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/

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

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,

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/`

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.

[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.