Beyond graph search: Exploring and exploiting rich connected data sets

Alonso, O.; Kamps, J.

DOI
10.1007/978-3-319-19890-3_1

Publication date
2015

Document Version
Final published version

Published in
Engineering the Web in the Big Data Era

Beyond Graph Search: Exploring and Exploiting Rich Connected Data Sets

Omar Alonso[1] and Jaap Kamps[2](✉)

[1] Microsoft Corp, Mountain View, CA, USA
omalonso@microsoft.com
[2] University of Amsterdam, Amsterdam, The Netherlands
kamps@uva.nl

Abstract. Modern Web data is highly structured in terms of entities and relations from large knowledge resources, geo-temporal references and social network structures, 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 information retrieval. Graph search-based systems offer 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 novel 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.

1 Introduction

With the explosion of social networks, the term *graph* has become more ubiquitous than ever. When people talk about the Facebook graph or the Twitter followers graph, to name a couple of examples, the focus is on the relationships and their semantic meaning (e.g., friend, like, re-tweet, etc.) instead of just web objects like a page or an image. On the more traditional search engine arena, there has a been a lot of work on extracting more useful information from web pages so they can reflect these relationships. These efforts have produced new experiences for users.

Bing and Google have now graphs (named Satori and Knowledge Graph respectively) that can contribute more information to the search engine results page (SERP) than just ten blue links. Figures 1 and 2 show different strategies to present more information that is not usually displayed by the typical blue links. In the case of Bing, the input box autocomplete incorporates bits of the web page (image and snippet) as the user enters the query. For Google, a similar effect takes place but instead of showing the content on the pull-down, the extra information is rendered next to the search results. The benefit for the user is very clear: instead of issuing a query, examining the results, modifying the query

P. Cimiano et al. (Eds.): ICWE 2015, LNCS 9114, pp. 3–12, 2015.
DOI: 10.1007/978-3-319-19890-3_1

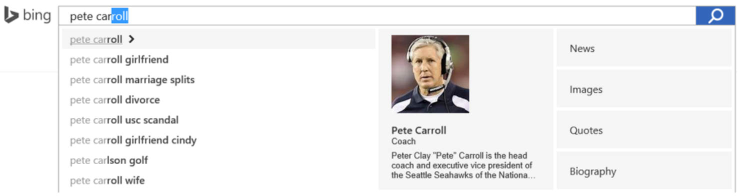

Fig. 1. Interactive autocomplete showing person name and basic information in Bing

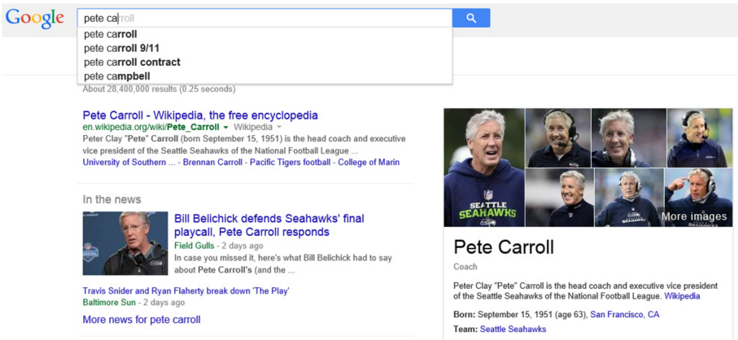

Fig. 2. Interactive autocomplete showing more information next to the SERP in Google

and examining again, this new model allows a more interactive *exploration* of what is possible before *exploiting* the relevant content that is being presented.

In 2013 Facebook introduced Graph Search, a search feature that allowed users to express more semantic queries for searching social content. The first version of Graph Search focused on four main areas (people, photos, places, and interests) and it was rolled out in English only. While still premature, it was possible to perform new types of queries and get results that were very different from what was the norm in a traditional SERP. Figure 3 shows examples of entity-based structured queries within Facebook graph search. At the time of writing, Facebook is replacing the novel graph search with a more traditional search over postings at the searcher's personal timeline[1].

Clearly, this approach 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 good graph structure linking every speech to the respective speaker, their role in parliament and their political party. Graph search allows us to explore politics from the viewpoint of individual members of parliament or government.

In this position paper we outline challenges, opportunities and possible research avenues for designing and building search services that can take full advantage of the potential of graphs and semantic information that can go beyond recommending friends or web links. Instead, we are interested in solutions that

[1] http://search.fb.com/

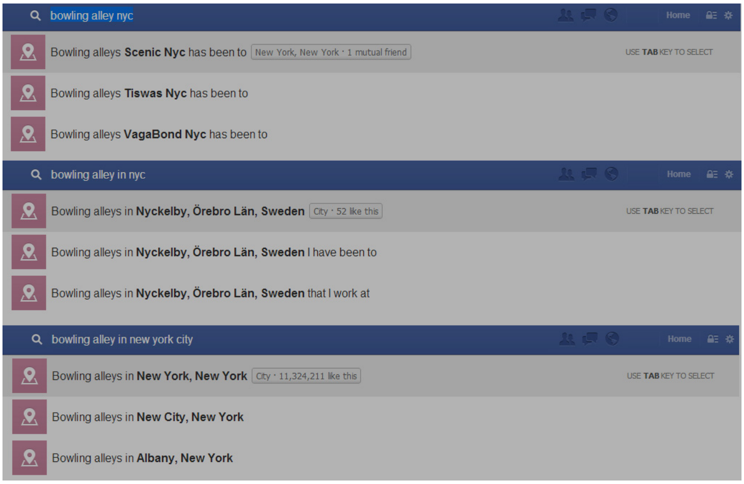

Fig. 3. Interactive autocomplete options within Facebook Graph Search

help users express information and navigate results better through the entire search space.

All the past and current efforts on using graph search have an exploratory information need in common. Information on the Web is becoming increasingly structured in terms of entities, types and relations from large knowledge resources, geo-temporal references and social network structures, resulting in a massive multidimensional graph that is very rich in content and connections. This graph essentially unifies both the searcher and the information resources that played a fundamentally different role in traditional information retrieval, and offers major new ways to access relevant information. In other words, the user is the default query. Or, *you* are the continuously running query on the system. Any user user input is added to this context.

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 location, 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 at various degrees of separation). On the other hand, it is an enabling mechanism for more powerful ways to explore the results from various aspects and viewpoints, by slicing and dicing the information using the graph structure, and also using the same structure for explaining why results are retrieved or recommended, and by whom.

This new graph based 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 [16], that is, a personal universe of information created just for a user using an array of personalizing filters. In the rest of the paper we outline some of these issues and shed light into potential solutions.

2 Open 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. It is also desirable to support graph search using an incognito mode with less available filters.

2.1 Two Step Interaction Model

Interaction plays a central role on this new model. The user starts by writing a query and the engine is expected to assist by providing potential query completions. At certain keystrokes, the engine will show results that will change as the user explores more the query formulations. The goal for the user is to maximize the knowledge gain while minimizing the cost of interaction. This is even more prominent in mobile scenarios where any assistance is welcome given the real estate limitations and potential input errors.

Incremental Structured Query Input. Creating a *graph query* requires incremental construction of a complex query using a variety of building blocks. Current search engines treat this as a form of query suggestion or query completion, which offers tailored suggestions trying to promote longer queries that cover multiple entity types and relations and various filters. Suggestions and entity types may be based on the user's own activity. This goes beyond prevailing autocompletion techniques, with previews and surrogates from traditional result pages or SERPs moving to a more dynamic query suggestion.

Dynamic Structured Result Set Exploration. Results are highly personalized: they are unique for the searcher at a given point in time. The result set is highly structured: rather than just showing the top-10 results from an almost infinite list, a faceted exploration based on user's interests or augmentation of the SERP is needed. The structure is dynamically derived from the graph structure and the user's point of view, rather than a rigid facet and facet value hierarchy.

2.2 Data Quality and Data Integration

Building a knowledge graph requires significant effort on data acquisition, cleaning, and 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 type of quality is needed and adequate? Is any noise level acceptable? How to deal with near duplicate detection, conflation, mappings, or entity disambiguation at scale?

2.3 Query Classification

Graph search also requires a new query classification scheme, beyond the traditional division into navigational, informational, and transactional queries. Is there a new way to characterize queries in this new model? Does the notion of information need change? It is the ultimate form of personalization, with the searcher not only responsible for the query but also determining the (slice of) the data being considered? What shifts in control and transparency are needed to accomplish this?

2.4 Graph Search Evaluation

This also presents a range of new evaluation problems. How to evaluate the overall process, given its personalized and interactive nature? How to evaluate the first stage as essentially a form of query autocomplete? And how to evaluate the second stage as to explore and exploit the result set? Can we evaluate user satisfaction and engagement differently?

2.5 Ethics and Privacy

Access to personal data is fraught with ethical and privacy concerns, is there 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?

2.6 When (not) to Use Graph Search?

Rather than a universal solution, the graph search is particularly useful for specific types of information needs and queries. This is also depending on the character of the data available. For example, the Facebook Graph Search emphasized the social network structure, friends and other persons, locations and location-tagged objects. Social network data is abundantly available (although getting access presents a major barrier) but also notoriously skewed. Rather the searcher's personal point of view, it can also be used to show results from the viewpoint of any person in the network. There are many interesting sets of data – both historically or modern – that capture both the persons and related information: think of parliamentary data in public government, or intranet data in enterprises and organizations. For 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.

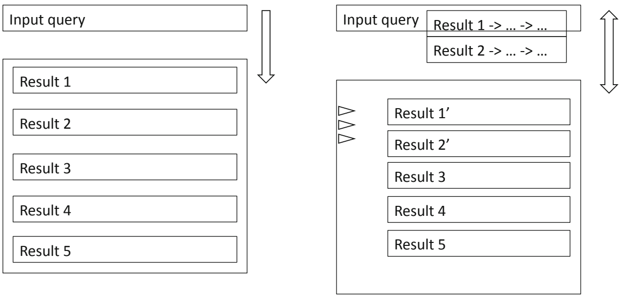

Fig. 4. Incremental Query Exploration Page (IQEP)

3 Potential Solutions

Graph search requires a highly interactive session covering both query formulation and result exploration. There is a seamless integration of what is shown as results depending of how far the user is expressing the query.

3.1 Query Exploration

There is a shift towards the control of the searcher, necessitating new tools that help the user construct the appropriate graph search query, and actively suggest refinements or filters to better articulate their needs, or explore further aspects [9]. This leads to a far more dynamic interaction than with traditional result lists, or modern hit lists showing summaries of a static set of results.

This suggests a new form of "query autocomplete" that invites and allows users to issue longer queries constructed based on entities, relationships, and templates. In contrast to SERP, we define IQEP as the *Incremental Query Exploration Page*. IQEP allows the user to explore more the result set as part of the input query. We can think of IQEP as an interactive mechanism that *promotes* relevant results selected by the user from the traditional SERP to the input box. In that sense, there is no static separation between query and search results. The more the user focuses or exploit a query term or entity, the more relevant content is moved upstream. Figure 4 shows IQEP as a bi-directional channel that moves results from the search list to the input box or vice versa (upstream/downstream). This mechanism enables the user to not only see what is possible but also to backtrack in case the content is not relevant or interesting enough anymore.

There are a range of suitable evaluation methods that we can use. The most obvious way is by direct evaluation of query suggestion and SERP. There is also a range of criteria useful for behavioral observation in the wild testing: users should issue longer queries, multiple filters, dwell-time, active engagement, structured-query templates. There are query segments where this type of querying is expected to be most useful: torso and tail queries (medium and rare occurrences based on query log distributions); exploratory scenarios. Traditional head or navigational queries, that is highly popular ones, seem less interesting, although these could be part of a more complex underlying information need.

This goes beyond Broder's taxonomy [4]: queries are all navigational, informational, and transactional *but* they are entity-focused. Queries may aim to return a single or a small set (not unlike traditional Boolean querying over structured data), or there is a need for data analytics on the whole set of results.

3.2 Result Exploration

This proposed shift towards the control of the searcher—small changes in the query can lead to radically different result sets—necessitating active exploration of slices of the data to explore further aspects.

This suggest a new form of search results unique for every user. Similarly to the query exploration mechanism, this interaction encourages users to explore over entities, relationships, and filters. Unlike traditional faceted search options, the result space is highly dynamic, and requires adaptive exploration options tailored to the context and searcher, at every stage of the process.

This is a radical departure from the traditional ten blue links available in any search engine. The IQEP moves from links to answers, and from answers to suggesting (expressions of) needs. This is an complete shift from the traditional dichotomy between query (the searcher's responsibility) and results (the system's responsibility). Traditional search results have moved to a hit list of result summaries (still a fix set of results, but the shown summaries are tailored to the searcher and her query). These summaries in terms of entities are now answers rather than links to answers. Now these results, or previews of them, are moving into the search box, in the form of structured query suggestions with some sort of preview indicating of the consequences on the result set (often in terms of numbers of results, or entity previews).

There are many options for the evaluation of components: (adaptive) captioning, (adaptive) filters, graph query templates. E.g., captioning should describe (relative to the entity), explain (relative to the user), and be contrastive (relative to the IQEP). There are standard experimental evaluation methods from Human-Computer Interaction and User Interface design [8]. With a running service, evaluation in the wild is very suitable. There are various implicit and explicit criteria: users should explore the result set, usage of multiple filters, dwell-time, active engagement, structured-query templates. Torso and tail queries, and exploratory scenarios are the most suitable query segments.

3.3 Barriers to Success

In order to work on graph search as a research community, we need to understand the problem better, and need to have sharable resources to do applied research and build common benchmarks.

The consequences of graph search are quite fundamental. Graph search has the potential to give users more power by unleashing semantically annotated information with many entities and relations between them. It brings the control back to the searcher, rather than leave it in the hands of the algorithm. Graph-based search systems also have the potential to solve part of the old information retrieval problem of conceptual search.

In terms of IR research and required evaluation methods, as discussed in the sections above, there are various open problems. What we need is to work on sharable research data, that exemplifies most of the characteristics we want to study. There is no need to be on Facebook or Twitter, or hand over your personal data. Similar small data sets and systems are available (e.g., so.cl, NYT, Parliamentary data, etc.) It will be hard to share a realistic subset of social network data but it should be possible to construct a simulated set.

What would be a concrete task to study on this data? Instead of implementing all features, it is would be useful to select a few components like query suggestion box, filters as facets, and captions to show the potential.

Search engine user interfaces has been very stable in the last 15 years. The input box and the ten blue links are the still the most optimal way to show search results. Can we do better in terms of user experience? This would give users a lot of flexibility and options. However, remains to be seen if users would adopt such dynamic interface.

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? Finally, there are a number of ethical issues such as privacy, transparency, bias and control, and filter bubbles.

4 Related Work

There has been published research on models for seeking information that are related to our approach, in particular information foraging [17] and information encountering [5].

There is considerable industrial activity around social graphs. The most famous example is Facebook Graph Search, a feature that allows users to perform more sophisticated searches on their social network. Bing has been integrating Facebook into their web search results for the last couple of years. Similarly, Google has been annotating search results with Google+ profiles. In terms of published literature, both Google and Microsoft have reported studies on different aspects of social annotations, e.g. [7,14,15].

Hearst [8] covers extensively issues with search user interfaces design and evaluation. In particular she addressed the user interface and user experience challenges of search results moving from the found links, to the HIT page as snippets, and now to the query auto-suggest box as previews of possible query extensions.

Graph search, or personalized search over a highly structured and curated information space, is closely related work on exploratory search and sense making

[13,18]. The graph structure provides natural facets for exploring the data, from a local point of view, allowing for a more dynamic structure than traditional faceted search using rigid, global, hierarchical structure.

There are crucial links with work on searching structured data [6,10], and work on the appropriate query languages [11,12]. These branches of research in particular focus on complex querying of structured text or data, whereas the graph search addresses also, and perhaps primarily, the process of constructing series of complex queries interactively.

5 Conclusions

This position paper introduced 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. Graph Search has fundamental consequences for web engineering and offers tremendous opportunities for building new systems and tools that allow users to explore information from many different angles. At the same time, using a graph to go beyond recommending friends or links requires solid knowledge of many components that need to work together. We presented the open questions and outlined a number of challenges and research directions that present some of the greatest challenges to work on in the coming years.

Acknowledgments. This research was supported in part by the Netherlands Organization for Scientific Research (ExPoSe project, NWO CI # 314.99.108; DiLiPaD project, NWO Digging into Data # 600.006.014).

We gratefully acknowledge discussion on this topic at the Exploiting Semantic Annotations in IR workshops (CIKM 2013 and CIKM 2014) [2,3], and at the Dagstuhl Seminar on Evaluation Methodologies in Information Retrieval [1]. A workshop devoted to this topic is planned at the SIGIR 2015 conference.

References

1. Alonso, O., Kamps, J.: Graph search and beyond. In: Agosti, M., Fuhr, N., Toms, E., Vakkari, P. (eds.) Evaluation Methodologies in Information Retrieval, Dagstuhl Reports, vol. 3 (10), chap. 4.3, pp. 108–111. Dagstuhl (2014)
2. Alonso, O., Kamps, J., Karlgren, J. (eds.): Proc. ESAIR 2014. ACM Press (2014). http://dl.acm.org/citation.cfm?id=2663712
3. Bennett, P.N., Gabrilovich, E., Kamps, J., Karlgren, J. (eds.): Proc. ESAIR 2013. ACM Press (2013). http://dl.acm.org/citation.cfm?id=2513204
4. Broder, A.: A taxonomy of web search. SIGIR Forum **36**(2), 3–10 (2002)
5. Erdelez, S.: Information encountering: a conceptual framework for accidental information discovery. In: Proc. of ISIC. Taylor Graham Publishing, London (1997)
6. Exploiting semantic annotations in information retrieval. Workshop (2009–2014)
7. Fernquist, J., Chi, E.: Perception and understanding of social annotations in web search. In: Proc. of WWW. International World Wide Web Conferences Steering Committee (2013)

8. Hearst, M.: Search User Interfaces. Cambridge University Press (2009)
9. Hearst, M.: How can search interfaces enhance the value of semantic annotations (and vice versa)? In: Proc. of ESAIR (2013)
10. Initiative for the evaluation of XML retrieval. Evaluation Forum (2002–2014)
11. Kamps, J., Marx, M., de Rijke, M., Sigurbjörnsson, B.: Articulating information needs in XML query languages. ACM TOIS **24**(4) (2006)
12. Libkin, L., Martens, W., Vrgoč, D.: Querying graph databases with XPath. In: Proc. of ICDT. ACM Press (2013)
13. Marchionini, G.: Exploratory search: from finding to understanding. Communications of the ACM **49**(4), 41–46 (2006)
14. Muralidharan, A., Gyongyi, Z., Chi, E.: Social annotations in web search. In: Proc. of CHI. ACM Press (2012)
15. Pantel, P., Gamon, M., Alonso, O., Haas, K.: Social annotations: utility and predictive modeling. In: Proc. of SIGIR. ACM Press (2012)
16. Pariser, E.: The Filter Bubble. Penguin (2011)
17. Pirolli, P.: Information Foraging Theory: Adaptive Interaction with Information. Oxford University Press (2009)
18. Pirolli, P., Russell, D.M.: Introduction to this special issue on sensemaking. Human-Computer Interaction **26**(1–2), 1–8 (2011)