Report on the First SIGIR Workshop on Graph Search and Beyond (GSB'15)

Omar Alonso[1] **Marti A. Hearst**[2] **Jaap Kamps**[3]

[1] Microsoft, Mountain View CA, USA
[2] UC Berkeley, Berkeley CA, USA
[3] University of Amsterdam, The Netherlands

Abstract

Modern Web data is highly structured in terms of entities and relations from large knowledge resources, geo-temporal references and social network structure, resulting in a massive multidimensional graph. This graph essentially unifies both the searcher and the information resources that played a fundamentally different role in traditional IR, and "Graph Search" offers major new ways to access relevant information. Graph search affects both query formulation (complex queries about entities and relations building on the searcher's context) as well as result exploration and discovery (slicing and dicing the information using the graph structure) in a completely personalized way. This new graph based approach introduces great opportunities, but also great challenges, in terms of data quality and data integration, user interface design, and privacy.

We view the notion of "graph search" as searching information from your personal point of view (you are the query) over a highly structured and curated information space. This goes beyond the traditional two-term queries and ten blue links results that users are familiar with, requiring a highly interactive session covering both query formulation and result exploration. The workshop brought together researchers from a range of areas in information access, who worked together on searching information from your personal point of view over a highly structured and curated information space.

1 Introduction

Information on the Web is increasingly structured in terms of entities and relations from large knowledge resources, geo-temporal references and social network structure, resulting in a massive multidimensional graph. This graph essentially unifies both the searcher and the information resources that played a fundamentally different role in traditional IR, and offers major new ways to access relevant information. In services that rely on personalized information like social networks, the graph plays an even more important role, in other words: *you are the query*.

Graph search affects both query formulation as well as result exploration and discovery. On the one hand, it allows for incrementally expressing complex information needs that triangulate information about multiple entities or entity types, relations between those entities, with various filters on geo-temporal constraints or the sources of information used (or ignored), and taking into account the rich profile and context information of the searcher (and his/her peers, and peers of peers, etc). On the other hand, it allows for more powerful ways to explore the results from various aspects and viewpoints, by slicing and dicing the information using the graph structure, and using the same structure for explaining why results are retrieved or recommended, and by whom.

This new graph based information seeking approach introduces great opportunities, but also great challenges, both technical ranging from data quality and data integration to user interface design, as well as ethical challenges in terms of privacy; transparency, bias and control; and avoiding the so-called filter bubbles. Graph search is already available today in many flavors with different levels of interactivity. Social network-based services like Facebook and LinkedIn provide flexibility to search their personal network form many diverse angles. Web search engines like Google and Bing rely more on using graphs to show related content as a mechanism to include other possible contexts for a given query. Clearly, it is not limited to web, and can be applied to other highly structured data. Just to give an example, the hansards or parliamentary proceedings are fully public data with a clear graph structure linking every speech to the respective speaker, their role in parliament and their political party. Graph search allows to explore politics from the viewpoint of individual members of parliament or government.

At a high level, graph search seems limited to familiar entity types (e.g., Facebook entities) and templates. How far can this scale? Will this work on truly open domains? There is a huge potential to use the graph to go beyond recommendations for new friends and contacts or semantically related content. Unlocking the potential of richer knowledge sources for new search strategies requires us to think outside the box, by combining different insights from IR, semantic search, data integration, query expansion and user interfaces to name a few.

The rest of this report will follow the program structure of the workshop. The workshop started with a round of introductions where attendees explained their own interest in the area. This was followed by an overview of the open research questions raised by searching highly structured data from a personal point of view (Section 2). Next, in Section 3) we summarize the four keynotes who helped frame the problems and reach a shared understanding of the issues involved amongst all workshop attendees. Rose Marie Philip talked about personalized post search at Facebook, Swee Lim about graph search at LinkedIn, Doug Oard about good uses for crummy knowledge graphs, and Alex Wade about Microsoft Academic Graph. Section 4 discusses the six contributed papers, which were presented in a boaster and poster session. In the next session, participants divided over two discussion groups preparing arguments against or in favor of the need of fundamental changes in information access, and fought this out as in an academic debate (discussed in §5). In the final session the results and progress of the workshop was discussed and preliminary conclusions were drawn (discussed in §6).

2 Open Research Questions

We view the notion of "graph search" as searching information from your personal point of view (you are the query) over a highly structured and curated information space. This goes beyond the traditional two-term queries and ten blue links results that users are familiar with, requiring a highly interactive session covering both query formulation and result exploration.

This raises many open questions:

IR Theory What happens if search gets personal? Does this break the classic dichotomy between users and documents, as users are nodes in the social network data themselves? What is the consequence of ultimate personalization, as the local graph differs for all users? As the local graph structure is key, does this obviate the need for large central indexes? Do these types of requests fit in the classic paradigm (e.g., Broder's taxonomy)? How does this shift the balance between the control of the searcher and the ranker over the result set?

Data Integration Building a knowledge graph requires massive data integration at many levels: are there trade-offs in simplicity and level of detail (such as the classic knowledge representation trade-off)? What levels of granularity and comprehensiveness are needed for effective deployment? What quality is needed: is any noise acceptable? How to deal with near duplicate detection, conflation, or entity disambiguation?

Use Cases and Applications Rather than a universal solution, graph search is particularly useful for specific types of information needs and queries. What are the data and tasks that make graph search works? What kind of scenarios that would benefit from a graph model? In what context can switching perspectives by showing results from the vista of other persons useful?

Query formulation How to move from singular queries to highly interactive sessions with multiple variant queries? What new tools are needed to help a searcher construct the appropriate graph search query using refinements or filters to better articulate their needs, or explore further aspects? How can we augment query autocompletion to actively prompt user to interactively construct longer queries exploring different aspects?

Result Exploration There is a radical shift towards the control of the searcher—small changes in the query can lead to radically different result sets—how can we support active exploration of slices of the data to explore further aspects? Unlike traditional facetted search options, the result space is highly dynamic, how can we provide adaptive exploration options tailored to the context and searcher, at every stage of the process?

Evaluation How do we know the system is any good? How to evaluate the overall process, given its personalized and interactive nature? Can we rely on the direct evaluation of query suggestions and query recommendations? Are there suitable behavioral criteria for in the wild testing, such as longer queries, multiple filters, longer dwell-time, more active engagement, more structured-query templates? Can we use are standard experimental evaluation methods from HCI and UI/UX design?

Privacy Access to personal data is fraught ethical and privacy concerns, is there is similarly structured public data for scientific research? As an extreme form of personalization, how

to avoid the uncanny cave, filter bubbles and echo chambers? How ethical is it to privilege a particular query refinement suggestion over the many other possible candidates?

Further discussion on the challenges of graph based search can be found in [1].

3 Keynotes

Four invited speakers helped frame the problems and reach a shared understanding of the issues involved amongst all workshop attendees.

3.1 Personalized Post Search at Facebook

The opening keynote was given by Rose Marie Philip (Facebook) on "personalized post search at Facebook" [7].

There are over a billion people and over two trillion posts on Facebook. Among these posts, there are uniquely personalized answers to many search queries. Facebook post search aims to help people find the most personally relevant posts for each individual query, tailored to the content of people's networks. This requires structured search over the entities and content, taking into account what is accessible to the user at hand. Textual queries undergo many query annotation and rewriting steps are made to construct a final query. Boolean'ish operators over sets of entities and posts deal with structural parts and an advanced ranking over many features produces a final ranking. Diversification based on the content and on the poster, relative to the social graph of the user, is essential to create an attractive ranking. There are still many challenges ahead, including better ways to learn from interaction data (how to generalize from highly personalized interactions), a deeper query understanding exploiting the structured "world view" of the type and scale of the available data.

3.2 Graph Search at LinkedIn

The second keynote in the morning was given by Swee Lim (LinkedIn) on "graph search at LinkedIn" [5].

LinkedIn is the largest professional social network. LinkedIn's graph and search systems help our users discover other users, jobs, companies, schools, and relevant professional information. LinkedIn has a complex architecture offering many front end services (to both members and recruiters), and different business and data layers. The graph adds connections between the member's data, the connections data, and the follows data allowing for fast and efficient cross domain joins. The current architecture is the 3rd generation, and a next generation graph called Liquid is about to be deployed. It adds considerable power: allowing for n-way relationships, fast joins, and rich properties, at greater flexibility (no cost schema evolution) and a more expressive graph-oriented query language. Currently graph operations are over all data, providing exact results in a database style, but search is based on a single index per domain using fast search engine technology. It is an ongoing effort to work on closer integration of the graph structure and search components, leveraging the best of what each system does best. Swee also covered the use of open source at LinkedIn.

3.3 Good Uses for Crummy Knowledge Graphs

The first keynote in the afternoon was given by Doug Oard (University of Maryland) on "good uses for crummy knowledge graphs" [6].

In 1993, Ken Church and Ed Hovy suggested that before we ask how well some new technology meets the need we envision for it, we should pause and first reflect on the question of whether —now that we know something about what can be built—we are envisioning the right uses for what we have. They titled their paper "Good Applications for Crummy Machine Translation" [3]. At about that same time, information retrieval researchers obliged them by (generally without having read their paper) starting to work on cross-language information retrieval; arguably the best application for crummy machine translation ever invented. Does the same argument hold for knowledge graphs? The main argument is that knowledge graphs are useful, but are also crummy —but IR is the art of making crummy things useful. Even imperfect or incomplete knowledge graphs can have value for different uses: interactive graph traversal (e.g., suggestions through auto completion, or learning to rank based on session or trajectory); multi hop reasoning (e.g., by making simple inferences); or explanation (e.g., provenance, attribution, accuracy, etc. or access to contextualized resources). The obvious way forward would be a co-design of knowledge graphs and their applications: the knowledge graph construction should reflect the needs of the application, and the application design should be informed about the error characteristics of the knowledge graph.

3.4 Overview of Microsoft Academic Graph

The second afternoon keynote was given by Alex Wade (Microsoft Research) with an "overview of Microsoft academic graph" [13]

The Microsoft Academic Graph is a heterogeneous graph containing scientific publication records, citation relationships between those publications, as well as authors, institutions, journals and conference "venues" and fields of study. It is the next generation of academic search, backed by the Bing and Cortana infrastructure, and used with graph search style natural language queries in Bing, e.g., "papers about maximum entropy in kdd after 2015 by ..." The use of autocompletion to support user to formulate complex queries in natural language is similar to Facebook's graph search feature [11]. The data consists of about 100 million papers, 700 million citations, and 25 million authors. It is available on Azure—free of charge—and is used in the WSDM 2016 Cup Challenge.

4 Accepted papers

We requested the submission of short, 4–5 page papers to be presented as boaster and poster. We accepted a total of 6 papers out of 8 submissions after peer review (a 75% acceptance rate).

Jadeja and Shah [4] investigate data driven ways to visualize and navigate graph or tree structured data. Navigating or traversing highly curated graph data is an understudied problem, and hierarchical or tree visualizations can help create order and overview. When visualizing the data from the viewpoint of a particular node makes any graph data (such as social network data) look like a tree with the starting node (a person with all context) as point of origin.

Sabetghadam et al. [8] investigates ways of "reranking" results based on a graph traversal approach for multimodal IR, that is hoped to be robust over different distributions of modalities. The use case of multimodal IR in a curated data space with rich context presents a challenge, as different features and scores on different modalities will be very differently distributed in very different probability spaces. An application of Metropolis-Hastings as sampling/estimation method is suggested as (partial) solution.

Sakamoto et al. [9] investigate captioning or summarizing results in highly curated graph data. Succint descriptions are essential for effective graph exploration, and requires to take the context and structure into account. The paper discusses a particular graph of words, sentences and documents, and also touches upon semantic annotations, which would move the document and text space to an entity space, with all documents and text linked to a particular category or entity.

Santisteban and Cárcamo [10] investigates a variant of the classic Tanimoto or Jaccard similarity measure able to deal with assymmetry in directed graphs and subsumption hierarchies. Similarity measures are central in IR, and related distance measures central in graph data. The discussion is motivated by a use case of "paradigmatic" structures.

Tong et al. [12] investigate category and word relation graphs for retrieving trouble shooting information/documents, addressing the classical IR problem of human assigned controlled terms versus document free text in the context of a curated data space and rich context (at least in principle). The paper offers an interesting graph approach is outlined, mapping terms to categories, for both requests and documents. Making this graph level explicitly available to users offers interesting new possibilities, and opens up ways to map the noisy term occurrence space to the curated, concept and entity based space of the category codes. Hence this paves the way to a semantic, entity based view.

Yu et al. [14] investigates the strength of connections in an entity graph, specifically a scholarly network with a rich entity graph available as public data. This is an interesing use case with a curated data space and rich context, plus an interesting dynamic structure over time. The paper proposes to take the strength (or weakness) of connections into account—here as simulated blind feedback—turning a network into a weighted network of the simple graph into a valued graph.

5 Debate

The lively discussion of the poster session continued with an academic debate on radical and incremental changes in information access. The participants were split into two teams and prepared the arguments in favor of, or against a proposition. The proposition was:

> *Everything changes when searching information from your personal point of view (you are the query) over a highly structured and curated information space.*

Each team had one hour to prepare the debate, and nominate four members with an active role in the debate in the next hour. The affirmative team advocated fundamentally new modes of information access, and consisted of Alex Wade (captain), Alexander Kotov (first affirmative speaker), Stefan Heindorf (second affirmative speaker), and Bin Tong (third affirmative speaker). The negative team advocated incremental changes over the existing data and search

experience, and consisted of: Josef Robeischl (captain), Hadi Hashemi (first negative speaker), Hosein Azarbonyad (second negative speaker), and Amir Fayazi (third negative speaker).

The debate started with a short introduction on the proposition and the fundamentally different views on the way forward, and then the speakers of each team took turns and made there arguments, both by reacting to the previous speaker and by new arguments. The affirmative team argued for the clear usefulness to take user's context into account and for the extra power of querying the graph data. The negative team stressed in particular stressed the negative sides of "over-personalizing" results, in terms of risks to privacy, and in terms of risks to create echo chambers and to lead to a narrow scoped and impoverished information consumption. The discussion was both strong and entertaining, with agreement on the potential of personalization over over highly structured social information, and fierce disagreement on the risks and degree of personalization needed.

Both teams managed to convince the audience for their position at different times, making it a difficult call to determine which side won the debate. In the eventual audience vote the affirmative team was declared the winner—plausibly due to their quite liberal interpretation of fundamental change, and including many current personalization approaches into their position. However, the best speaker of the debate was Amir Fayazi of the negative team.

The debate led to further insight in the trade-off between using rich contextual profiles and information for personalized search, and the need to diversify and offer unfamiliar perspectives, and on how to give control and transparency to the user—let them be in the driver's seat—in order to avoid undesirable side effects.

6 Conclusions

The workshop brought together researchers from a range of areas in information access, who worked together on searching information from your personal point of view over a highly structured and curated information space. There was a feeling of clear progress being made, and that there is something potentially revolutionary changing.

Graph Search has fundamental consequences for information access and offers tremendous opportunities for building new systems and tools that allow users to explore information from many different angles, shifting control back to the user. This is a radical departure from current systems where machine learning dominates the interaction: the entire information space is determined by the user, and the user is in the driver's seat when expressing her needs and exploring the space of options interactive.

Last, but certainly not least, the workshop continued over a social event in the bar "The Clinic" in Santiago, `http://www.bartheclinic.cl/`, owned by the popular satirical/investigative newspaper with the same name, attended by workshop participants and other SIGIR attendees interested in the workshop's topic, combining great discussion with a sheer endless supply of food and drinks. Intense discussion about novel information access approaches and (scientific) life in general continued far into the Santiago night...

Acknowledgments We would like to thank ACM and SIGIR for hosting this workshop, the SIGIR workshop chairs Fernando Diaz and Diane Kelly, and in particular local arrangements chair

Diego Arroyuelo, for their outstanding support in the organization.

Details about the workshop including the presentations and slides are online at http://humanities.uva.nl/~kamps/gsb15/. The proceedings are available online at http://ceur-ws.org/Vol-1393/.

References

[1] O. Alonso and J. Kamps. Beyond graph search: Exploring and exploiting rich connected data sets. In *ICWE'15: Engineering the Web in the Big Data Era*, volume 9114 of *LNCS*, pages 3–12. Springer, 2015. URL http://dx.doi.org/10.1007/978-3-319-19890-3_1.

[2] O. Alonso, M. A. Hearst, and J. Kamps, editors. *GSB'15: Proceedings of the SIGIR'15 Workshop on Graph Search and Beyond*, 2015. CEUR-WS. URL http://ceur-ws.org/Vol-1393/.

[3] K. W. Church and E. H. Hovy. Good applications for crummy machine translation. *Machine Translation*, 8:239–258, 1993. URL http://dx.doi.org/10.1007/BF00981759.

[4] M. Jadeja and K. Shah. Tree-map: A visualization tool for large data. In Alonso et al. [2], pages 9–13. URL http://ceur-ws.org/Vol-1393/.

[5] S. Lim. Graph search at linkedin. In Alonso et al. [2], page 5. URL http://ceur-ws.org/Vol-1393/.

[6] D. W. Oard. Good uses for crummy knowledge graphs. In Alonso et al. [2], page 6. URL http://ceur-ws.org/Vol-1393/.

[7] R. M. Philip. Personalized post search at facebook. In Alonso et al. [2], page 7. URL http://ceur-ws.org/Vol-1393/.

[8] S. Sabetghadam, M. Lupu, and A. Rauber. Leveraging metropolis-hastings algorithm on graph-based model for multimodal ir. In Alonso et al. [2], pages 14–18. URL http://ceur-ws.org/Vol-1393/.

[9] K. Sakamoto, H. Shibuki, T. Mori, and N. Kando. Fusion of heterogeneous information in graph-based ranking for query-biased summarization. In Alonso et al. [2], pages 19–22. URL http://ceur-ws.org/Vol-1393/.

[10] J. Santisteban and J. T. Cárcamo. Unilateral jaccard similarity coefficient. In Alonso et al. [2], pages 23–27. URL http://ceur-ws.org/Vol-1393/.

[11] N. V. Spirin, J. He, M. Develin, K. G. Karahalios, and M. Boucher. People search within an online social network: Large scale analysis of facebook graph search query logs. In *CIKM'14*, pages 1009–1018. ACM, 2014. URL http://doi.acm.org/10.1145/2661829.2661967.

[12] B. Tong, T. Yanase, H. Ozaki, and M. Iwayama. Information retrieval boosted by category for troubleshooting search system. In Alonso et al. [2], pages 28–32. URL http://ceur-ws.org/Vol-1393/.

[13] A. D. Wade. Overview of microsoft academic graph. In Alonso et al. [2], page 8. URL http://ceur-ws.org/Vol-1393/.

[14] Y. Yu, Z. Jiang, and X. Liu. Random walk and feedback on scholarly network. In Alonso et al. [2], pages 33–37. URL http://ceur-ws.org/Vol-1393/.