Trustworthy Metadata for Decentralized Search

Marcel Gregoriadis

***Abstract*—Abstract goes here.**

I. Introduction

Search engines like Google control what content is visible, influencing public access to information. Centralized search engines track and store vast amounts of personal data because of their data-driven business model, compromising user privacy. The *United States vs. Google* trail evidence shows that Google has a 90% market share in sell-side advertisement inventory and 80% in buy-side demand [29]. The US court ruled that "Google is a monopolist, and it has acted as one to maintain its monopoly". According to this case, Google leveraged its dominant market position to maximize profits while actively preventing new competitors from entering the market. This underscores the need for a decentralized search engine.

Decentralized search has been investigated for multiple decades [3, 12, 13, 14, 17, 30]. It has proven to be extremely challenging to create a search engine that is truly decentralized, has acceptable performance, and offers effective spam resilience. Thus, while decentralized services like Bitcoin, IPFS, and BitTorrent are flourishing, research on decentralized search has stalled. Many published approaches contain central elements to make the problem easier to solve, such as central indexes [7, 24]. We argue that the main barrier to adopting decentralized search is the lack of trustworthy and descriptive metadata. Files are often non-textual and merely described by their name, as users lack incentives to annotate them with more searchable metadata, e.g., through tagging. Moreover, the lack of content moderation in decentralized systems causes the proliferation of spam.

We revisit the unsolved **trustworthy metadata problem** with the ongoing progress in machine learning. That problem is defined by ensuring the accuracy, reliability, and authenticity of metadata in decentralized systems. By using the pattern recognition and emerging language capabilities of artificial intelligence, we propose to leverage implicitly available metadata. Implicit metadata encompasses user preferences learned from previous searches, such as interests, language, and demographics. Furthermore, it includes file attributes, such as size, creation date, number of seeders, and network latency. Thereby, we are reducing the dependence on explicit metadata that is solicited from users, such as votes or tags. Specifically, our solution targets the relevance ranking of search results in decentralized storage systems. To this end, we propose the employment of Learning-to-Rank (LTR). LTR describes a class of machine learning techniques widely and successfully adopted in centralized information retrieval [2, 16]. We further leverage language models for the semantic embedding of queries and file names.

With our method, we eliminate the problem of motivating user participation by automatically extracting metadata from user behavior, allowing it to be generated "effortlessly". In this work, we propose and evaluate two implementations of LTR for decentralized search:

- **Local-Only:** Each peer has their personal LTR model, which they train on locally generated data, i.e., past search interactions.
- **Collaborative:** Peers train their local LTR model on locally generated data, and periodically gossip model updates; incoming model updates are aggregated with the local model.

Our performance measurements are based on real user interactions gathered from gossip exchanges in the decentralized file-sharing network Tribler [25].

The remainder of this article is structured as follows. In Section II, we present the challenges of decentralized search engines that our paper addresses. Section III reviews existing solutions and, at the same time, provides some background to the field of information retrieval and relevance ranking.

II. Problem Description

Big Tech makes extensive use of advanced machine learning for targeted advertisements, fighting spam, and organizing marketplaces. The field of decentralized learning has only recently emerged and is still taking shape.

Big Tech AI has massive computing power and data points across many types of human activities from billions of consumers. Transforming this into a collective search infrastructure, which is distributed across donated computational resources, introduces numerous research problems.

Trustworthy user metadata. User metadata, which tracks popular searches and trends among similar users, is essential in refining search algorithms. A key source of this information is the clicklog, which captures user behavior, including clicks and navigations. This implicit feedback enables the generation of accurate user profiles without requiring explicit metadata. However, to ensure privacy, it is critical that user profiles disseminated via the clicklog are anonymized and shared in a privacy-preserving manner.

Trustworthy document metadata. Effective content discovery in decentralized systems hinges on the accuracy and completeness of item metadata. Important metadata fields include the name of the content, its type, the language in which it is written, and the date of its creation.

To further enhance content categorization, microtagging enables detailed tagging, allowing users to attach granular descriptors to content. This level of precision in metadata ensures more effective and trustworthy search results.

Active attackers. In decentralized systems, it is critical to anticipate that attackers may possess an advanced understanding of the system's architecture, on par with its original designers. These attackers can flood the network with malicious content, such as spam, fraudulent advertisements (e.g., promoting Viagra under popular keywords), and misinformation. The scale of this issue is vast, with internet fraud and misinformation reaching levels that can influence significant societal events, such as elections [34]. For instance, in 2022, Facebook reported the removal of 4.8 billion fake accounts, underscoring the magnitude of this challenge [8].

True decentralized learning. In a true decentralized learning context, there is no central authority or single point of failure. All learning and data processing occurs entirely on the client side, ensuring that the system remains resilient and autonomous. This architecture eliminates dependency on any central server or coordinating entity, distributing responsibility and computation equally across all participants in the network.

To conclude, the problem of decentralized search engines may be formulated as finding trustworthy information, while under active attack, and preserving decentralization.

III. Background and Related Work

Metadata scarcity is a cardinal problem in online communities, which has been studied in the context of centralized [15, 18] as well as decentralized services [5, 20]. Specifically, this problem prevails with multimedia search, where content lacks textual descriptions [19]. Efforts to motivate voluntary user contributions have often relied on altruism and socio-psychological rewards [31, 9]. As users are usually busy, and annotating documents requires time and effort, When motivation is made extrinsic, e.g., through crypto-economic incentives, this encourages low-quality contributions and spam [26]. Likewise, in centralized applications, where users are incentivized by ad revenue or view counts, clickbait tactics emerge [4, 33].

A. Search Engines and Relevance Ranking

When a user submits a search query, the search engine's task is to retrieve a set of possible result candidates and then rank them based on their relevance to the query. Relevance ranking presents a core problem in information retrieval (IR). Search engines rank documents based on many criteria. Term-based techniques such as the classical BM25 [27] incorporate statistical measures like term frequency and document length to estimate relevance. Recently, neural approaches to IR are becoming more prevalent [21]. Large language models are capable of generating deep semantic embeddings of both queries and documents [10]. Embeddings allow for a richer understanding of semantic similarities, but they can lack the precision of term-based methods. Mitra et al. [22] demonstrated that the best results are achieved when embeddings and term-based techniques are used in conjunction. There are, however, also metrics that look beyond the query or document content, which can further improve retrieval performance. Google famously employs PageRank [1], which capitalizes on the intricate link structure of the web to infer the relevance of a webpage. As metadata is scarce, and because of the complexity of understanding user intent, search engines also turn to analyzing user engagement. For example, YouTube correlates watch time to the associated search query to assess the relevance of a video with the provided query [23]. Further, platforms such as Amazon and Netflix [6, 28, 32] use collaborative filtering to infer user preferences based on user or document similarity.

Given the wide range of metrics that can be derived from queries, documents, and user signals, weighing these parameters for optimal ranking is a nontrivial task [35, 36]. Learning-to-Rank (LTR) provides a machine learning-based method for solving this problem. It has been extensively researched and applied in various search engines to refine the ranking order of a retrieved set of result candidates [2, 16, 19].

B. Decentralized Search Engines

Centralized systems have a natural advantage, aggregating user data to fine-tune search algorithms. Decentralized systems face unique challenges, such as security, scalability, incentivization, and content moderation. A recent survey by Keizer et al. [12] revealed that no current system adequately addresses these issues in a comprehensive manner. Although numerous projects for decentralized search on decentralized data have been proposed [30, 13, 14], they generally focus on narrow aspects of the problem. Consequently, in practice, users still rely on centralized indices to locate files within decentralized storage networks. For instance, IPFS Search [11] provided such an index by using a crawler that tracks updates in IPFS and using Apache Tika for metadata extraction. However, due to the high cost of maintenance and the lack of a business model, the service was shut down in 2023 [7]. Some researchers have proposed to decentralize the process in IPFS Search and maintain the extracted metadata on the DHT [13, 37]. Wang and Wu [30] extend the metadata stored in the DHT by network metrics such as freshness, proximity, resource quantity, and bandwidth, and incorporate them in their ranking function. The decentralized file-sharing software Tribler [25] maintains an index of every torrent's number of seeders and leechers and their creation time. Similarly, these metrics are used in the search result ranking as they serve as indicators for the document's popularity.

-grank

References

[1] Sergey Brin and Lawrence Page. "The anatomy of a large-scale hypertextual web search engine". In: *Computer networks and ISDN systems* 30.1-7 (1998), pp. 107–117.

[2] Zhe Cao et al. “Learning to rank: from pairwise approach to listwise approach”. In: *Proceedings of the 24th international conference on Machine learning*. 2007, pp. 129–136.

[3] Yatin Chawathe et al. “Making gnutella-like p2p systems scalable”. In: *Proceedings of the 2003 conference on Applications, technologies, architectures, and protocols for computer communications*. 2003, pp. 407–418.

[4] Yimin Chen, Niall J Conroy, and Victoria L Rubin. “Misleading online content: recognizing clickbait as” false news””. In: *Proceedings of the 2015 ACM on workshop on multimodal deception detection*. 2015, pp. 15–19.

[5] Philippe Cudré-Mauroux et al. “PicShark: mitigating metadata scarcity through large-scale P2P collaboration”. In: *The VLDB Journal* 17.6 (2008), pp. 1371–1384.

[6] Michael D Ekstrand, John T Riedl, Joseph A Konstan, et al. “Collaborative filtering recommender systems”. In: *Foundations and Trends® in Human–Computer Interaction* 4.2 (2011), pp. 81–173.

[7] Frido Emans. *Bump in the road — web.archive.org*. https://web.archive.org/web/20240422190159/https://blog.ipfs-search.com/bump-in-the-road/. [Accessed 10-09-2024]. 2023.

[8] *Facebook fake account deletion per quarter 2023 | Statista — statista.com*. https://www.statista.com/statistics/1013474/facebook-fake-account-removal-quarter/. [Accessed 10-09-2024].

[9] Paulo B Goes, Chenhui Guo, and Mingfeng Lin. “Do incentive hierarchies induce user effort? Evidence from an online knowledge exchange”. In: *Information Systems Research* 27.3 (2016), pp. 497–516.

[10] Po-Sen Huang et al. “Learning deep structured semantic models for web search using clickthrough data”. In: *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. 2013, pp. 2333–2338.

[11] *ipfs-search.com*. ipfs-search.com. [Accessed 21-08-2024].

[12] Navin Keizer et al. “A Survey on Content Retrieval on the Decentralised Web”. In: *ACM Computing Surveys* 56.8 (2024), pp. 1–39.

[13] Nawras Khudhur and Satoshi Fujita. “Siva-the ipfs search engine”. In: *2019 Seventh International Symposium on Computing and Networking (CANDAR)*. IEEE. 2019, pp. 150–156.

[14] Mingyu Li et al. “Bringing decentralized search to decentralized services”. In: *15th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 21)*. 2021, pp. 331–347.

[15] Kimberly Ling et al. “Using social psychology to motivate contributions to online communities”. In: *Journal of Computer-Mediated Communication* 10.4 (2005), pp. 00–00.

[16] Tie-Yan Liu et al. “Learning to rank for information retrieval”. In: *Foundations and Trends® in Information Retrieval* 3.3 (2009), pp. 225–331.

[17] Boon Thau Loo et al. “The case for a hybrid P2P search infrastructure”. In: *Peer-to-Peer Systems III: Third International Workshop, IPTPS 2004, La Jolla, CA, USA, February 26-27, 2004, Revised Selected Papers 3*. Springer. 2005, pp. 141–150.

[18] Ana-Maria Manzat, Romulus Grigoras, and Florence Sèdes. “Towards a user-aware enrichment of multimedia metadata”. In: *2nd Workshop focusing on Semantic Multimedia Database Technologies (SMDT 2010)*. Vol. 680. CEUR-WS: Workshop proceedings. 2010, pp. 30–41.

[19] Tao Mei et al. “Multimedia search reranking: A literature survey”. In: *ACM Computing Surveys (CSUR)* 46.3 (2014), pp. 1–38.

[20] Michel Meulpolder et al. “Public and private BitTorrent communities: a measurement study.” In: *IPTPS*. Vol. 4. 2010, p. 5.

[21] Bhaskar Mitra, Nick Craswell, et al. “An introduction to neural information retrieval”. In: *Foundations and Trends® in Information Retrieval* 13.1 (2018), pp. 1–126.

[22] Bhaskar Mitra et al. “A dual embedding space model for document ranking”. In: *arXiv preprint arXiv:1602.01137* (2016).

[23] *Navigating YouTube Search - How YouTube Works — youtube.com*. https://www.youtube.com/intl/ALL_en/howyoutubeworks/product-features/search/. [Accessed 07-09-2024].

[24] Joost Poort et al. “Baywatch: Two approaches to measure the effects of blocking access to The Pirate Bay”. In: *Telecommunications Policy* 38.4 (2014), pp. 383–392.

[25] Johan A Pouwelse et al. “TRIBLER: a social-based peer-to-peer system”. In: *Concurrency and computation: Practice and experience* 20.2 (2008), pp. 127–138.

[26] Dandan Qiao et al. “Mitigating the adverse effect of monetary incentives on voluntary contributions online”. In: *Journal of Management Information Systems* 38.1 (2021), pp. 82–107.

[27] Stephen E Robertson et al. “Okapi at TREC-3”. In: *Nist Special Publication Sp* 109 (1995), p. 109.

[28] Muhammed Sütçü, Ecem Kaya, and Oğuzkan Erdem. “Movie Recommendation Systems Based on Collaborative Filtering: A Case Study on Netflix”. In: *Erciyes Üniversitesi Fen Bilimleri Enstitüsü Fen Bilimleri Dergisi* 37.3 (2021), pp. 367–376.

[29] U.S. Department of Justice. *Justice Department Sues Google for Monopolizing Digital Advertising Technologies*. Accessed: 2024-09-10. 2023. URL: https://www.justice.gov/opa/pr/justice-department-sues-google-monopolizing-digital-advertising-technologies.

[30] Feng Wang and Yanjun Wu. “Keyword search technology in content addressable storage system”.

In: *2020 IEEE 22nd International Conference on High Performance Computing and Communications; IEEE 18th International Conference on Smart City; IEEE 6th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*. IEEE. 2020, pp. 728–735.

[31] Xiahua Wei, Wei Chen, and Kevin Zhu. "Motivating user contributions in online knowledge communities: virtual rewards and reputation". In: *2015 48th Hawaii international conference on system sciences*. IEEE. 2015, pp. 3760–3769.

[32] Chao-Yuan Wu et al. "Using navigation to improve recommendations in real-time". In: *Proceedings of the 10th ACM Conference on Recommender Systems*. 2016, pp. 341–348.

[33] Savvas Zannettou et al. "The web of false information: Rumors, fake news, hoaxes, clickbait, and various other shenanigans". In: *Journal of Data and Information Quality (JDIQ)* 11.3 (2019), pp. 1–37.

[34] Xichen Zhang and Ali A Ghorbani. "An overview of online fake news: Characterization, detection, and discussion". In: *Information Processing & Management* 57.2 (2020), p. 102025.

[35] Le Zhao and Jamie Callan. "Term necessity prediction". In: *Proceedings of the 19th ACM international conference on Information and knowledge management*. 2010, pp. 259–268.

[36] Guoqing Zheng and Jamie Callan. "Learning to reweight terms with distributed representations". In: *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*. 2015, pp. 575–584.

[37] Liyan Zhu, Chuqiao Xiao, and Xueqing Gong. "Keyword search in decentralized storage systems". In: *Electronics* 9.12 (2020), p. 2041.