Search in BigData^2 - When Big Text meets Big Graph

Christos Giatsidis, Fragkiskos D. Malliaros, François Rousseau, Michalis Vazirgiannis
Computer Science Laboratory, École Polytechnique, France
{giatsidis, fmalliaros, rousseau, mvazirg}@lix.polytechnique.fr

1. Introduction – State of the Art on Big Data

We have entered the *Era of Big Data.* The explosion and profusion of available data in a wide range of application domains rise up new challenges and opportunities in a plethora of disciplines – ranging from science and engineering to biology and business. One major challenge is how to take advantage of the unprecedented scale of data – typically of heterogeneous nature – in order to acquire further insights and knowledge for improving the quality of the offered services. To exploit this new resource, we need to *scale up* and *scale out* both our infrastructures and standard techniques.

The World Wide Web – along with the content and provided services – plays a prominent role towards shifting to the Big Data paradigm. We consider the *textual* and *social* contents of the Web as the most distinctive and significant with regards to Big Data. That is, on one hand you have the textual Web content – typical example of *Big Text* – that people want to easily consult and search to get answers and quench their thirst of knowledge. In 2004, a survey from the Pew Research Center reported that already "92% of Internet users say that Internet is a good place to go for getting everyday information" [4]. Recent surveys have shown that about half of the population of the US gets their news online, and about one third goes online every day for news [14]. But the Web now consists of hundreds of billions of pages – Google reported in 2008 that it just hit the trillion unique URLs visited [1]. Indexing this tremendous amount of data while allowing instant responses to user queries are two major challenges to tackle in the context of Web Search and Big Data.

On the other hand, you have the social structures formed over the Web – mainly represented by the online social networking applications such as Facebook, Google+ or Twitter. Typically, the interactions of the users within a social networking platform form graph structures, leading to the notion of *Big Graph*.
Everybody is different; except in a community – assuming that you can find that community! And the task can be hard when as of October 4th, 2012, Facebook reached the billion monthly active users, the hundred of billions of connections between these users and an average of 130 online friends per user. Furthermore, more than 3.2 billion likes and comments are performed every day, while more than one million websites are integrated with Facebook [15].

All these points stress out the importance of text and graph data in the Big Data era (mainly applied on the context of the Web, but also in other domains where text and graph data co-exist). Next, we will refer to some important application domains of Big Text and Big Graph, and then we will proceed with our vision and research, regarding the convergence of graph and text data.

2. Applications and Challenges

Based on the previous discussion, we focus our attention on two specific application domains of Big Data, namely *Web Search (Big Text)* and *Social Networks (Big Graph)*. In the case of Web Search, a user has an *information need* that he translates into a *free text query* and he expects to be returned by the search engine an ordered set of the most relevant documents for that query. In Social Networks, a user has a *social need* that is expressed by *links* with other people – indicating interactions – and he expects to share contents, discover people like him or get recommendations for stuff that he is interested in. Next, we emphasize on some key-challenges for Web Search and Social Networks that can be addressed in the paradigm of Big Data.

2.1. Challenges

We highlight here some challenges for the above applications that are still open problems in the context of Big Data and for which our research brings new insights and solutions.

Phrasal indexing: Should you index *big data* as two separate words or as a whole phrase in Web Search? Knowing at *indexing time* which phrases of a document are going to be searched at *query time* is a very hard problem. In a recent paper [3], two Googlers were still reporting that "there is no obvious mechanism for accurately identifying which phrases might be used in queries, and the number of candidate phrases is enormous since they grow far more rapidly than the number of distinct terms".

Document summarization: On the results page, we show document snippets corresponding to the query. Most search engine users now expect to find dynamic document summaries (query-dependent) that help them decide if a result is relevant to their information need or not without actually consulting it. Since the query is not known, it is hard to pre-compute the best summaries for each document in advance. Actually, good static summaries are already a problem for web pages.

Recommender systems: How can a movie recommendation system – with millions of users and thousands of movies – effectively recommend movies to users? Similarly, how to recommend products to potential customers in e-commerce platforms (targeted advertising)? In the case of online outsourcing marketplaces – such as Amazon Mechanical Turk – given a pool of individuals with different skills, how to form a team of experts for completing a specific task? These are only a few challenging applications that can be benefited by exploiting the properties of social graphs. Typically, users within such systems form social structures – in the sense that they create an underlying social network based on friendship or interaction relationships. Thus, several applications, including recommendations on the web and expert finding, can take advantage of these rich data structures to become more effective – increasing user's satisfaction level.

3. Our Vision and Research

We claim that research in graph theory can help solve both Web Search's and Social Networks' challenges. The most important phrases (for phrasal indexing) and set of keywords (for document summarization) in a document are the ones with terms sharing the strongest relations in a graph representation of the document. A graph whose nodes represent terms and whose edges represent relations between the terms – such as co-occurrence in a sliding text window – can be used to model this problem. Similarly, for targeted advertising campaign and recommendation, it is essential to find groups of people strongly connected to each other; the underlying social structure that is imposed by user interactions can contribute on this. Thus, the aforementioned open problems can be reduced to the following two:

- Community detection and evaluation
- Graph summarization

That way, graph theory and mining concepts can be applied on the graphs produced by the corresponding applications. However, the additional challenges concern the scale (size) of the graphs, as well as additional features associated with them (i.e., the graphs can be weighted or unweighted, directed or undirected, and they may contain signed edges capturing positive/negative interactions between entities). For the first point, we need tools and methods that are able to scale-up with respect to the size of the data. For the latter point, it is essential to extend graph-related concepts (e.g., density, degeneracy) to capture effectively the additional features.

In order to solve those problems, we are applying and extending state-of-the art techniques from graph theory, namely k-core degeneracy [11, 12, 13, 6] – an easy-to-compute reduction of a graph to its most connected nodes – and center-piece subgraph [10] – that consists of a small subgraph that best captures the connections between the source nodes, recently successfully applied to query recommendation for Web Search [2]. In particular, in our recent work, we have explored large-scale community evaluation based on the concept of degeneracy for two- and one- way relationships (undirected and directed networks respectively) and for trust networks with signed edges.

As for the role of big data in education we claim offering the students with equipment to handle the challenges of analyzing and mining Big Data and Big Graph. Methods we teach in a relevant course in a master course of Ecole Polytechnique include advanced machine learning techniques suitable for the requirements of Big Data.

4. Our Predictions and Recommendations

We claim that research in social networks and web mining will merge inevitably into one paradigm that will include information from both domains. Given the fact that both areas are evolving – if not already evolved – into Big Data, their combination will bring forth greater challenges. The merge has the potential to enrich search results in applications such as personalized search, recommender systems and social aware Web Search. Our future research will focus on improving the quality and scalability of data mining methods in those two areas.

References

[1] J. Alpert and N. Hajaj. We knew the web was big..., Official Google Blog, July 2008.

[2] F. Bonchi, R. Perego, F. Silvestri, H. Vahabi, and R. Venturini. Efficient query recommendations in the long tail via center-piece subgraphs. In: SIGIR, page 345–354, 2012.

[3] A. Das and A. Jain. Indexing the World Wide Web: The Journey So Far. In Next Generation Search Engines: Advanced Models for Information Retrieval , pages 1–28. IGI-Global, 2012.

[4] D. Fallows. The internet and daily life. Technical report, Pew Internet & American Life Project, 2004.

[5] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford University, 1999.

[6] S. Seidman. Network structure and minimum degree. Social Networks , 5:269–287, 1983.

[7] A. Singhal. Introducing the Knowledge Graph: things, not strings, Official Google Blog, May 2012.

[8] A. Singhal. Search, plus Your World, Official Google Blog, January 2012.

[9] T. Stocky and L. E. Rasmussen. Under the Hood: Building Graph Search Beta, Facebook Newsroom, January 2013.

[10] H. Tong and C. Faloutsos. Center-piece subgraphs: problem definition and fast solutions. KDD ’06, page 404–413, 2006.

[11] C. Giatsidis, D. Thilikos, M. Vazirgiannis,"D-cores: Measuring Collaboration of Directed Graphs Based on Degeneracy", Knowledge and Information Systems Journal, Springer, 2012.

[12] C. Giatsidis, K. Berberich, Dimitrios M. Thilikos, Michalis Vazirgiannis: Visual exploration of collaboration networks based ongraph degeneracy. In: KDD, 2012.

[13] C. Giatsidis, D. M. Thilikos, M. Vazirgiannis: Evaluating Cooperation in Communities with the k-Core Structure. In: ASONAM, 2011.

[14] http://www.people-press.org/2012/09/27/section-2-online-and-digital-news-2/

[15] http://visual.ly/facebook-2012-facts-figures

Work Group (please mark your position paper's objective(s) with a X):

University/Business driven Applications	**X**
University/Public driven Applications	
Data Challenges	**X**
Impacts on Organization Design	**X**
Partnership/Fellowship	**X**