SEARCH IN A REDECENTRALISED WEB

Thanassis Tiropanis[1], Alexandra Poulovassilis[2], Adriane Chapman[1], and George Roussos[2]

[1]School of Electronics and Computer Science, University of Southampton, UK

[2]Department of Computer Science and Information Systems, Birkbeck, University of London, UK

ABSTRACT

Search has been central to the development of the Web, enabling increasing engagement by a growing number of users. Proposals for the redecentalisation of the Web such as SOLID aim to give individuals sovereignty over their data by means of personal online datastores (pods). However, it is not clear whether search utilities that we currently take for granted would work efficiently in a redecentralised Web. In this paper we discuss the challenges of supporting distributed search on a large scale of pods. We present a system architecture which can allow research, development and testing of new algorithms for decentralised search across pods. We undertake an initial validation of this architecture by usage scenarios for decentralised search under user-defined access control and data governance constraints. We conclude with research directions for decentralised search algorithms and deployment.

KEYWORDS

Redecentralisation, search, decentralised systems

1. INTRODUCTION

The provision that third parties can maintain indexes of Web resources has been a key architectural choice of the Web from the beginning [7], and has played a significant role in its growth by enabling search engines and supporting the discovery of user generated content. However, in recent years, a large part of user activity and generated data has been concentrated on a small number of online platforms that have evolved into data silos, raising concerns and leading to proposals for the redecentralisation of the Web [6]. SOLID[1] [15] is a proposed suite of technologies to support such redecentralisation by envisaging that user data be always maintained in user-controlled personal online datastores (pods) as opposed to online platform-controlled data silos. Online applications need to request and obtain access to user pods in order to function according to this paradigm. This can enable users to share their data with multiple online application providers, fostering data-driven innovation and AI. Nevertheless, this would also require support for large-scale data search across pods.

There has been a large body of previous work on topics closely related to search in such redecentralised environments but currently there is no conceptual model of what search functionalities across SOLID pods would require. We first review relevant literature and then propose an architecture to support search in a Web that has been redecentralised based on the concept of pods as proposed in SOLID. To that end we describe a logical pod structure and identify components that could effectively support search across a large scale of pods. We

[1] https://solidproject.org

perform an initial high-level scenario-driven validation of these proposals and we propose further research and development roadmaps.

2. RELATED WORK

Search in decentralised data ecosystems is a non-trivial problem since it can involve both keyword and database-type queries distributed over a scale of thousands of datastores to which different search parties can have different access rights and different data governance constraints on data storage or migration may apply. Earlier work has explored distributed queries making use of database schemas and statistics across database endpoints which can have varying types of autonomy [27,20,1,12]. There has also been work on distributed information retrieval using meta-information about databases [8], peer-to-peer (P2P) data management [2, 14, 29, 17], search optimisation in P2P systems [24], social-graph-informed query routing [18], and socio-aware P2P search including work in the Haggle project that used a distributed index for search within groups [30, 23]. There has also been research on schema-based P2P data management with semantic links between data shared by peers [21, 11, 16]. Large-scale distributed search, architectures, query propagation and performance have been explored in Gaian databases [4, 28, 5]. IPFS [3] is a more recent approach to storing and retrieving data on a global, P2P decentralised file system. However, varying access control to available data resources, query endpoints and indexes has not been central to the design of these earlier approaches.

Prior work on securing distributed queries on personal repositories has approached the issue from the angle of addressing privacy threats [13]. Other work has focused on architectures for enforcing access control policies in P2P environments [26] but not in the context of distributed queries. Also proposed is attribute-based search on encrypted data with access control focusing on centrally stored data on the cloud [32]. The complexity of dealing with identities for access control in large-scale fog/edge computing has led to proposals on using distributed hash tables as a substitute for access control lists [31].

There is also a body of work on distributed indexing techniques [9, 10] often focusing on sensor networks and events. Decentralised search engines such as BitClave leverage blockchain to let Web users share their data directly with advertisers, removing intermediaries, but the emphasis is on users protecting their own data rather than on search algorithms across a large scale of users' personal datastores.

3. PROPOSED SYSTEM ARCHITECTURE

3.1. Stakeholders

To meaningfully explore search in a decentralised Web we identify stakeholders and a decentralised search architecture based on SOLID. The first stakeholder to acknowledge is the ***pod user*** who effectively owns and controls their pod, setting the desired access policies over their pod data. There can be more complex ownership models, e.g.: a minor's data held in trust by a guardian; data that is held jointly by two parties, such as marriage-related data between two individuals; or community owned data. Another stakeholder is the ***pod provider*** that provides the digital service and infrastructure to manage and host a pod. There can be many distinct providers in an open marketplace. Search providers can offer an interface for searches to be requested, providing optimisation on queries and metadata shared by pod owners; however, we note that decentralised search does not necessarily require a search provider but can operate on a peer-to-peer basis. We can also identify the ***search issuer*** as the individual or organisation that issues a search query via a search interface. Finally, ***regulating entities*** create and enforce

access to information based on local and global laws, such as GDPR[2]. Any system that searches over information in pods must be able to accommodate the needs and requirements of each of these stakeholders.

3.2. Architecture

In the SOLID framework [15] individuals can identify themselves using WebID[3] and maintain their data resources within pods that can be stored in Web-accessible, user-owned or user-rented equipment, e.g. on local hard drives or on the cloud. Users decide who gains access to data in their pods by means of Access Control Lists and the Web Access Control system[4]. Pods can be hosted in pod servers (SOLID enabled Web servers) and their data can be accessed via RESTful interfaces, as in the Linked Data Platform recommendation[5]. Third-party applications can access data in users' pods if their WebID is linked to specific rights recorded for that data in the pod. URLs can be used to identify individuals or groups of individuals and their access rights to resources. SOLID pods may also offer SPARQL support including link-following SPARQL for specific applications, such as the Contracts app [15]. An underlying problem here is that, in allowing access to data in a pod, the user must pre-define who has access to their data; however, search in this case has a multi-faceted, undefined group of users and purposes. Other frameworks for cloud-based personal online datastores include that of the HAT project [22] which enables hosting of user-owned data, especially IoT data, in user-controlled containers similar to pods, and supports micro-services to enable an ecosystem of applications on that data.

Our proposal is for a refined logical structure for pods, compatible with the structure offered by SOLID and HAT, in order to support search algorithms across pods and foster research on optimisation techniques for both keyword and query-type search on a very large scale. There are two main themes required for searching within pods: a) to find appropriate pods that may contain the required data - but not share the data itself, merely identify it so that a contractual access negotiation can take place; and b) access to any publicly available data. In the first case, we need to make accessible to search algorithms possibly available data without impinging upon the privacy of the individual; efforts such as [19] could be drawn upon. We describe the second case in more detail below.

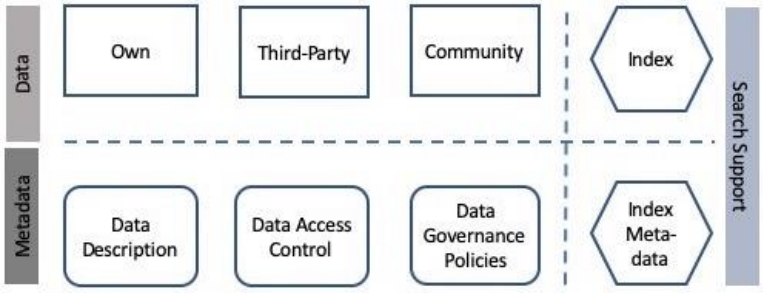

Figure 1. Pod logical structure to support decentralised search.

The logical architecture that we propose (illustrated in Figure 1) distinguishes between three types of pod data: *owner data*, *third-party data* stored in the pod with the consent of the user, and *community data* that are co-owned and potentially co-created by a community of users. SOLID envisages that a user may have several pods [25] but given that the same individual is always in full control of them, we can conceptually work with a single pod. Another distinction

[2] https://gdpr-info.eu
[3] www.w3.org/2005/Incubator/webid/spec/
[4] www.w3.org/wiki/WebAccessControl
[5] www.w3.org/TR/ldp/

in the proposed model is that of *data resources* and *metadata* to describe those resources; that distinction is made so that if a party issues a distributed query across pods and has access to pod metadata, query planning algorithms can decide whether a search query will be executed on a specific pod. Metadata in this sense are similar to terms like *meta-information*, *statistics* or *summaries* in the literature. In addition to such data description metadata, we envisage data *access control metadata* to distinguish metadata that can support optimisation when planning distributed search based on access criteria, for which agreement on schemas can be easier to reach. Finally, we also envisage *data governance policies* as another type of structured and potentiality widely agreed metadata on licensing, copyright, and data storage, migration and retention policies. To emphasise support for distributed keyword-based queries we also identify *indexes* as another type of data resource as well as index metadata to describe information on index access and use by search algorithms.

SOLID enables querying pods for access to locally stored data or for data in other pods using link-following [25]; the issuer of the query is responsible for retrieving links to other pods that can be queried. Our proposed search components architecture shown in Figure 2 aims to be compatible with search as either a third-party or a P2P application. For this reason, we distinguish between index building and index distribution components for optimisation in ways similar to those described in the literature but with additional capabilities to respect user-imposed data governance constraints such as access control, data indexing, and data migration across pods. We also distinguish between distributed keyword search engines and query engines as distinct components that could be used as P2P or third-party applications but could also be combined for more complex hybrid search cases. Finally, we envisage distinct components for query planning, optimisation and API interaction. We differentiate between optimisation and query planning since, especially for hybrid search, optimisation software can determine which query planning algorithms are most suitable each time based on the types of query, pod metadata, network topology etc. Adjacency of components in Figure 2 indicates where possible interfaces are likely to be defined (e.g. optimisation components interfacing with index distribution, distributed query engine and query planning ones) but it does not exclude other possibilities.

4. SUPPORT FOR USAGE SCENARIOS

We consider two scenarios for a high-level validation of our architecture that cover both top-k and exhaustive search under different types of access constraints of pod owners in terms of access control, data export and data processing.

Usage Scenario 1: top-k search to form a community of users who have a common interest in combating the risk of developing Type 2 diabetes by accessing appropriate recipes for meals. The users in this scenario are the *app developer* and a number of *users*; all users have one or more pods, holding non-personal application data for the developer and personal health and nutrition data for the other users. Suppose Eric has enabled access control on the personal data in his pod but has made some metadata public: the fact that he has personal health and nutrition data, in which schemas those data are available, licensing and copyright information on those metadata. He has also authorised an indexer app that maintains a local index for those metadata. The app developer rolls out a Web app which, using search components compatible with pod indexes, can first discover all users (including Eric) with potentially relevant health and nutrition data, and then offer them recipes that present an appropriate nutritional profile. Eric receives and approves a request to provide the Web app with access to his relevant pod data and to allow the app to use existing indexes in his pod or to create new ones for the purposes of the application. Eric can then start using the app; for example, he can request it to search for nutritional information from the *top-k* other users with a similar health profile to him and who have been showing a stable or improving health status.

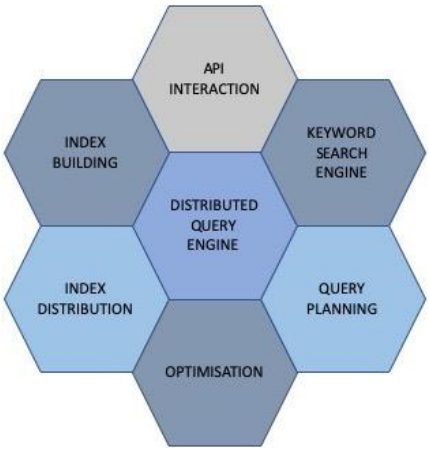

Figure 2. Search service components.

In this scenario, the provision of search components that can be integrated in applications, and indexers that can independently run on user pods, supports the deployment of a Web app that can provide users with relevant information without storing centrally their health and nutrition data. The indexers will use index building components, while the app will use keyword search on indexes to first identify users with relevant data and, once permission is granted, to execute distributed queries to obtain health profile and nutritional data using LDP/HTTP requests and/or SPARQL. Optimisation and query planning components can be integrated into the app to support these queries and link-following features to route queries across pods may be used.

Usage Scenario 2: exhaustive search over thousands of pods to investigate air quality in the city of Birmingham. Helen has a pod with geo-tagged data that she collects on her bike when cycling in Birmingham. She wants to contribute to the improvement of air quality but does not wish to compromise her privacy. She maintains her data in a pod and exposes metadata on the geo-range and time period covered in her pod. A developer provides a Web app that reports on levels of particulates around the city for different times of the day. Helen consents to the app using her pod with the condition that her data or query results cannot be stored by parties beyond her city boundary, and that they will always be aggregated with other query results to reduce the probability of triangulation by a certain factor. The app starts issuing queries to the user network, storing intermediate results in pods respecting their users' settings. Aggregated results are collected in the application pod and reported via the Web interface.

The application can use SPARQL link-following across pods for query propagation inspired by Gaian database approaches [28,5] but with extensions to support access control and restraints on the caching of pod data. Index building and distribution components specific for the app are used for query planning and optimisation. Optimisation algorithms need to consider network structure, user preferences on data aggregation and transfer, and routing distances for exhaustive search. A distributed query engine in combination with keyword search engine components and APIs support the Web app functionality.

5. CONCLUDING REMARKS

We have proposed an architecture to enable research and deployment of decentralised search at scale across SOLID pods, arguing that current distributed search techniques do not fully cover its requirements. We therefore propose that the community revisits distributed search in the context of decentralisation where access control and other data governance constraints are under the control of individuals. There is also scope to validate new decentralised search algorithms and optimisation approaches in existing or emergent ecosystems. Future work also requires

supporting the definition of data governance policies in user pods and monitoring their observance.

REFERENCES

[1] Abedjan, Z., Golab, L., Naumann, F.: Profiling relational data: a survey. VLDB Journal 24(4), 557–581 (2015)

[2] Androutsellis-Theotokis, S., Spinellis, D.: A survey of peer-to-peer content distribution technologies. ACM Comp. Surv. 36(4), 335–371 (2004)

[3] Benet, J.: IPFS - Content Addressed, Versioned, P2P File System. Tech. rep., Protocol Labs (2014)

[4] Bent, G., Dantressangle, P., Stone, P., Vyvyan, D., Mowshowitz, A.: Experimental evaluation of the performance and scalability of a dynamic distributed federated database. In: Proc. 3rd Ann. Conf. Int. Tech. Alliance (2009)

[5] Bent, G., Dantressangle, P., Vyvyan, D., Mowshowitz, A., Mitsou, V.: A dynamic distributed federated database. In: Proc. 2nd Ann. Conf. Int. Tech. Alliance (2008)

[6] Berners-Lee, T.: Long Live the Web. Scientific American 303 (2010)

[7] Berners-Lee, T., Cailliau, R., Groff, J.F., Pollermann, B.: World-wide web: the information universe. Internet Research (1992)

[8] Callan, J.: Distributed Information Retrieval, pp. 127–150. Springer (2000)

[9] Danzig, P.B., Ahn, J., Noll, J., Obraczka, K.: Distributed indexing: A scalable mechanism for distributed information retrieval. In: Proc. 14th Ann. Int. Conf. Res. and Dev. in Inf. Retrieval. p. 220–229. SIGIR '91, ACM (1991)

[10] Greenstein, B., Ratnasamy, S., Shenker, S., Govindan, R., Estrin, D.: Difs: a distributed index for features in sensor networks. Ad Hoc Networks 1(2), 333 – 349 (2003)

[11] Halevy, A.Y., Ives, Z.G., Suciu, D., Tatarinov, I.: Schema mediation in peer data management systems. In: 19th Int. Conf. Data Eng. pp. 505–516. IEEE (2003)

[12] Haslhofer, B., Klas, W.: A survey of techniques for achieving metadata interoperability. ACM Comp. Surv. 42(2), 1–37 (2010)

[13] Loudet, J., Sandu-Popa, I., Bouganim, L.: Dispers: Securing highly distributed queries on personal data management systems. Proc. VLDB Endow. 12(12), 1886–1889 (2019)

[14] Lua, E.K., Crowcroft, J., Pias, M., Sharma, R., Lim, S.: A survey and comparison of peer-to-peer overlay network schemes. IEEE Comm. Surv. Tutorials 7(2), 72–93 (2005)

[15] Mansour, E., Sambra, A.V., Hawke, S., Zereba, M., Capadisli, S., Ghanem, A., Aboulnaga, A., Berners-Lee, T.: A demonstration of the solid platform for social web applications. In: Proc. 25th Int. Conf. Companion on WWW. pp. 223–226 (2016)

[16] McBrien, P., Poulovassilis, A.: P2P Query Reformulation over Both-As-View Data Transformation Rules. In: Databases, Information Systems, and Peer-to-Peer Computing, pp. 310–322. Springer (2006)

[17] Meshkova, E., Riihija ̈rvi, J., Petrova, M., Ma ̈ho ̈nen, P.: A survey on resource discovery mechanisms, peer-to-peer and service discovery frameworks. ERCIM News 52(11), 2097–2128 (2008)

[18] Mislove, A., Gummadi, K.P., Druschel, P.: Exploiting social networks for internet search. In: In Proc. 5th Workshop on Hot Topics in Networks (HotNets-V) (2006)

[19] Mork, P., Smith, K., Blaustein, B., Wolf, C., Sarver, K.: Facilitating discovery on the private web using dataset digests. In: Proc. 10th Int. Conf. Information Integration and Web-Based Applications & Services. p. 451–455. ACM (2008)

[20] Naumann, F.: Data profiling revisited. SIGMOD Rec. 42(4), 40–49 (2014)

[21] Nejdl, W., Wolf, B., Qu, C., Decker, S., Sintek, M., Naeve, A., Nilsson, M., Palmer, M., Risch, T.: EDUTELLA: A P2P Networking Infrastructure Based on RDF. In: Proc. 11th Int. Conf. WWW. pp. 604–615 (2002)

[22] Ng, I., Maull, R., Parry, G., Crowcroft, J., Scharf, K., Rodden, T., Speed, C.: Making Value Creating Context Visible for New Economic and Business Models: Home Hub-of-all-Things (HAT) as Platform for Multisided Market powered by Internet-of-Things. In: HICSS (2013)

[23] Nordström,E.,Rohner,C.,Gunningberg,P.:Haggle:Opportunisticmobilecontent sharing using search. Computer Communications 48, 121 – 132 (2014)

[24] Reynolds, P., Vahdat, A.: Efficient peer-to-peer keyword searching. Springer (2003)

[25] Sambra, A.V., Mansour, E., Hawke, S., Zereba, M., Greco, N.: Solid: a platform for decentralized social applications based on linked data. Tech. rep., MIT CSAIL & Qatar Computing Research Institute (2016)

[26] Sandhu, R., Zhang, X.: Peer-to-peer access control architecture using trusted computing technology. In: Prof 10th ACM Symp. on Access Control Models and Technologies. p. 147–158. SACMAT '05, ACM (2005)

[27] Sheth, A., Larson, J.: Federated database systems for managing distributed, heterogeneous, and autonomous databases. ACM Comp. Surv. 22(3), 183–236 (1990)

[28] Stone, P.D., Dantressangle, P., Bent, G., Mowshowitz, A., Toce, A., Szymanski, B.: Query propagation behaviour in gaian database networks. In: Proc. Ann. Conf. Int. Technology Alliance (2012)

[29] Suryanarayana, G., Taylor, R.: A survey of trust management and resource discovery technologies in peer-to-peer applications. Tech. rep., UC Irvine (2004)

[30] Yoneki, E., Hui, P., Chan, S., Crowcroft, J.: A socio-aware overlay for publish/subscribe communication in delay tolerant networks. In: Proc. 10th ACM Symp. on Modeling, Analysis, and Simulation of Wireless and Mobile Systems. pp. 225–234 (2007)

[31] Zaghdoudi, B., Kaffel-Ben Ayed, H., Harizi, W.: Generic access control system for ad hoc mcc and fog computing. In: Foresti, S., Persiano, G. (eds.) Cryptology and Network Security. pp. 400–415. Springer (2016)

[32] Zheng, Z., Xie, S., Dai, H.N., Chen, X., Wang, H.: Blockchain challenges and opportunities: a survey. Int. J. of Web and Grid Services 14(4), 352–375 (2018)

Authors

Thanassis Tiropanis has a research track record on e-Infrastructures, starting with network management, following on with research on Grid infrastructures for content delivery, and more recently infrastructures for Web Observatories. He has worked on distributed queries across federated and distributed data repositories using semantic and peer-to-peer techniques, and on conceptu
al models of Data Observatories. At the same time, he engaged in interdisciplinary research exploring the impact of the Internet and the Web on society and the economy. More recently he has been contributing to reports and proposals for infrastructures and policies to support data sovereignty. He is a senior member of IEEE, and a chartered IT professional with BCS.

Alexandra Poulovassilis has been Professor of Computer Science in Birkbeck's Department of Computer Science and Information Systems (CSIS) since 2001, becoming Professor Emeritus in May 2021. Her research is in data integration, querying, visualisation and personalisation. She has published widely in these areas and held numerous grants from the UK Research Councils, EU and industry. She is Founding Director of the Birkbeck Knowledge Lab (Feb. 2016-March 2021), and prior to that Co-Director of the London Knowledge Lab (2003-2015), and has championed and engaged in interdisciplinary research across the sciences, social sciences and arts for many years.

Adriane Chapman is an internationally recognised expert in the field of data provenance. She won ACM SIGMOD's 2016 Test of Time Award for her work on provenance (Buneman et al. 2006), and has over 30 peer-reviewed publications spanning provenance, data integration, and trust. She has designed software systems for agencies across the US government. Relevant to this proposal is her work with the US Food and Drug Administration on searching for related data across independently operated and governed systems. She received a Program Recognition Award for her work on incorporating provenance into the National Geospatial-Intelligence Agency Human-Geography Pilot. She is currently the Director of the ECS Centre for Health Technologies. Data discovery, access and handling are major components of this work.

George Roussos is Professor of Pervasive Computing at CSIS, where he heads the Experimental Data Science Group and the IoT lab. He has over 20 years experience in leading and successfully delivering research projects supported among others by the AHRC, EPSRC, the European Commission and the Michael J Fox Foundation. His work pioneered participatory cyberphysical computing as the predominant methodology for the construction of mobile and pervasive computing systems. His current research interests include applications of the IoT in mobile healthcare, the effects of social activity on mobile system architectures, and exploring mechanisms to support navigation and findability. He has authored four books and over 100 research papers, and since 2004 serves on the ACM US Public Policy Committee.