DATA ANALYSIS

Aristides Gionis

Aalto University, Espoo, Finland
*Email:*Aristides.gionis@aalto.fi

1 STATE OF THE ART

The objective of this report is to highlight opportunities for enhancing global research data infrastructures from the point of view of data analysis. We discuss various directions and data-analysis functionalities for supporting such infrastructures.

We group the proposed data-analysis challenges around two main themes: (1) text analytics, information retrieval, filtering, aggregation, and dissemination via social platforms and (2) mining data consortiums. A broad distinction of the two themes is that the first theme refers to data that contain textual content and involve some amount of text processing while the second theme refers mostly to structured data. We proceed by providing a short description of each of the two themes and place them with respect to the state of the art.

1.1 Text analytics, Information retrieval, filtering, aggregation, and dissemination via social platforms

The traditional framework of information retrieval assumes a clear distinction of roles between *information producers* and *information consumers* (or *information seekers*). For example, in a repository of news articles, journalists (information producers) write news stories, and readers (information consumers) search for relevant information. Similarly, in a repository of scientific articles, documents are written by scientists and searched by a wide audience.

In recent years, however, with the advent of social-media and user-generated content platforms, we are witnessing a tremendous paradigm shift on how information is generated and disseminated (Solis, 2007). It is no longer the case that there is a clear separation between information producers and information consumers. Information consumers are not just passive receivers of search results. Instead they consume information by engaging in a wide range of activities, such as commenting, reposting, declaring favorite items, tagging resources with short keywords, sharing interesting content with friends, and so on. Such actions not only help us to understand the interests of the users, but they also help us to understand better the available content. It is fair to say that participating in information consumption while enhancing the available content with additional attributes can be perceived as an act of information production by itself. This new ecosystem of information generation and dissemination offers novel opportunities for information retrieval, filtering, aggregation, and dissemination, and at the same time it poses new research challenges. It has been an active area of research in the fields of information retrieval, text mining, user modeling, Web analytics, social-media analysis, and more.

We believe that global data research infrastructures can benefit immensely by this new social-media paradigm. Consider as an example the current system of scientific knowledge dissemination. It has not changed much during the last century, and it very much resembles the framework of traditional information retrieval that we described above: the roles of information producers and information consumers are clearly separated. Published articles form static entities, and scientists search for information without having their information-seeking activities provide any feedback in the actual scientific content. However, scientific endeavor is a collaborative activity where personal interaction, argumentation, and feedback play a vital role. Thus we envision an infrastructure for collaborative scientific research that encompasses many social-media capabilities: scientists share their articles, discuss and comment on past or current work, build personal profiles, get organized in communities, search for collaborators, search for outlets of their scientific outcome, seek the help of experts on specific problems of interest, and so on. The same paradigm can be extended to other contexts and application domains not only scientific collaboration. So, in general we envision an infrastructure that is able to support communities whose members can actively generate ideas and products, search for relevant personalized information, search for people, interact with each other, and collaborate towards common goals.

1.2 Mining data consortiums

Traditionally, data-analysis techniques have focused on the analysis of a single type of data. Examples include relational data, documents, transactional datasets, graphs, event sequences, and so on. On the other hand, the current digital revolution has enabled a large-scale collection of huge amounts of data regarding all kinds of human activity, and all kinds of measurements in scientific domains (Economist, 2010; Han, Altman, Kumar, Mannila, & Pregibon, 2002; Toffler, 1984). This abundance of available data has challenged not only the state-of-the-art in performing large-scale analysis on single-typed datasets but also the demand to develop data-analysis techniques that operate on multiple heterogeneous datasets, jointly analyze those datasets, and discover interesting patterns and correlations. Cross-analysis of heterogeneous datasets has already brought out a number of success stories. For example, Google Flu Trends has introduced an interesting technique to predict flu epidemics by analysis of Web query logs (http://www.google.org/flutrends/). As another example, in computational biology, the combination of microarray data with genetic-sequence data has been a key factor in understanding genetic variation. **In general, we believe that by combining multiple heterogeneous datasets, the possibilities of making interesting discoveries and examining new scientific hypotheses grow tremendously.**

Consequently, we envision a global research data infrastructure where many different datasets of extremely large scale are collected and stored. Sophisticated data-analysis techniques can then be applied to these datasets, either analyzing each dataset separately or performing joined analysis of appropriately selected subsets of datasets. We envision collecting datasets from diverse application domains, such as Web activity of users, macroeconomic variables, climate and environmental indicators, population and demographics, sensor recordings, telecommunication logs, biomedical engineering, ecology, and many more. For lack of a standard terminology we refer to such diverse collections of data as *data consortia*. Collecting the datasets in a unified infrastructure has many advantages. First, high-quality data-analysis software can be made available and shared among all researchers. In other words, the infrastructure provides not only the capability of collecting and storing the data but also toolboxes to analyze those data. Second, the ability to share data, set benchmarks, and deliver reproducible results promotes and fosters the scientific endeavor. In the same spirit, researchers can collaborate in order to improve the wealth of shared data; for example, not only raw-data become available but also the (meta-) data resulting from performing an interesting analysis on existing datasets. Researchers can also coordinate in order to improve the quality and coverage of data; for example, they can identify inconsistencies or missing data and work towards remedying those problems. Third, and in accordance with our previous discussion, performing data analysis on multiple heterogeneous datasets forms a basis for making new discoveries, discovering interesting new associations, falsifying scientific hypotheses, and in general, better understanding human activity and natural phenomena.

2 TEN YEAR VISION

We identify specific data-analysis activities within the two themes that we discussed above. Some of the proposed activities can be addressed directly by standard data-analysis methods for which good solutions are handily available. Other activities refer to more recent and more challenging problems and are on-going work by researchers in the field of data analysis; nevertheless, in many cases preliminary solutions already exist, and much better solutions are expected to be developed in the next 10 years.

2.1 Text analytics, Information retrieval, filtering, aggregation, and dissemination via social-media platforms

The following data-analysis functionalities are fairly general, and they will be quintessential ingredients of user-generated content platforms in the next decade.

— *User modelling:* The actions of the users in the system (information they produce, links they share, feedback they provide, discussions, queries, tags, etc.) can be used to build accurate user profiles (http://research.yahoo.com/workshops/umwa2011/). These profiles model not only the interests of the users but

also their skills, expertise, trustworthiness, how influential they are as well as the modes of interaction of the users with the system—for instance, whether a user likes to read content about a certain topic but almost never produces any content about it. Building user profiles that accurately model those characteristics is a very interesting research problem.

— *Personalization and recommendation*: The main application of building accurate user profiles is to put those models in the service of the user. *Personalized search* (Haveliwala, 2003), (Jeh & Widom, 2003) and *collaborative filtering* (Breese, Heckerman, & Kadie, 1998; Malone, Grant, Turbak, Brobst, & Cohen, 1987) are two primary examples of personalization. By personalized search, we mean that when a user performs a search action, the system provisions to offer results that are tailored to the interests and the level of expertise of the user. By collaborative filtering we mean that the system makes concrete suggestions for certain items, which it is likely the user will find highly relevant and interesting. In the current era of *information overload* (Toffler, 1984) where users do not necessarily know what to search for, collaborative filtering is an essential tool that allows users to navigate in information space and find interesting items.

— *Information dissemination:* By information dissemination, we refer to the following problem: assume that a user has produced a new item, such as a blog post. Who would be the best audience to send this new item to in order to receive the maximum possible attention and relevant feedback? This problem is related to *expert finding* (Balog, Azzopardi, & de Rijke, 2006), and it is also a form of personalized recommendation. Viewing the user-generated content platform as an information market where supply needs to match demand (De Francisci Morales, Gionis, & Sozio, 2011), the problems of information dissemination and collaborative filtering complement each other nicely.

— *Composite-item retrieval:* The typical setting of information retrieval is to retrieve top-*k* relevant items for a given query. Usually there are no interrelational constraints for the retrieved items. A more challenging setting is to request a bundle of items that are all relevant while at the same time satisfy certain constraints; for example, they are diverse or they cover a complex issue from many different points of view (Basu Roy, Amer-Yahia, Chawla, Das, & Yu, 2010).

— *Social interactions:* Certain information-processing tasks do not depend on any context (say, searching for the properties of a certain chemical compound) while others are very local and social (searching for an expert in a certain field who has the reputation of being "approachable" and to whom I can be introduced by a common collaborator). The system should be able to support both types of activities, social and non-social, as well as to operate in interim modes.

— *Communities:* The system should be able to identify meaningful communities for the users (Newman, 2006). Different communities should be identified for different target concepts, for example, similar interests, social connectivity, types of social connectivity, professional ties, geographic proximity, and so on. Overlaps between the communities should be possible because, by nature, people belong to different types of communities simultaneously (Banerjee, Krumpelman, Ghosh, Basu, & Mooney, 2005). Ideally, the system should provide explanations or short summaries of the communities discovered. The users should have the possibility to provide feedback and correct the proposed communities, and the system should be able to take the feedback into account and improve the community-detection algorithms.

— *Collaboration:* The system should provide tools that enable people to collaborate effectively. Collaboration can be enabled at various degrees, from simple tasks to very complicated. An example of a simple task is for the system to support mechanisms so that a set of users can collectively edit a document (such as a version-control system or Google documents). Finding an expert on a given topic can also be viewed as a subtask towards collaboration. A more difficult collaboration problem is to find a team of experts whose expertise covers all the skills required to accomplish a complex task (Lappas, Liu, & Terzi, 2009). The problem can be extended to incorporate various constraints, such as availability and load balancing (the team of experts are not too loaded), geographic proximity, social ties (the team of experts have worked in the past and proved that they can collaborate effectively), and so on.

— *Data aggregation and data mining:* In addition to the functionalities that the system provides to its users, it should be possible to generate analytics that reflect global behavior as well as to support mining for interesting usage patterns. Examples of such analytics include identifying global and local trends, analysis of sentiments on different topics, identifying the most influential users, explaining evolution aspects (how entities and relations in

the system change over time), and so on. Such analytics may also need to be available in interactive mode; for example, a user asks for the major trends in his/her social neighborhood or with respect to a subset of users with certain interests.

2.2 Mining data consortiums

As we discussed in the state-of-the-art section, the vision reflects a global research data infrastructure where researchers can collect and share their data, coordinate in improving the data quality, perform various data-analysis tasks, and combine the multiple sources of heterogeneous data in interesting ways. Some of the concrete functionalities that should be provided by such an infrastructure are the following.

— *Support standard data-mining and machine-learning algorithms:* As a minimum requirement, efficient implementations of general-purpose data-analysis algorithms should be available through the data consortium, and it should be possible to employ them effortlessly on a wide range of datasets. Examples of data-analysis tasks include clustering, classification, correlation analysis, regression, feature selection, frequent-pattern mining, and so on. For each task, a variety of algorithms should be available; for example, decision trees, naïve Bayes, *k*-nearest-neighbor classifier, support-vector machines, and more should be available for the task of classification. This activity involves a large amount of engineering effort (or the import of some existing data-analysis solution).

— *Data visualization:* Similar to the previous functionality, the infrastructure should provide ways to visualize the available data as well as the resulting data-analysis models. Examples include histograms, scatter plots, projections to low dimensions, graph visualizations, clustering, classification visualizations, and so on.

— *Information-integration capabilities:* The infrastructure should provide methods to integrate heterogeneous data and make it possible to link and associate the data. Some of the relevant tasks here refer to finding dependencies and associations among the attributes of different data tables as well as matching schemas and performing data cleaning. Most of those problems have been studied extensively in database research.

— *Mining complex data types:* Many of the activities described above refer to relational data types, such as tables, with data instances being the rows and attributes being the columns. A more challenging environment is to extend the data-analysis environment for more complex types of data, such as graphs (social networks, protein-interaction networks, etc.), data including geographic coordinates, images, etc.

— *Privacy:* It should be possible to share datasets that contain sensitive information, without compromising the privacy of the individuals whose information is represented in the data (Agrawal & Srikant, 2000; Sweeney, 2002). Privacy considerations are very important when sharing data in applications such as biomedicine, finance, social networks, Web-user activity, and others. Conservation of privacy can be achieved by anonymization techniques. The name of the game is to anonymize a dataset while not distorting the data entries too much and maintaining the utility of the dataset for data-analysis purposes. Developing privacy-preserving data-analysis techniques is an active area of research.

— *Collaboration:* The combination of data and data-analysis platforms requires not only the functionalities described above, but it also requires collaboration between data providers, engineers, and researchers. The infrastructure should provide tools that facilitate the collaboration of researchers and experts. As an example, consider a group of environmentalists who want to collaborate in order to collectively gather observation data about an endangered species, perform a correlation analysis with variables related to human-development activity, and then produce a report summarizing their findings.

— *Mining multiple and heterogeneous datasets:* The last activity is the most challenging and the one that it is the furthest from the state of the art. The vision is to develop capabilities of analyzing jointly many diverse and heterogeneous datasets. For example, given a dataset on species occurrence over a geographic region and a dataset of environmental variables over the same region, one is interested in mappings and interesting associations between combinations of species and combinations of environmental variables. As another example,

one may want to analyze how economic growth factors in different cities relate to literacy rates or pollution indices. The system should be flexible enough to allow specifying the type of analysis and the way that the data sources should be linked from a pool of available task joint operators.

3 CURRENT CHALLENGES

A major challenge for both of the proposed themes, text analytics and data consortiums, is to support collecting, storing, and analyzing very large and heterogeneous datasets. Implementing general-purpose data-analysis methods that work efficiently and accurately for datasets of different properties, different types, and different distributions requires significant engineering effort and thorough understanding of the problem at hand. Perhaps the largest challenge is to implement methods that analyze jointly multiple datasets. This is a problem for which currently there is no known general theory, and existing solutions are application-driven and ad hoc.

4 RESEARCH DIRECTIONS PROPOSED

In the previous sections we suggested an extensive and diverse list of research activities. To provide a more detailed formalization of these activities and propose concrete solutions exceeds the purpose of this document. Instead we provide some general remarks on the research directions and methodology that should be followed.

The research methodology should be fairly standard. First the research activities need to be understood and specified in detail. At the same time, the relevant literature needs to be reviewed carefully as existing methodologies and existing solutions should provide the basis for extensions and improvements. Then solutions need to be designed, implemented, and evaluated.

Some of the activities we propose can be addressed with different methodologies, for example, probabilistic methods, combinatorial algorithms, graph theoretic approaches, and so on. We would not like to make concrete suggestions on which methodology to be followed for each task; the researchers should be free to choose the methodology of their preference and the one they think is more appropriate for a give task. However, it would be valuable to investigate more than one solution and always compare with baseline approaches.

Some of the proposed activities (e.g., *support standard data mining and machine-learning algorithms* or *data visualization*) do not represent novel research problems; they require implementing existing solutions to standard problems. Nevertheless, we propose emphasizing the implementation of such basic activities as they can provide the necessary platform for developing the more advanced tasks. We also propose giving priority to collecting good-quality datasets and building benchmarks. Quite often, gaining insight on the real data not only can shape the solution to a problem but can also help in defining the right problem to solve.

Some of the proposed activities (e.g., *personalization and recommendation*, *information dissemination*, *composite-item retrieval*, etc.) have strong emphasis on information retrieval aspects; that is, the systems should not only be efficient, but they should also produce results that are useful for the users. For such problems we recommend conducting thorough user studies. These days, with the use of crowd-sourcing platforms, conducting user studies has been simplifying considerably.

5 RECOMMENDATIONS

As we mentioned, some of the proposed activities are well understood, and there exist standard methods while others correspond to open research questions. Naturally we would recommend to the participants of the GRDI2020 project that they start their design and implementation by addressing the easy questions first, before gaining expertise and moving to the more difficult problems.

Another recommendation is to not try implementing everything from scratch but to first investigate existing solutions, for example, existing data-analysis toolboxes.

Finally, we recommend striving for standardization and an open-development environment.

6 REFERENCES

Agrawal, R. & Srikant, R. (2000) Privacy Preserving Data Mining. In *Proceedings of the ACM SIGMOD Conference Management of Data*.

Balog, K., Azzopardi, L., & de Rijke, M. (2006) Formal models for expert finding in enterprise corpora. In *Proceedings of SIGIR*.

Banerjee, A., Krumpelman, C., Ghosh, J., Basu, S., & Raymond, M. (2005) Model-based overlapping clustering. In *Proceedings of the 11th International Conference on Knowledge Discovery in Data Mining (KDD).*

Basu Roy, S., Amer-Yahia, S., Chawla, A., Das, G., & Yu, C. (2010) Constructing and exploring composite items. In *Proceedings of the ACM SIGMOD Conference Management of Data.*

Breese, S. J., Heckerman D., & Myers, K. C. (1998) Empirical Analysis of Predictive Algorithms for Collaborative Filtering. In *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence*.

De Francisci Morales, G., Gionis A., & Sozio M. (2011) Social content matching in map-reduce. *In Proceedings of the VLDB Endowment (PVLDB)*.

Cukier, K. (2010) Data, data everywhere. Special Report, Managing Information. *The Economist*. Retrieved from the World Wide Web May 7, 2013: http://www.economist.com/node/15557443?story_id=15557443

Han, J., Altman, R., Kumar, V., Mannila, H., & Pregibon D. (2002) Emerging Scientific Applications in Data Mining. *Communications of the ACM 45*(8), pp 54-58.

Haveliwala, H. T. (2003) Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search. *IEEE Transactions on Knowledge and Data Engineering.*

Jeh, G. & Widom, J. (2003) Scaling Personalized Web Search. In *Proceedings of the 12th International World Wide Web Conference.*

Lappas, T., Liu, K., & Terzi, E. (2009) Finding a Team of Experts in Social Networks. In *Proceedings of the 15th International Conference on Knowledge Discovery in Data Mining (KDD).*

Malone, T., Grant, K., Turbak, F., Brobst, S., & Cohen, M. (1987) Intelligent Information Sharing Systems. *Communications of the ACM 30*(5), pp 390-402.

Newman, M. (2006) Modularity and community structure in networks. In *Proceedings of the National Academy of Sciences, 103*(23)

Solis, B. (2007) The Social Media Manifesto. Retrieved from the World Wide Web May 7, 2013: http://www.briansolis.com/2007/06/future-of-communications-manifesto-for/

Sweeney, L. (2002) k-anonymity: a model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10*(5), pp 557-570.

Toffler, A. (1984) *Future shock*. Random House Publishing Group.

(Article history: Available online 1 July 2013)