_WEB_final.pdf.

[7] Feil Ch, Gieger Ch, Grobbing A. *Projekt: Informationsverhalten von Kindern im Internet – eine empirische Studie zur Nutzung von Suchmaschinen. München: Deutsches Jugendinstitut*; 2013. Abrufbar unter: https://www.dji.de/fileadmin/user_upload/www-kinderseiten/898/1-BMBF-Fkz%2001PF08017.pdf.

[8] Le Deuff O. Search engine literacy. In *Proceedings of the European Conference on Information Literacy* 2017; 359–365.

[9] Aharony N, Gur H. The relationships between personality, perceptual, cognitive and technological variables and students' level of information literacy. *Journal of Librarianship and Information Science* 2017; 51(2): 527–544.

[10] Beltran-Sanchez J A, Lopez R I G, Ramirez-Montoya M S, Quintana J T. Factors influencing the Integration of the Digital Literacy and Inclusion Program into Primary School Teaching. *Revista Electronica De Investigacion Educativa* 2019, 21: 1–11.

[11] Beheshti J, AlGhamdi M J, Cole C, Abuhimed D, Lamoureux I. Designing an Intervention Tool for Students with Students. New Directions in Children's and Adolescents' Information Behavior Research. *Library and Information Science* 2017, 10: 295–331.

[12] Nuxoll F. *Medienwelten Grundschule Arbeitsheft 3/4.* Braunschweig: Westermann; 2021

[13] Wilson M L, Ye C, Twidale M B, Grasse H. Search literacy: Learning to search to learn. In *Proceedings of SAL@ SIGIR*; 2016; Pisa, Italy.

[14] Moreno-Morilla C, Guzman-Simon F, Garcia-Jimenez E. Digital and information literacy inside and outside Spanish primary education schools. Learning Culture and Social Interaction 2021, 28(1): 1–21.

[15] Stoilova M, Nandagiri R, Livingstone S. Children's Understanding of Personal Data and Privacy Online – A Systematic Evidence Mapping. *Inf Commun Soc* 2019; 6(1): 1–19.

[16] Batool S H, Webber S. Mapping the state of information literacy education in primary schools: The case of Pakistan. *Libr Inf Sci Res* 2019; 41(2): 123–131.

[17] Peschel M. Welterschließung als sachunterrichtliches Lernen mit und über digitale Medien - Lernen mit und über digitale Medien als Ausgangspunkt einer umfassenden Sachbildung. In: *Thumel M, Kammerl R, Irion T, Hrsg. Digitale Bildung im Grundschulalter—Grundsatzfragen zum Primat des Pädagogischen.* München: kopaed; 2020. S. 341–355.

[18] Bach S. *Subjektiver Kompetenzerwerb von Schülerinnen und Schülern beim unterrichtlichen Einsatz von kidi-Maps - Eine Studie zum Einsatz digitaler Karten am Beispiel von kidi-Maps im Vergleich zu analogen Karten bei Schülerinnen und Schülern einer vierten Jahrgangsstufe im geographisch-orientierten Sachunterricht.* Saarbrücken: Universität des Saarlandes; 2018.

[19] Schirra S, Peschel M. Recherchieren, Dokumentieren und Präsentieren mit kidipedia im Zeitalter von Tablet & Co. In: Peschel M, Irion T, Hrsg. *Neue Medien in der Grundschule 2.0. Grundlagen – Konzepte - Perspektiven.* Frankfurt am Main: Grundschulverband; 2016. S. 235–246.

[20] Klan F, Kyba C, Schulte-Römer N, Kuechly H. Co-Designing Mobile Applications for Citizen Science Projects - First Results of the Nachtlicht-BüHNE Project. In *Proceedings of the 3rd Intl. ECSA Citizen Science Conference;* 2020; Online.

[21] Etzold, H., Noack, S. & Jurk, A. (2019). *Algorithmen im Alltag. Leitfaden für Lehrerinnen und Lehrer. Teil 1: Hintergrund und Theorie. Digitales Lernen Grundschule, Universität Potsdam.* Abgerufen von https://dlgs.uni-potsdam.de/sites/default/files/u3/Leitfaden_Algorithmen_Teil_1.pdf

[22] Lewandowski, D. *Suchmaschinen verstehen.* Berlin/Heidelberg: Springer Vieweg; 2021.

[23] Nagulendra S, Vassileva J. Understanding and controlling the filter bubble through interactive visualization: a user study. *Proceedings of the 25th ACM conference on Hypertext and social media*; 2014 September; Santiago, Chile. New York: Association for Computing Machinery.

[24] OECD. *PISA 2018 Assessment and Analytical Framework.* Paris: OECD Publishing; 2019.

[25] Selwyn N, Pangrazio L. Doing data differently? Developing personal data tactics and strategies amongst young mobile media users. *Big Data & Society* 2018; 5(1): 1–12.

[26] Cook-Sather A, Bovill C, Felten P. *Engaging Students as Partners in Learning and Teaching: A Guide for Faculty.* San Francisco: Jossey Bass; 2014.

[27] Eisenberg, M. B. & Berkowitz, R. E. *Information problem-solving: the big six skills approach*; 2003.

https://doi.org/10.5281/zenodo.8066392

[28] Prediger S, Link M, Hinz R, Hußmann S, Thiele J, Ralle B. Lehr-Lernprozesse initiieren und erforschen – Fachdidaktische Entwicklungsforschung im Dortmunder Modell. *MNU* 2012; 65(8): 452–457.

[29] Peffers K, Tuunanen T, Gengler C E, Rossi M, Hui W, Virtanen V et al. The design science research process: a model for producing and presenting information systems research. In *Proceedings of the 1st international conference on design science research in information systems and technology* 2006; 83–106.

[30] Platz M. „Forscher spielen" und mathematisches Beweisen in der Primarstufe. *transfer Forschung - Schule* 2020; 6: 30–43.

https://doi.org/10.5281/zenodo.8349455

OPENING THE PANDORA'S (BLACK) BOX OF AI MADE IN EUROPE : INTERPRETABILITY - EXPLAINABILITY - COMPREHENSIBILITY

Christophe Denis, Sorbonne University, LIP6 [75252] Paris, also at ERIAC, [76821], Mont-Saint-Agnan

Anaëlle Martin, CEIE, University of Strasbourg [67046] Strasbourg, France

Abstract

In the already extensive literature connecting « black-box effect » and explainability, it is well established that some *deep learning methods* (Dl), although successful from the accuracy point of view, are opaque in terms of understanding how they make decisions. While the lack of explicability of *machine learning* (ML) techniques raises operational, ethical and legal problems, the technical and political solutions provided by researchers and policy-makers to make algorithms more « transparent » may also appear problematic in terms of clarity. We argue that there is a strong paradox in listing numerous requirements that are conceptually indeterminate[1], in order to address the challenges posed by « black box algorithms ». We suggest that before any prescriptive or normative ambition in the field of AI ethics and regulation, some epistemological prolegomena are required to distinguish the epistemic functions and uses of descriptive, predictive and explicative models.

PROBLEM STATEMENT

"*In the modern system it should appear as though everything were explained"*, Wittgenstein, Tractatus Logico-Philosophicus (6.372)

"*Our civilisation is characterized by the word progress. (. . .) Its activity is to construct a more and more complicated structure. And even clarity is only a means to this end and not an end in itself. For me on the contrary clarity, transparency, is an end in itself. I am not interested in erecting a building but in having the foundations of possible buildings transparently before me. "* Wittgenstein (MS 109 200: 5.11.1930)

External properties like explainability and fairness are commonly associated to deep neural network to try to overcome its "black box" effects. It is therefore necessary to clarify the particularity of the black box associated with deep learning, which is not always negatively connoted. For example, in software engineering, the use of black boxes facilitates maintenance and reduces the code programming time. The concept of black box is not the prerogative of deep learning as it is used in science and software engineering. In everyday life, an institution can also be considered as a black box since we do not know its internal functioning, for example when we ask for an administrative act. From a functional point of view, a black box, represented by Figure 1, compute output values from inputs ones without any knowledge of is internal workings.

Input
Black Box
Output

Figure 1: Functional point of view of a black box

[1] Indeed, it is not always clear if the rules that are mentioned are legal or ethical and if the concepts refer to binding principles or « values ».

https://doi.org/10.5281/zenodo.8349455

TRANSPARENCY AND ITS COROLLARIES AS SOLUTIONS TO THE "BLACK BOX" PROBLEM: TRACEABILITY, AUDITABILITY AND COMMUNICATION

Ethical principles & requirements set out by the Guidelines for Trustworthy AI

The aim of the Guidelines is to promote Trustworthy AI (lawful, ethical and robust) and provide guidance on how such principles can be operationalized in socio-technical systems.

The ethical principles are as follows:

1. respect for human autonomy;

2. prevention of harm;

3. fairness;

4. explicability.

The ethical requirements are as follows:

1. human agency and oversight;

2. technical robustness and safety;

3. privacy and data governance;

4. transparency;

5. diversity, non-discrimination and fairness;

6. environmental and societal well-being;

7. accountability.

Explicability is viewed as a « principle » — closely linked with the rights relating to Justice — while transparency is considered as a « requirement ». The latter encompasses transparency of elements relevant to an AI system: data, system and business model. It includes traceability, explainability and communication. The principle of explicability means that « processes need to be transparent, the capabilities and purpose of AI systems openly communicated, and decisions – to the extent possible – explainable to those directly and indirectly affected ». Traceability involves that data sets and processes should be documented to allow for an increase in transparency. This enables identification of the reasons why an AI-decision was erroneous. Communication implies that AI systems should not represent themselves as humans to users. Beyond this, the AI system’s capabilities and limitations should be communicated. Explainability requires that whenever AI has a significant impact on people’s lives, it should be possible to demand a suitable explanation of the AI’s decision-making process. It should be timely and adapted to the expertise of the stakeholder.

Legal obligations laid down by the AI Act

In its preamble, the *AI Act* asserts that the proposal lays down obligation that will apply to providers and users of high-risk AI systems. It promotes public trust in the use of AI by facilitating audits of the AI systems with new requirements for documentation, traceability and transparency. In accordance with Article 13, "transparency and provision of information to users" are required for high-risk AI systems. High-risk systems should be designed in such a way "to ensure that their operation is sufficiently transparent to enable users to interpret the system’s output and use it appropriately".

https://doi.org/10.5281/zenodo.8349455

The proposal calls for « an appropriate type and degree of transparency » and provision of « concise, complete, correct and clear information that is relevant, accessible and comprehensible to users ». The proposal states that « to address the opacity that may make certain AI systems incomprehensible to or too complex for natural persons, a certain degree of transparency should be required for high-risk AI systems. Users should be able to interpret the system output. It is also recalled that the exercise of important procedural fundamental rights could be hampered, « where such AI systems are not sufficiently transparent, explainable and documented ». In case of fundamental rights violation, effective redress will be made possible « by ensuring transparency and traceability of the AI systems ». According to Article 12, high-risk AI systems should be designed with capabilities enabling the automatic recording of events while the systems is operating.

THE NEED FOR CONCEPTUAL CLARIFICATIONS

After having proposed a clarification and an adaptation of the notions of interpretability and explainability such as one encounters them in the already abundant literature on the subject, we recall in this article the interest of implementing the epistemological distinctions between the *different epistemic functions* of a model, and between the *epistemic function* and the *use* of a model[2]. We argue that systematically explaining *deep learning* to all its users is not always justified, could be counterproductive and even raises ethical issues. For example, how to assess the correctness of an explanation that could even be unintentionally permissive or even manipulative in a fraudulent context? There is therefore a need to revisit the theory of information (Fisher, Shannon) and the philosophy of information (Floridi) in the light of *deep learning*. This information will allow certain users to produce their own reasoning (surely an abductive one) rather than receiving an explanation. Last but not least, should we trust a *machine learning* model ? Trust means handing over something valuable to someone, relying on them. The corollary is that "the person who trusts is immediately in a state of vulnerability and dependence", and all the more so on the basis of an explanation whose correctness is difficult to assess. We believe that using human relationship terms, like trust or fairness in the context of machine learning, necessarily induces anthropomorphism, whose bad effects could be addiction (Eliza effect) and persuasion rather than information. In contrast, our philosophical and mathematical research direction tries to define conviviality criteria in machine learning based on Ivan Illich's thought.

According to Illich, a convivial tool must have the following properties:

- it must generate efficiency without degrading personal autonomy;
- it must create neither slave nor master;
- it must widen the personal radius of action.

[2] C. Denis; F. Varenne. Interprétabilité et explicabilité de phénomènes prédits par de l'apprentissage machine. ROIA, Volume 3 (2022) no. 3-4, pp. 287-310.

https://doi.org/10.5281/zenodo.8349455

As presented in the last part of the talk, neural differential equations, by providing trajectories rather than predictions, seem to be an efficient mathematical formalism to implement convivial deep learning tools.

Clarification of the notion of explainability

According to the High-Level Expert Group on AI, « explainability » is contained in the requirement of transparency, as well as « traceability » and « communication »[3]. These concepts are described in the guidelines as « other explicability measures » (e.g. traceability, auditability and transparent communication on system capabilities) and are required when an explanation as to why a model has generated a particular output is not always possible. These cases are referred to as « black box algorithms ». Explainability which seems to be a kind of « sub-requirement » concerns the ability to explain « both the technical processes of an AI system and the related human decisions »[4]. As for « explicability », it is an « ethical principle » that is, according to the High-Level Expert Group, « crucial for building and maintaining users' trust in AI systems ».

It appears that the concepts of explicability and explainability, although terminologically distinct, are not easily distinguished. The experts refer to both explicability and explainability, which is quite confusing given that the doctrine debates the status and definition of these notions. According to Floridi, for example, explicability is a richer notion than explainability, as it is combining demands for intelligibility and accountability[5]. As a result, it enables both people working with and those affected by AI systems to understand and challenge outcomes. Herzog on his part considers that explicability means both more and less than explainability understood in a sense that considers only mechanistic explanations. It means more because explicability demands explanatory interfaces tailored to the recipient and use-case that focus on putting the respective stakeholders in a position to take responsibility. It means less because the principle remains flexible enough to not strictly demand mechanistic explanations at every level of usage and may even allow for none if not available[6]. Others challenge the very existence of a principle of explicability, regardless of the term used: explainability, transparency, understandability.

We suggest to apply here a « grammatical investigation » in the sense that Wittgenstein understood it. In his time, Wittgenstein denounced in a radical way the

[3] Ethics Guidelines for Trustworthy AI, p. 14.

[4] Trade-offs might have to be made between enhancing a system's explainability (which may reduce its accuracy) or increasing its accuracy (at the cost of explainability).

[5] Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., et al.: AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recom- mendations. Minds Mach. 28, 689–707 (2018). https://doi.org/10. 1007/s11023-018-9482-5

[6] Herzog, C. On the risk of confusing interpretability with explicability. *AI Ethics* **2**, 219–225 (2022). https://doi.org/10.1007/s43681-021-00121-9

https://doi.org/10.5281/zenodo.8349455

syncretism which reigned in science, in ethics (and philosophy) and in epistemology. According to the Viennese philosopher, this attitude encouraged scholars to provide justifications and explanations for phenomena that science had difficulty to understand (what he called the « mythological power of explanation »). To counter this metaphysical drift, Wittgenstein developed a « grammatical-therapeutic philosophy ». Taking into account the wittgensteinian epistemology, we will adopt the definitions proposed by C. Denis and F. Varenne to define the notions of interpretability[7], explainability[8] and understandability[9] (comprehensibility) and to determine their mutual relationship. We claim that before enshrining a principle of explicability, whether at the ethical or legal level, or both, a clear (and unambiguous) definition of the concept must be provided. Making AI explainable must be an epistemic requirement before a moral obligation and a binding principle (principle of explicability).

In the field of AI, the epistemological question is fundamental because the validation of a « black box » differs from that of the mathematical and causal modeling of a physical phenomenon. Indeed, contrary to the previous one, machine learning methods do not pretend to represent a causality between the input and output parameters, despite the use of misleading terms, derived from the statistical theory : the so-called « explanatory » variables[10].

The explanation could also be considered as a « colossus with feet of clay » on the methodological level. Indeed, learning methods are often used when it is difficult or impossible to define the functional specifications of a process. In particular, one interprets the question posed to the machine learning algorithm, and one then wishes to obtain an explanation of the prediction obtained on a question which is not the one solved by the algorithm. There is a famous example commonly used to underline the need for explainability. Suppose you want to implement an algorithm that detects on an image that an animal is a wolf or a husky. The machine learning method uses half images of wolves or husky. The results obtained are spectacular until the day a wolf without a snowy backdrop is detected as a husky. A more in-depth study provides the following explanation: "The learning algorithm did not "learn" to recognize the

[7] Interpretability of a model: set of symbols having the property of being composed of elements (signs, figures, concepts, data, etc.) that each have a meaning for a human subject. A model is interpretable when all its symbols are interpretable

[8] The explainability of a model is the ability to deploy and explain the outputs of the algorithm in a series of steps that are linked together by what a human being can meaningfully interpret as causes or reasons.

[9] The notion of understandability must be defined on the basis of the notion of interpretation (and not the opposite), and must be distinguished from the definition of explanation. There is comprehension of a phenomenon when the human being has the possibility of grasping the whole of it and of unifying its successive manifestations under a single representation easy to conceive and to recall.

[10] Christophe Denis, Franck Varenne. Interprétabilité et explicabilité de phénomènes prédits par de l'apprentissage machine. Revue Ouverte d'Intelligence Artificielle, Association pour la diffusion de la recherche francophone en intelligence artificielle, 2022, 3 (3-4), pp.287-310. ff10.5802/roia.32ff. ffhal03640181

https://doi.org/10.5281/zenodo.8349455

difference between a wolf and a husky but to recognize the presence or absence of snow on the image".

As a corollary, the first reflex is to blame the machine learning method, which deceived us. On the contrary, we believe that the algorithm has done its job correctly, that is to say finding a robust criterion for distinguishing the images. It is therefore not a search for an explanation but the definition of the functional specifications of the black box without degrading its performance that one could not expect with a transparent model.

Last but not least, does a more transparent model facilitate access to the world of knowledge? We do not think so, on the contrary. Following the work of Herbert Simon and cyberneticians, intelligence is obtained by a feedback loop by acting on a black box taking into a simpler world representation. The deep neural network allows, without our yet knowing how to explain it mathematically, to find regularities and symmetries in the complexity of the world.

To conclude, we believe it is a little schizophrenic to think that "opening" the machine learning black boxes would permit us to access to knowledge about our physical world. The functional description, as fine as it is, of the Galilee's telescope does not make it possible for us to understand the ins and outs of the theory of heliocentrism. Our current research consists in placing deep neural networks in this appropriate scientific and epistemic paradigm, the cybernetic one, for which the notion of black box is an asset. Indeed, from signals measured in the visible domain, deep learning manages an informational structure that an human abductive-inductive reasoning allows;

https://doi.org/10.5281/zenodo.8070916

Five years of Open Search Initiative – where do we stand?

S. Voigt, C. Plote, C. Geminn and F. Hauser, Open Search Foundation, Starnberg, Germany

Abstract

The Open Search Initiative advocates for an independent, free and self-determined access to information on the internet. Its goal is a web search that is not profit-driven but based upon ethical values and fundamental rights - a web search ecosystem that offers manifold choices, including public as well as commercially driven services. In the year 2022 the Open Search Initiative is in its fifth year of existence and many scientific organisations, universities, computing centres, dedicated individuals, administrative bodies, foundations, NGOs, start-ups and companies are cooperating therein and are developing concepts, components, and general rules for the envisaged open search ecosystem. Their common driver is not profit, however, the goal of an internet search that benefits everyone.

The basic idea of the initiative is to build the web search ecosystem based on open and auditable source code; distributed crawling, indexing and hosting across many different computing centres and language spaces; as well as on public moderation and democratic oversight of the infrastructure. It is the aim to make web search a common good that is implemented and maintained by public and partially maybe also private actors, as it is the case with other relevant infrastructure such as roads, water supply or electricity. It will allow the building of a diversity of search frontends and will furthermore enable establishing many other open web analysis services, both, public/free of charge or commercially driven.

During the first years, the Open Search Initiatives focussed on spreading and jointly sharpening the concepts, as well as on building an interdisciplinary network of supporters, addressing the many different technical, organisational, legal, ethical and societal aspects of cooperative and open web search. After a first stakeholder workshop in 2017 and establishing the Open Search Foundation as non-profit organisation in 2018, a growing community built up during the past five years. Currently more than 120 individuals from more than 50 organisations, from 12, mainly European, countries are actively contributing to it. With several specialist groups cooperating in different disciplines and a number of events being organised for fundraising, community building, awareness raising and scientific exchange. The initiative has developed an sound momentum momentum already and the central event is the yearly International Open Search Symposium, with the first #ossym held in 2019 at Leibniz Supercomputing Center and in subsequent years at CERN.

It is clear, that web search is more than a technological challenge. Many aspects play a role, such as legal, ethical, economic, and educational disciplines. In order to address this diversity, six specialist working groups were establishing helping to structure the work, better coordinate the developments and improve the communication: Awareness, Application, Economy, Ethics, Legal and Tech.

A key activity is raising awareness for the open search idea with organisations, public bodies, politics as well as the general public. To this end, the initiative could already successfully inspire the European Commission to start supporting research on open search and it raised interest of, and is in dialogue with, members of the European and German parliament, it gained support by several federal and 'Länder' Ministries in Germany and liaised with many NGOs, start-ups, universities, foundations etc. Further more, several partner organisations established internal open search related projects and activities, to strengthen and coordinate their contribution (e.g. DLR, LRZ, CERN and several universities). Furthermore, teams of the the community could raise dedicated project support and funds from public donors and foundations, e.g. European Commission, Stiftung Mercator, BMW-Foundation and NLNet. These range from a few ten-thousand Euros for individual projects to several million Euro of research funds for the OpenWebSearch.EU EC Horizon Europe project.

Where do we stand with the Open Search Initiative? Reflecting the developments and the achievements described above, the Open Search Initiative can be considered as having left its initial phase. A substantial group of relevant organisations, science centres, universities, computing centres and concerned individuals are engaging and is cooperating. Awareness and concern about single platforms controlling access to information on the Internet is significantly raising. Through the work of different specialist groups within the initiative, key aspects and development tasks for the envisaged open web search ecosystem could already be identified and are being currently further developed.

Next important steps: With the basic technical, organisational and societal dimensions of an open search ecosystem being identified, it will be important during this next phase of the initiative to connect these different dimensions through: a) Intensifying the dialogue between the involved disciplines, b) Elaborating first concrete principles, rules of engagement and interfaces for operating the open search ecosystem, technically and organisationally and c) Implementing and operating a first substantial and lasting prototype, to demonstrate the full potential of open web search. In parallel to this, it will be important to bring open web search on the political agenda at European level as well as in Member States, to ensure concrete commitments in terms of funding and legislative support. Beyond this, it will important in the coming phase to develop the financing and operating models, as well as the governance framework for an open and cooperative web search infrastructure in Europe. It will be important to develop the technical, organisational, societal, financial and legislative aspects of this open

https://doi.org/10.5281/zenodo.8070916

search ecosystem in parallel and with close linkages and exchange. By doing so, it will be possible to implement and scale the open search idea at the necessary pace and to bring it to routine operation across Europe and for different language spaces and top-level domains. All this can only be achieved if a cooperative and team-play approach can be maintained. Together, for a better net.

https://doi.org/10.5281/zenodo.8075712

OPEN WEB SEARCH FOR AI AND NLP IN EUROPE

Jelena Mitrović, Chair of Data Science, University of Passau
Tomáš Mikolov, CIIRC, Czech Technical University
Michael Granitzer, Chair of Data Science, University of Passau

INTRODUCTION

Artificial Intelligence (AI) and Machine Learning (ML) are becoming an increasingly important part of today's businesses. Web search, spam filters and ads recommendation systems rely heavily on AI and ML. However, this "AI Industry" is currently mainly located in Silicon Valley, which creates an imbalanced situation in the research community: while funding, computational resources, and data are all plentiful and available for the researchers located in US and Canada, the situation in Europe is less stellar.

Natural Language Processing (NLP) is an important, and some might argue, the most popular area of AI research and industry nowadays. The goal of NLP is to computationally analyze natural language through simpler, almost accomplished tasks such as part of speech tagging, to complex tasks such as rhetorical figure detection, discourse analysis, or implicit hate speech detection. Most of these tasks are heavily dependant on the presence of well curated data sets, i.e. corpora.

The arguably biggest trend in NLP from the last decade is the adoption of models based on various artificial neural network architectures. Statistical language models (LMs) in particular were dominated for decades by n-gram techniques, to the point that the research community stopped believing it would be possible to significantly overcome these models on large datasets. This changed a decade ago when several important discoveries related to neural language models have been made – especially through open-sourcing these technologies via RNNLM [1] and word2vec [2] projects which were used by many researchers and engineers in companies such as Google, Facebook, and Microsoft as the starting point in using neural models for NLP tasks. While this paradigm shift has been started mainly by researchers from Europe, the neural language models were massively popularized by US companies in recent years.

FUTURE OF NEURAL LANGUAGE MODELS

Recently developed large LMs trained on publicly available data show impressive results. Interestingly, the pretraining dataset for LaMDA [4], e.g., consists of 2.97B documents, 1.12B dialogs, and 13.39B dialog utterances, for a total of 1.56T words. There is an imbalance in research potential at academic institutions because at the moment, access to large amounts of structured, curated data is a reality only for the big companies. When it comes to domain-specific LMs, e.g. LegalBERT [5] or HateBERT [6], it is possible to re-train them successfully on well-curated, comparably smaller corpora. Still, if we wish to bring the performance of these models on par with LaMDA, and have truly valuable specialized models, we should be able to employ frequent and proactive updates using the current Internet data, incorporate the relevant knowledge from social media, and be able to update the performance of our models accordingly - all of which requires a lot of computational power and resources. Furthermore, a proper ethical framework for using the new data for re-training should be at the centre of such efforts. It is not clear how these issues are handled with large LMs at the moment. If we want to build domain-specific language models that can deal with non-literal language, rhetorical figures, negation, humor, we need access to a vast amount of publicly available data, along with literary works, political debates, social media data. Current language models are only partly capable of truly dealing with these phenomena.

CONCLUSION

The chase for building bigger and better AI models is not and should not be seen as the only way forward in science, but the fact that European institutions are lagging behind due to the lack of resources, should not be a defining factor behind our success. Basic research is crucial for progress in AI and NLP, but we also need access to vast amounts of data to test our hypotheses and come to scientific breakthroughs.

It could be argued that to create balance in the current situation, Europe needs to become a leader in areas that are currently being dominated by US monopolies. An example is web search and social networks: while Google and Facebook generate hundreds of billions of dollars in revenue every year, the current business landscape seem to imply that most of the investments are aimed towards the US. This is harmful for the research progress - the scientists no longer compete over having the most innovative ideas, but rather win various benchmarks by using more computational resources to train larger and larger models. This is why having an Open web search paradigm and ultimately an EU-based organisation based on its principles would be an important step towards demonopolization of the AI industry and in creating fairer opportunities in science.

ACKNOWLEDGEMENTS

SPONSORED BY THE
Federal Ministry
of Education
and Research

The project on which this report is based was partly funded by the German Federal Ministry of Education and Research (BMBF) under the funding code 01|S20049. The author is responsible for the content of this publication.

https://doi.org/10.5281/zenodo.8075712

REFERENCES

[1] Mikolov et al., Recurrent neural network based language model, INTERSPEECH, 2010.

[2] Mikolov et al., Distributed Representations of Words and Phrases and their Compositionality, NIPS, 2013.

[3] Brown et al., Language Models are Few-Shot Learners, `https://arxiv.org/abs/2005.14165`

[4] Thoppilan et al., LaMDA: Language Models for Dialog Applications, `https://arxiv.org/abs/2201.08239`

[5] Chalkidis, Ilias et al., LEGAL-BERT: The Muppets straight out of Law School, ACL, `https://aclanthology.org/2020.findings-emnlp.261`

[6] Caselli et al., HateBERT: Retraining BERT for Abusive Language Detection in English,Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021).

https://doi.org/10.5281/zenodo.8075727

Maritime Data Platform for increasing data accessibility via a data space architecture

Jankowski, Dennis, German Aerospace Center (DLR e.V.), Oldenburg, Germany

Abstract

The amount and heterogeneity of data that needs to be managed is continuously increasing. As a result, it is becoming more and more difficult for people to find suitable data. In addition, the data is often not provided in a uniform format and is made available via a wide variety of different interfaces (e.g. via the web, in databases or CSV files). So how can a centralized and uniform search be realized that enables users to find and access suitable data despite the high degree of technical heterogeneity?

For this purpose, a Maritime Data Platform has been developed that enables unified data access to distributed data sources while considering maritime challenges such as volatile connectivity (c.f. Figure 1). In the following, the essential aspects that are in the focus of this work are described in more detail:

Data space architecture with Data Space Support Platform (DSSP) for access to distributed data sources while maintaining the sovereignty of the data providers - no physical migration of the data itself. Access to distributed data is controlled by so-called “connectors” that are connecting the data source to the platform. The connectors can be configured by the data provider so that he/she can decide which data should be exposed to which data consumer.

Handling volatile connections: Management for handling volatile connections and low bandwidth - Possibilities to tackling this are a cache for the intermediate storage of data during a connection interruption, a guarantee that messages are not only sent but have also arrived at the recipient (e.g. handshakes), and the reduction of the transmitted data volume through compression.

Monitoring of connected data sources: Continuous measurement of key figures that allow an assessment of the reliability or quality of a data source and the platform architecture itself - latency, downtime, transmission rate. If data sources show anomalies, such as excessive latency for a service or a dropped connection, a warning should also be displayed to the consumer.

Management of changing data providers: Responsibility for providing certain data (e.g., weather data) may vary from one geographic location to another. For a maritime user, however, it is (relatively) uninteresting who provides the data. The data consumers only need access to the data. If a vessel travels out of port and thereby the responsible data provider changes, the data supply should continue unchanged in the background for the data consumer (provided he has access to both data sources).

Secure authentication and authorization: Users can authenticate themselves within the platform and also authorize themselves to access the data provided for them. For this purpose, the link with the Maritime Connectivity Platform shall be established. Each user should be able to authenticate himself using his/her MCP credentials/certificate.

Encrypted communication: Since the transmitted data may be sensitive information, encrypted communication via the platform is relevant. The transmitted data must be encrypted at the connectors and forwarded to the recipient of the data. The recipient should then be able to decrypt the data again. Men-in-the-middle attacks should be made more difficult in this way.

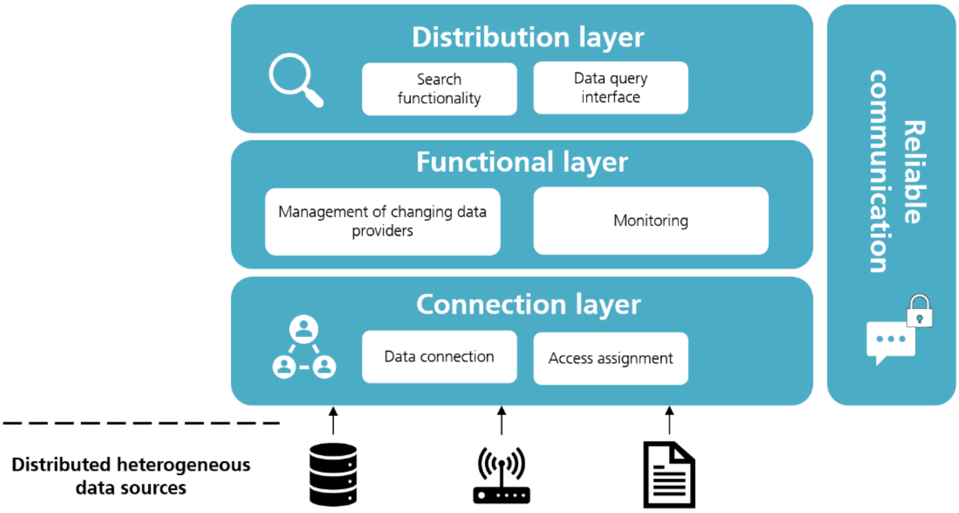


Figure 1: Data Space Architecture for accessing heterogenous data sources

https://doi.org/10.5281/zenodo.8075853

Open Search for Science - Science for Open Search

S. Voigt[†] and T. Hecking, Dennis Jankowski and Maximilian Schwinger,
German Aerospace Center,
Oberpfaffenhofen/Cologne/Oldenburg, Germany

Abstract

Science is about the discovery of new things, knowledge, concepts, contexts, systematics, relations etc. The more data and information we render and collect in the digital sphere and use as the basis for science, the more important it is to ensure open, unbiased and public access to it. Science has a vital role in and responsibility for this. It profits significantly from the ease of exchange of data and information and thus also has a responsibility in keeping it open, accessible, findable and curated.

Large scientific organisations consume large quantities of information and data from external sources into their research environments. They generate, store, manage and maintain large amounts of information in their intranet, research systems, data repositories and information corpora as well as they generate large volumes of scientific artefacts which they output to the public domain. The German Aerospace Centre, with a bit less than 10.000 scientific, technical and administrative staff has a vital interest in ensuring that access to digital information is unbiased, objective and as efficient as possible. Internally, within the organisation, as well as externally, in the web. The thus established ‘OpenSearch@DLR’ project is currently in its second year and in this talk we present first findings, approaches and lessons learnt from the project as well as we demonstrate pilot applications for scientific search.

During the course of the project we also found the following: Efficient searching and finding of information within the different repositories of the organisation is as important as searching for external resources on the web. In particular for scientists, finding and retrieval of research data is very important and needs better technical tools and support. This requires thematically pre-structured and pre-processed information on existing data and information to better search in internal repositories and corpora. Often enough, information needs to be found in heterogeneous artefacts (such as papers, software tools, manuscripts, etc.) which are often distributed over various different subsystems. Techniques for automatic content analysis have to be advanced to identify relationships between scattered pieces of information and data, which enable integrated search and discovery of scientific knowledge simultaneously and synergistically across different repositories and corpora: e.g. to search for authors, scientific concepts, tools or techniques. To this end, representing extracted information and resources as well as semantic relations between them in graph-databases seems to be a promising approach for this. Sophisticated search involves efficient visualisation and analysis tools, allowing deeper insight in to search results and going far beyond of what is offered in today's standard search tools. This includes for example that the availability and analysis capabilities of geospatial search features need to be significantly improved for all kinds of search and analysis tasks in the web. Along with this, time-stamps and versioning of web artefacts are equally important factors. Versions of knowledge artefacts have to be tracked in order to avoid inconsistencies and ensure that data of known provenance and up-to-date-ness are used for scientific analysis. Furthermore, there is a need for the identification of emerging topics and to anticipate future trends of scientific and technological or even societal developments.

All in all, it can be established that good internal search features, good internal knowledge management and thus, sophisticated intranet search is a critical infrastructure for science centres, which need better care and which by far should not be considered a by-product. In conclusion we argue that improving the intranet search features and linking them with sound public and distributed open web search infrastructures will substantially improve the scientific work and organizational efficiency of many large-scale science organizations. Thus, many major science and research organisation need to actively engage in the building and maintaining of openly searchable information repositories internally, as well as where possible, also externally, in cooperation with others and for searching the web as a whole.

https://doi.org/10.5281/zenodo.8083527

SEARCHING FOR EDUCATIONAL RESOURCES IN A STUDY ASSISTANT: AN OPEN SEARCH USE CASE

L. Martin*, F. Engl, T. Hirmer, M. Ochs, A. Henrich,
University of Bamberg, 96047 Bamberg, Germany

Abstract

In recent years we have developed a study planning assistant, for which the extension to a more general study assistant represents an interesting open search use case due to the subjects of search, i.e., (Open) Educational Resources. Considering this use case deepens the understanding of vertical providers' requirements for an Open Search Infrastructure.

STUDY (PLANNING) ASSISTANT

Currently, our study planning assistant supports the process of selecting modules as a student as well as the order in which to attend these. Furthermore, it provides an overview of compulsory and elective modules as well as the students‘ individual progression within their program. Our planned research aims to extend the capabilities of the current assistant to provide a more general assistant with vertical search features. The possible use case is to provide a (personalized) recommendation system for (open) learning material, which requires the integration of Educational Resources (ERs).

(OPEN) EDUCATIONAL RESOURCES

ERs can be defined as “learning, teaching and research materials in any format and medium” [1]. If those resources “reside in the public domain or are [...] released under an open license” [1], they are called Open Educational Resources (OERs). However, OERs often lack findability [2], which is addressed by various OER search projects. Known representatives for vertical search over OER are *OER Commons*[1], targeting various educational levels, and Open Educational Resources Search Index (OERSI)[2], focused on higher education resources. Apart from vertical searches, projects such as *OERhörnchen*[3] execute search queries via Google that are specially configured for OER. Our approach aligns with these projects‘ common goals of making curated OER more accessible, yet it allows for direct access of educational material not only within the domain of education, but within the context of a study (planning) assistant application by means of an Open Search Infrastructure (OSI).

OPEN SEARCH INFRASTRUCTURE

An OSI, as envisioned by the contributors of the OSSYM, provides an interface for accessing a decentralized pool of information sources such as crawls, databases, and indices. A central requirement for implementing the study assistant using the OSI is an easy integration of ERs or ER providers to the decentralized pool. For this reason, the said integration and management of external information sources in the OSI has to be a benefit for vertical providers and external information source providers as well because most ERs are using laboriously curated indices. Possible ways of incorporating external information sources are integrating the data directly into the OSI, adding the external information sources to the first-hand crawls of the OSI, or leveraging meta searches, i.e., relaying the search to external search providers and post-processing the search engine result pages.

PLANNED RESEARCH

As noted before, ERs are often structured, highly curated, domain specific, and tailored to the users, e.g., regarding previous knowledge. Therefore, we plan to extend the existing study planning assistant with search engine and recommender system features, e.g., matching anonymized user profiles with an ER knowledge graph to identify possible resources of interest for a user's current information need. To establish a ranking of relevant ERs, voting techniques for calculating relative scores [3] are to be applied where previously visited ERs vote for candidates retrieved from the ER knowledge graph. For instance, if computer science students search for ERs with regards to personal preference and previous knowledge, our study assistant is to provide a ranking of ERs that are strongly related to their user profile. This planned research complements our other project investigating the requirements for the interface of an OSI in the context of a company search [4].

REFERENCES

[1] UNESCO. “Recommendation on open educational resources (oer).” (2019), `http://portal.unesco.org/en/ev.php-URL_ID=49556&URL_DO=DO_TOPIC&URL_SECTION=201.html`

[2] M. D. et al., “Whitepaper Open Educational Resources (OER) in Weiterbildung/Erwachsenenbildung. Bestandsaufnahme und Potenziale 2015,” 2015. `https://open-educational-resources.de/wp-content/uploads/Whitepaper-OER-Weiterbildung-2015.pdf`

[3] A. Henrich and M. Wegmann, “Search and evaluation methods for class level information retrieval: Extended use and evaluation of methods applied in expertise retrieval,” in *SAC '21: Proceedings of the 36th Annual ACM Symposium on Applied Computing*, New York, NY, USA: ACM, 2021, pp. 681–684. DOI: `10.1145/3412841.3442092`.

[4] L. Martin, F. Engl, and A. Henrich, “Requirements for an Open Search Infrastructure from the Perspective of a Vertical Provider,” Zenodo, Oct. 2021. `https://doi.org/10.5281/zenodo.6139647`

[*] leon.martin@uni-bamberg.de

[1] *OER commons*. `https://www.oercommons.org` [03.06.2022].

[2] *OERSI*. `https://oersi.org` [03.06.2022].

[3] *OERhörnchen*. `https://oerhoernchen.de` [03.06.2022].

ISBN: 978-92-9083-646-9 Open Search Symposium 2022 - #ossym2022 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.8309761

WARC-DL: SCALABLE WEB ARCHIVE PROCESSING FOR DEEP LEARNING

Niklas Deckers Martin Potthast
Leipzig University

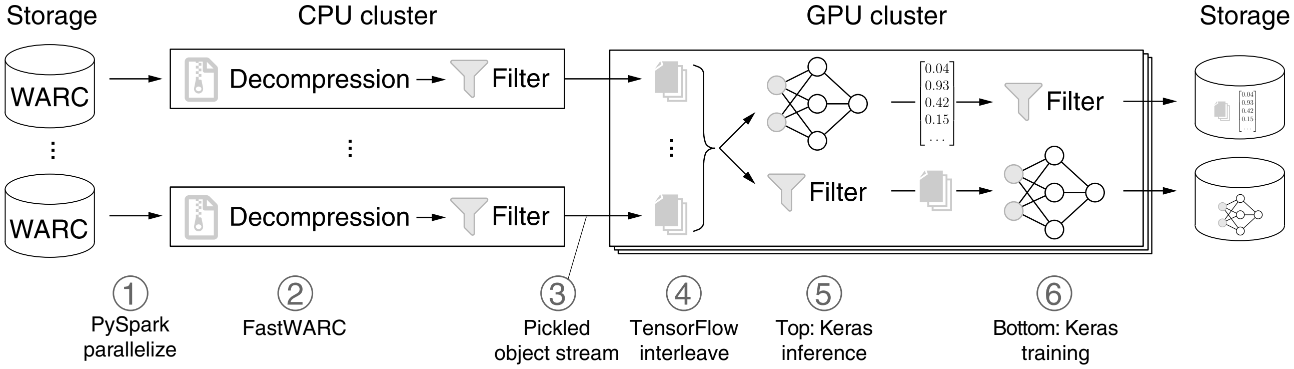


Figure 1: The six steps of the WARC-DL web archive processing pipeline: Filters and the Keras training are customizable.

Introduction. Web archives have grown to petabytes. In addition to providing invaluable background knowledge on many social and cultural developments over the last 30 years, they also provide vast amounts of training data for machine learning. To benefit from recent developments in Deep Learning, the use of web archives requires a scalable solution for their processing that supports inference with and training of neural networks. To date, there is no publicly available library for processing web archives in this way, and some existing applications use workarounds [1]. This paper presents WARC-DL,[1] a deep learning-enabled pipeline for web archive processing that scales to petabytes.

Technical Setting. In many "traditional" data centers, mass storage, CPU-bound processing, and GPU-bound processing are separate clusters. Since the storage capacity of the latter two does not usually match that of the former, large web archives are traditionally processed in batches. However, batch processing of web archives on GPU clusters is wasteful: since web archive data is raw data, it must be preprocessed before being passed to a GPU. This preprocessing is usually CPU-bound and highly parallelizable, so using a CPU cluster for this is desirable before streaming the preprocessed data to the GPU cluster. In addition, the data relevant to a particular task is usually sparse across the archive (e.g., only certain images are needed for the training of image representations, and only certain plaintexts are needed for argument mining), resulting in a variable data flow after preprocessing. To optimize GPU usage, a constant flow of data is required. WARC-DL solves both problems: After loading web archive data from the memory cluster into the CPU cluster for preprocessing, it streams the useful data into the GPU cluster for interleaved processing.

Web Archive Processing Pipeline. Figure 1 provides an overview of the six-step pipeline implemented by WARC-DL: (1) WARC files[2] are distributed to the CPU workers using PySpark. (2) FastWARC [2][3] is used to decompress and iterate the records. The records can optionally be filtered and CPU-bound preprocessing like feature extraction or tokenization can be performed. (3) Pickled record streams are sent to the GPU Cluster via TCP. (4) The streams are converted into TensorFlow datasets and interleaved. (5) One option is to use a pre-trained Keras model to batch process the samples. The results are filtered and stored, including the original data extracted from the WARC datasets. (6) Alternatively, the samples can be used to train a Keras model after an optional filtering step, e.g., for duplicate removal. The filtering in (2) and (6) implements a basic MapReduce.

The preprocessing steps and the pre-trained models used are fully customizable (an extension to frameworks other than TensorFlow is planned). The pipeline enables extraction of data from multiple modalities, including text for language models and images for computer vision models. It also supports the simultaneous extraction of linked web archive records, such as the text of a web page and the image that was originally linked on the page. This should enable efficient multimodal learning, where the pipeline transparently solves the problem of matching an image with its associated text. To optimize the ratio of allocated CPU to GPU resources depending on the model, a profiling method based on the TensorFlow Profiler is provided.

REFERENCES

[1] H. Yang, L. Liu, I. Milligan, N. Ruest, and J. Lin, "Scalable content-based analysis of images in web archives with tensorflow and the archives unleashed toolkit," in *JCDL 2019*.

[2] J. Bevendorff, M. Potthast, and B. Stein, "Fastwarc: Optimizing large-scale web archive analytics," in *OSSYM 2021*.

Contact: <firstname>.<lastname>@uni-leipzig.de

This work was supported by the German Federal Ministry of Education and Research (BMBF, 01IS18026B) by funding the competence center for Big Data and AI "ScaDS.AI" Dresden/Leipzig.

[1] https://github.com/chatnoir-eu/chatnoir-warc-dl

[2] https://www.iso.org/standard/68004.html

[3] https://github.com/chatnoir-eu/chatnoir-resiliparse

ISBN: 978-92-9083-646-9

ISSN: 2957-4935

https://doi.org/10.5281/zenodo.8349307

CLASSIFYING ADULT CONTENT USING NAÏVE BAYES

O. Behrendt*, Munich, Germany

Abstract

In today's World Wide Web any kind of content is always at everyone's fingertip. But not every content meets ethical, cultural or legal norms. In particular, regular use of online pornography is associated with sexually aggresive behaviour. Since web search engines are pivotal as an entry point to the web, they need to address this issue. This article presents a naïve Bayes classifier for detecting adult web sites in context of a web crawler. The classifier is solely based on textual content of home pages. The approach is fast and easy to implement and experiments show high accuracy.

INTRODUCTION

Since frequent use of violent pornography by adolescent is suspected to increase sexually coersive and aggressive behavior [1], it appears ethically indicated to filter adult content. Moreover, ignoring cultural or legal norms pose reputational and financial risk to search engine owners.

Many reported filtering techniques try to detect adult media content, others are based on textual content and some use a combination of techniques, see [2] for a more thorough discussion. The presented naïve Bayes classifier, trained on the text of home pages, allows a web crawler to apply a simple and effective strategy: Whenever an unseen domain is met, we first fetch the home page for classification.

NAÏVE BAYES CLASSIFICATION

Following [3] a *multinomial naïve Bayes* classifier estimates the class $\hat{c}(d)$ of a document $d = (w_i)_{i=1}^n$ ("bag-of-words" representation) by maximizing the posterior $P(c|d)$:

$$\begin{aligned}\hat{c}(d) &= \arg\max_{c\in C} P(c|d) = \arg\max_{c\in C} \frac{P(c)P(d|c)}{P(d)} \\ &= \arg\max_{c\in C} P(c)\textstyle\prod_{i=1}^n P(w_i|c) \qquad (1)\end{aligned}$$

The prior $P(c)$ and conditional probabilities $P(w_i|c)$ are estimated as relative frequencies:

$$P(c) = \frac{N_c}{N}, \quad P(w_i|c) = \frac{f(w_{i,c}) + 1}{\sum_{w\in V}(f(w_c) + 1)},$$

where N_c, N is the number of documents in c resp. in total and $f(w_{i,c})$, $f(w_c)$ denote token counts in class c. Note the Laplace-smoothing '+1' to avoid zeros in equation (1).

EXPERIMENTAL SETUP

Training data consists of 217 adult and 4351 non-adult home pages. For each retrieved home page all HTML tags were removed and the remaining text added to the vocabulary. Words that occured only once or in more than 50% of the pages were removed. The remaining 4568 pages were randomly split into ten folds for cross-validation. Each validation run builds a vocabulary from the nine training folds and classifies the pages in the test fold using equation (1).

EXPERIMENTAL RESULTS

The results in Tab. 1 show a high accuracy of $\frac{210+4322}{4568}$ = 99.2%. Note that the precision depends on the ratio of adult and non-adult pages in the training set, as shown in Fig. 1. Bearing this in mind, our results are very competitive with results from Largillier et al. [2], who report an accuracy of 97.2% (precision 98.3%) for 839 adult and 314 "safe" pages.

		true values		
		adult	¬adult	**precision**
out	adult	210	29	$\frac{210}{210+29}$=87.9%
	¬adult	7	4322	$\frac{4322}{4322+7}$=99.8%
	recall	$\frac{210}{210+7}$=96.8%	$\frac{4322}{4322+29}$=99.3%	

Table 1: Confusion matrix for 10-fold cross-validation including derived micro-average precision and recall.

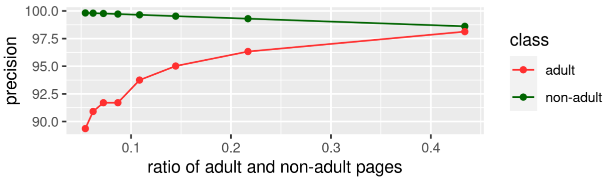


Figure 1: Precision versus ratio of classes in training set.

CONCLUSION

To further improve accuracy and especially the unwanted misclassification of non-adult pages as adult pages, a number of improvements could be explored in future work. One straightforward extension that comes to mind, is not only to analyse home pages, but additionally a fixed number of linked pages for each domain.

REFERENCES

[1] N. Stanley *et al.*, "Pornography, sexual coercion and abuse and sexting in young people's intimate relationships: A european study," *Journal of interpersonal violence*, vol. 33, no. 19, pp. 2919–2944, 2018.

[2] T. Largillier, G. Peyronnet, and S. Peyronnet, "Efficient filtering of adult content using textual information," *arXiv preprint arXiv:1512.00198*, 2015.

[3] H. Schütze, C. D. Manning, and P. Raghavan, *Introduction to information retrieval*. Cambridge University Press Cambridge, 2008, vol. 39.

* olaf.behrendt@neurolab.de

https://doi.org/10.5281/zenodo.8349341

STRATEGY COMPARISON FOR SEMANTIC ZERO-SHOT TAXONOMY FILTERS

A. Hamm*, German Aerospace Center (DLR), Cologne, Germany

Abstract

This contribution extends and tests ideas from sentence-transformer-based zero-shot text classification to the problem of building a taxonomy filter which can be used for assigning large quantities of documents to user-defined thematic groups.

EXTENDED ABSTRACT

Motivation

In information retrieval, categorised filtering is a way of supporting users in formulating their information needs in an efficient way: Instead of having to provide an explicit specification of the search request, the user can select from an existing list of available categories or tags.

Many collections of documents provided by publishers, libraries, or archives are structured in terms of subject-specific categories that can be used within their domains. Traditionally, assigning documents to categories is an effortful manual process. Progress in machine learning classification algorithms has made it possible to automatize this task in a generally acceptable manner, provided a sufficient number of labelled example documents from all categories is put into the training process.

The latter requirement, however, is a serious obstacle for a flexible use over a broad range of domains and in areas with limited amount of training data available.

It is therefore attractive to explore how the recently proposed method of transformer-based zero-shot text classification [1] can be applied to building taxonomy filters.

Objective

The aim of this contribution is to suggest and compare methods with the following characteristics, which will be explicated below:

1. The method should work with any user-provided commented taxonomic category system.
2. The method should not require taxonomy-specific training.
3. Time-consuming pre-processing steps for each document should not depend on the individual taxonomy categories.

A taxonomy for documents is a hierarchical tree-like system of categories (groups of documents) which covers a domain of interest. Taxonomies are abundant in scientific, economic, and normative classification schemes. Formalised, they are part of the W3C-recommended Simple Knowledge Organization System (SKOS). A commented taxonomy adds a short description to each taxonomy label.

If users want to set up specialised taxonomies for their purposes, it is much easier to provide a short description than to collect a sufficiently large set of labelled examples. General purpose language models which are able to detect semantic similarity can then be used to match documents to taxonomic descriptions.

However, if this matching for *N* documents and *M* categories requires the encoding of *NM* text sequence pairs it becomes inefficient. For large scale applications it is therefore important to employ the bi-encoder strategy of sentence transformers [2] which needs just *N*+*M* encodings – individually on documents and on category descriptions.

Method

A simple baseline implementation of a taxonomy filter obeying the first two objectives can be easily realised by directly using the *zero-shot classification pipeline* of *Hugging Face* [3]. The third objective can be achieved by confining to its *sentence similarity pipeline*.

This contribution reports on ongoing work regarding the refinement of taxonomy filters through category assignments which go beyond a gross similarity score by

a) Breaking down the documents into single sentences and computing a weighted aggregated category vote,
b) Taking into account hierarchical consistency as a criterion for category assignment.

Various variants of such taxonomy filters will be tested and compared on the basis of datasets and taxonomies from different domains.

These examples also show the degree to which approximate nearest neighbourhood computations in the underlying embedding space can replace exact calculations as a further performance improvement for large scale applications.

REFERENCES

[1] W. Yin, J. Hay, and D. Roth. “Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach”, in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, Nov. 2019, pp. 3914–3923. doi:10.18653/v1/D19-1404

[2] N. Reimers and I. Gurevych. “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks”, in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, Nov. 2019, pp. 3982–3992. doi:10.18653/v1/D19-1410

[3] https://github.com/huggingface/transformers

[*] Email: andreas.hamm@dlr.de

https://doi.org/10.5281/zenodo.8349371

THE ROBOTS.TXT STANDARD – IMPLEMENTATIONS AND USAGE

S. Nagel*

Extended Abstract

The robots.txt standard allows web masters to signalize operators of web crawlers how to best crawl their sites. The standard was initially proposed in 1994 [1] and is implemented in the following way: A file `robots.txt` is deployed in the root folder of a web site (eg. `http://example.org/robots.txt`). The text file is readable for web crawlers and contains policies how the crawlers ("robots") shall access the site's content. Access policies can be given for individual crawlers by "user-agent" name or by a wildcard rule block catching all crawlers not addressed in a named policy.

As a convention based on consensus not a legally binding regulation – the robots.txt standard nevertheless was adapted by all major web search engines and was extended in multiple ways to be able to specify

- the desired delay between successive requests to the web site
- more fine-grained access rules
- or rules for URL canonicalization
- the location of a sitemap [2] which allows to communicate to the crawler the exhaustive list of pages, images, videos, news pages while avoiding duplicates

The various extensions lead to differences how search engine crawlers implement the robots.txt standard. To further formalize the standard, in 2019 researchers at Google [3] submitted a RFC proposal [4] accompanied by a example robots.txt parser implementation [5].

The talk oversights the robots.txt standard as such and the varying extensions introduced over time. Starting with the perspective of a web crawler, we first look into actual implementations (eg. [5], [6]) of a robots.txt parser and difficulties matching rules and URLs.

We then analyze at the usage of the robots.txt in the web by web masters and site owners: which features or extensions are actually used, which crawlers are addressed as individual user-agents and is there a bias in favor of some search engine crawlers? We'll use six years of archived robots.txt files from Common Crawl for our analysis and also compare these results with earlier work (eg. [7], [8], [9], [10], [11]).

We will also showcase examples how a polite crawler respecting the robots.txt standard can use the robots rules to optimize the implementation of URL queues.

* sebastian@commoncrawl.org

REFERENCES

[1] M. Koster, *A standard for robot exclusion*, 1995. `https://www.robotstxt.org/`

[2] *Sitemaps.org*. `https://www.sitemaps.org/protocol.html`

[3] H. Zeller, L. Sassman, and G. Illyes, *Formalizing the robots exclusion protocol specification*, 2019. `https://developers.google.com/search/blog/2019/07/rep-id`

[4] M. Koster, G. Illyes, H. Zeller, and L. Sassman, *Robots exclusion protocol*, 2019. `https://datatracker.ietf.org/doc/draft-koster-rep/`

[5] *Google robots.txt parser and matcher library*. `https://github.com/google/robotstxt`

[6] *Crawler-commons*. `https://github.com/crawler-commons/crawler-commons/`

[7] Y. Sun, Z. Zhuang, I. Councill, and C. Giles, "Determining bias to search engines from robots.txt," in *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, WI 2007*, 2007, pp. 149–155. DOI: `10.1109/WI.2007.98`.

[8] Y. Sun, Z. Zhuang, and C. L. Giles, "A large-scale study of robots.txt," in *Proceedings of the 16th international conference on World Wide Web*, 2007, pp. 1123–1124.

[9] S. Kolay, P. D'Alberto, A. Dasdan, and A. Bhattacharjee, "A larger scale study of robots.txt," in *Proceedings of the 17th international conference on World Wide Web*, 2008, pp. 1171–1172.

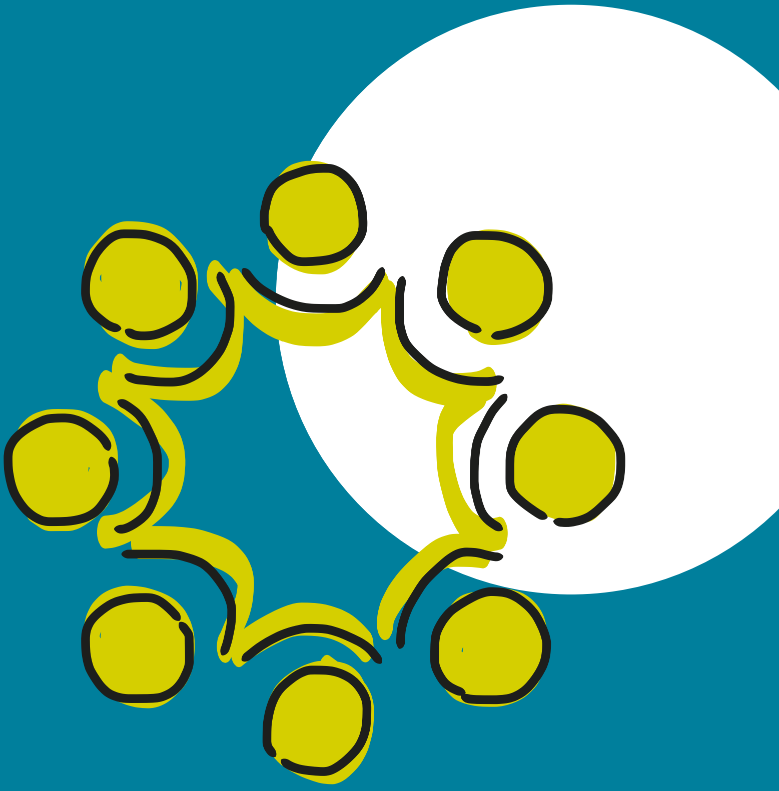
[10] U. Schonfeld and N. Shivakumar, "Sitemaps: Above and beyond the crawl of duty," in *Proceedings of the 18th international conference on World wide web*, 2009, pp. 991–1000.

[11] *Knuckleheads' club – the evidence we've found so far*, 2020. `https://knuckleheads.club/the-evidence-we-found-so-far/`

[12] M. Schellekens, "Are internet robots adequately regulated?" *Computer Law & Security Review*, vol. 29, no. 6, pp. 666–675, 2013. DOI: `https://doi.org/10.1016/j.clsr.2013.09.003`. `https://www.sciencedirect.com/science/article/pii/S0267364913001659`

Appendix

List of Autors



ISSN: 2957-4935
ISBN: 978-92-9083-646-9