Project 1. (Ioana Manolescu, LIX) Summary-based optimization in semantic graph databases

Modern data is increasingly represented as graphs, with RDF (the Resource Description Format) pioneered by the W3C being the most popular such data format; the Linked Open Data Cloud (http://lod-cloud.net/) is one of its best-known example.
RDF graphs are express not only data (nodes, or resources, and their properties) but also semantics, that is: the types to which resources belong, and the relationships between the properties of a resource and the types which the resource has.

RDF graphs can be exploited by means of path or graph queries, where the users look for nodes having certain properties, e.g. searching for "X such that the type of X is Computer and the brand of X is Mac". Users may also specify property values for the resources they are looking for, either exactly (e.g., OperatingSystem="X Yosemite") or approximately (e.g., OperatingSystem contains "Yosemite"). There are many different ways to process such a query, and their performance may be drastically different.

The purpose of the project is to devise novel strategies for chosing the most efficient way to evaluate a given RDF query, given on a structural summary describing the possible paths in the data graph. Toward this goal, to each summary node should be attached statistic or synopsis information, compactly specifying the set of values found in the data graph nodes corresponding to that summary node. Then, given a query, an algorithm needs to be developed which enumerates possible query evaulation strategies and choses one likely to be the most efficient.

References

1. Resource Description Format Primer: https://www.w3.org/TR/2014/NOTE-rdf11-primer-20140225/
2. Yannis Ioannidis: "Query optimization", ACM Computing Surveys (CSUR) Surveys Volume 28 Issue 1, March 1996 (available also at http://graal.ens-lyon.fr/~yrobert/henri3/ioannidis96query.pdf)
3. Roy Goldman, Jennifer Widom: "Dataguides: Enabling Query Formulation and Optimization in Semistructured Databases", http://infolab.stanford.edu/lore/pubs/dataguide_vldb97.pdf

Project 2. (Ioana Manolescu, LIX) Statistics for JSON databases

JSON (Javascript Object Notation) is a document format overwhelmingly used in modern applications; Twitter messages are a notable example of JSON documents. A JSON document can be seen as a tree whose nodes are either text fragments, attribute-value pairs, or arrays; the model supports nesting and thus JSON documents can be arbitrary-depth trees.
Several NoSQL and "Big Data" systems have been developed for storing and querying JSON data, but most of these (such as MongoDB) only support simple forms of look-ups by searching top-down or through a mechanism akin to keyword search.

The purpose of the project is to devise and implement: (1) a set of value and structure statistics which can be efficiently gathered from a JSON corpus; and (2) a cardinality estimation module which, based on the statistics and a query, efficiently estimates the number of results which we expect the query to have on the dataset on which statistics have been computed. Error bounds and confidence in the cardinality estimations should also be characterized.

References

1. Introduction to JSON: https://www.w3schools.com/js/js_json_intro.asp

2. Yannis Ioannidis: "Query optimization", ACM Computing Surveys (CSUR) Surveys Volume 28 Issue 1, March 1996 (available also at http://graal.ens-lyon.fr/~yrobert/henri3/ioannidis96query.pdf)

3. Jason McHugh, Jennifer Widom: Query Optimization for XML, Very Large Databases Conference 1999. http://www.vldb.org/conf/1999/P32.pdf

Project 3. (Ioana Manolescu, LIX and Le Monde) Search on heterogeneous journalistic data sources

Fact-checking is hot data management applications today, as witnessed by the development of fact-checking organizations (http://www.factcheck.org, https://crosscheck.firstdraftnews.com) and media teams (http://www.lemonde.fr/les-decodeurs/, http://www.liberation.fr/auteur/15236-service-desintox). Data journalism is a related area, whereas journalists produce a piece of news based on high-quality, trusted data, such as that produced by national or international statistics institutes, think INSEE or EuroStat, and possibly other sources; high-profile recent data journalism efforts include the Panama Papers analysis (https://offshoreleaks.icij.org/, https://panamapapers.icij.org/20160403-panama-papers-global-overview.html).

The success of both data journalism and fact checking depends on the availability of flexible and efficient data management tools, capable of taking advantage of rich and heterogeneous data sources, such as: text, data graphs, relational data (tables), structured documents (JSON) etc. Each such type of data requires some technical knowledge and skills to query; using all of them together is extremely difficult today for journalists. The same issues are also encountered by domain experts (not journalists) familiar with their area (e.g. a branch of economy, or air pollution etc.) which need to work with many sources of heterogeneous data.

The goal of the project is to devise keyword search techniques across a set of heterogeneous data sources, deploy and validate them in collaboration with the Les Décodeurs team from the Le Monde newspaper. We collaborate with them within the ANR ContentCheck project (https://team.inria.fr/cedar/contentcheck/), and have identified an application scenario related to the analysis of incoming news based on the Decodex plugin produced by Les Décodeurs.

The topic is related to the areas of data management, information retrieval, data integration and query optimization.
The project will be supervised by Ioana Manolescu for the scientific aspects and guided by the application provided by Le Monde (reference people: Samuel Laurent and Adrien Vaudano).

References

[1] Efficient Keyword Search Across Heterogeneous Relational Databases, M. Sayyadian, H. LeKhac, A. Doan and L. Gravano, IEEE International Conference on Data Engineering, 2007

[2] Mixed-instance querying: a lightweight integration architecture for data journalism Raphaël Bonaque, Tien Cao, Bogdan Cautis, François Goasdoué, Javier Letelier, Ioana Manolescu, Oscar Mendoza, Swen Ribeiro, Xavier Tannier, Michaël Thomazo *VLDB*, Sep 2016, New Delhi, India. Very Large Databases Conference, 2016

Project 4. (Ioana Manolescu, LIX) Efficient summarization of large data graphs

Many high-value data collections are naturally organized as directed labeled graphs; in particular, Linked Open Data sources (some of which are listed at http://lod-cloud.net/) are typically organized in the Resource Description Format model, itself a variant of labeled directed graphs. Information about companies, stakeholders, and owners, published by the International Consortium of Investigative

Journalism (ICIJ), the organization behind the Panama Papers series, is another natural example of data graph.

Manipulating and exploring large graphs is challenging for users due to the high complexity and heterogeneity of the data. To help with this, different algorithms have been devised to summarize (compress) the structure of such labeled graphs into smaller, more compact ones, which serve as representative structures, easier to grasp than the original graphs.

We are interested in particular in summaries defined as quotient graphs: based on an initial graph and an equivalence relation that holds between the graph nodes, the summary is the quotient of the original graph by the given equivalence relation.

Such summarization frameworks have been introduced for semistructured data graphs in [1,2] and more recently adapted to semantic graphs in [3]; several equivalence notions can be defined, resulting in summaries that are more or less compact. The complexity of building the summaries also varies, depending on the chosen equivalence notion.

The purpose of the project is to propose two classes of algorithms to make quotient-based summarization more efficient:

a. Incremental algorithms, whereas given a summary S(G) of a graph G and a modification Delta brought to G, we seek to compute S(G+Delta) based on S(G) and Delta, without re-summarizing the graph G;

b. Parallel algorithms, whereas we seek to compute the summary S(G) of a graph G using a parallel computation framework.

References

[1] Roy Goldman, Jennifer Widom: "Dataguides: Enabling Query Formulation and Optimization in Semistructured Databases", VLDB 1997
http://infolab.stanford.edu/lore/pubs/dataguide_vldb97.pdf

[2] Chen, Q., Lim, A., Ong, K.W.: "D(K)-index: An adaptivestructural summary for graph-structured data", SIGMOD 2003
https://pdfs.semanticscholar.org/f5b2/9e844b2ed9440053863ee8de0dccbb1d8c34.pdf

[3] A Framework for Efficient Representative Summarization of RDF Graphs. Šejla Čebirić, François Goasdoué, Ioana Manolescu Research Report 9090, Inria Saclay Ile de France; Ecole Polytechnique,; Université de Rennes 1 [UR1]. 2017, pp.11

Project 5. (Yanlei Diao, LIX) Algorithms for Augmenting Text Documents with Images

Full description:

Imagine that you are an editor. You first write an article in text. Before publishing, you would like to insert several images from a large image database such as ImageNet [1] to make the article interesting to the reader. For example, the article that you are writing is about lavender fields in City A. It would be interesting to insert images about lavender fields that were taken from City A, and perhaps also oil and soap made from such lavender, tourism in City A, etc. Given the variety of articles that an editor may write, a manual process to retrieve such images is tedious and time-consuming.

In this project, we aim to design a software tool with intelligent algorithms for automating the processing of augmenting text documents with images. We have two specific goals: (1) The images retrieved must match the topic of a given article that an editor is editing. (2) Often times retrieving images only based on the

topic of the document is not sufficient. The editor is likely to have her own interpretation of whether an image is relevant or not. Therefore, we woud like to take an "Explore-by-Example" approach [2,3]: the algorithm should first recommend a few images based on the topic, and solicit the user feedback on the images. Then the algorithm should incorporate such feedback to adjust the internal model of the editor's interest, and retrieve a few more images for feedback. This process goes in iterations. We would like to design an algorithm that requires a minimum number of iterations to return a specified number of images that the editor deems relevant and interesting.

References:

[1] ImageNet. http://image-net.org/

[2] Yanlei Diao, Kyriaki Dimitriadou, Zhan Li, Wenzhao Liu, Olga Papaemmanouil, Kemi Peng, and Liping Peng. AIDE: an automatic user navigation system for interactive data exploration. PVLDB, 8(12):1964–1967, 2015.

[3] Kyriaki Dimitriadou, Olga Papaemmanouil, and Yanlei Diao. Explore-by-example: an automatic query steering framework for interactive data exploration. In SIGMOD Conference, pages 517–528, 2014.

Contact:

Prof. Yanlei Diao (Ecole Polytechnique), yanlei.diao@polytechnique.edu

Project 6. (Yanlei Diao, LIX) Insight Discovery from Knowledge Bases

Full description:

Ontologies and knowledge bases such as WordNet [1], Yago [2], and the Google Knowledge Graph contain not only known facts but also rules that encode human knowledge and can be used to derive additional (implicit) facts. They have been highly valuable resources for many tasks including question answering (Siri), information retrieval, or providing structured knowledge to users.

Recently, there is a growing demand to derive interesting "insights" [3] from such knowledge bases in an automated fashion. For example, over a knowledge base on scientific publications, an insight may be that certain topics have gained a significant number of submissions in recent years while the acceptance rates have remained low, hence revealing the possible issue of low quality submissions on these topics. As an another example, over a knowledge base about online retailing, an insight could be that while the total sales of the products in the sports category have declined last year in a district, the sales of a particular product in that district have increased sharply. An automated way to extract such insights from knowledge bases will be crucial to a wide range of applications across business, government, and society.

The goal of this project is to develop a formal definiton of "insights", efficient algorithms to extract such insights from large knowledge bases (based on both existing and derived facts), and evaluation methods to demonstrate that these insights indeed bring interesting and useful information to the end user.

References:

[1] G.A. Miller. WordNet: A Lexical Database for English. Communications of the ACM, 1995.

[2] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: a core of semantic knowledge. In Proceedings of the 16th international conference on World Wide Web, 2007.

[3] Bo Tang, Shi Han, Man Lung Yiu, Rui Ding, Dongmei Zhang. Extracting Top-K Insights from Multi-dimensional Data. Proceedings of the ACM Conference on Management of Data (SIGMOD), Chicago, Illinois, USA, May 2017.

Contacts:

Prof. Yanlei Diao (Ecole Polytechnique), yanlei.diao@polytechnique.edu

Project 7. (Yanlei Diao, LIX) Predicating and Explaining Interesting Patterns in Real-time Stream Analytics

Full description:

Recent applications such as the Internet of Things, data center monitoring, and market trend analysis present a pressing demand for perpetual, low-latency analytics to support a wide range of time-critical tasks and decisions. Today's data stream systems (e.g. [2]) support passive monitoring by requesting the monitoring application (or user) to explicitly define patterns of interest. However, a growing number of applications demand a new service beyond passive monitoring, that is, the ability of the monitoring system to automatically identify interesting patterns (including anomalous behaviors), produce a concrete explanation for the anomalies from the raw data, and based on the explanation enable a user action to prevent or remedy the effect of the anomaly, or to develop better strategies in the future. This task presents a more ambitious goal than what the state of the art [1] can support.

Toward this goal, our research explores the following research directions, in collaboration with real-world enterpreneurs.

1. Feature Generation. An explanation for the anomaly must be built on appropriate features, and such features may not exist int the raw event streams. We would like to explore a few alternative methods for feature generation, including (a) a general framework for feature extraction, such as deep learning; (b) custom feature extraction techniques, such as window-based frequent itemsets, or window-based aggregates.

2. Finding Explanations. When we have a good feature space, the next problem is how to build a model on a subset of the features that can best "explain" the anomalous instances. While regression models and decision trees can be used to build such a model, a good model for explaining anomalies must satisfy a broader set of requirements: (a) high accuracy of the model, as captured by the prediction error rate when the model is applied to a test dataset; (b) a compact, human-readable explanation: the model needs to be as compact as possible, i.e., using the minimum number of features among redundant or correlated features, and in a human-readable format, hence providing valuable insights to the user and increasing user confidence in triggering critical actions in demand-response management.

References:

[1] Manish Gupta, Jing Gao, Charu C. Aggarwal, and Jiawei Han. Outlier detection for temporal data: A survey. IEEE Trans. Knowl. Data Eng., 26(9):2250–2267, 2014.

[2] Haopeng Zhang, Yanlei Diao, and Neil Immerman. On complexity and optimization of expensive queries in complex event processing. In SIGMOD Conference, pages 217–228, 2014.

Contact:

Prof. Yanlei Diao (Ecole Polytechnique), yanlei.diao@polytechnique.edu

Project 11. (Frank Nielsen, LIX) Geometric methods for deep learning

Full description:

Deep learning consists in training neural networks built by stacking up many layers [1].
We are interested in stochastic neural networks where the set of parameters can be
interpreted as a point lying on a neuromanifold: learning is then interpreted as a trajectory on this manifold.
We shall build efficient and intrinsic supervised and unsupervised deep learning algorithms [2].

References:
-[1] http://www.deeplearningbook.org/

-[2] "Relative Natural Gradient for Learning Large Complex Models", ICML 2017

Contact:

Frank Nielsen (nielsen@lix.polytechnique.fr)

Project 12. (Frank Nielsen, LIX) Learning generative models

Full description:

Deep learning consists in training neural networks built by stacking up many layers [1].
We are interested in stochastic neural networks where the set of parameters can be
interpreted as a point lying on a neuromanifold: learning is then interpreted as a trajectory on this manifold.
We shall build efficient and intrinsic supervised and unsupervised deep learning algorithms [2].

References:
-[1] http://www.deeplearningbook.org/

-[2] "Relative Natural Gradient for Learning Large Complex Models", ICML 2017

Contact:

Frank Nielsen (nielsen@lix.polytechnique.fr)

Project 13. (Leo Liberti, LIX) Bootstrapping ontologies

Full description:

By "ontology" I mean here a body of knowledge organized in the form of a data structure, such as a graph. The nodes represent words or concepts, are labeled by denotations (semantic attributes: whether the concept has the property P or not for a set of "natural language" properties P). The edges, which could be (multiple) arcs and loops, are labelled by connotations (again semantic attributes, which establish a contiguity or other relation between the two adjacent concepts). Altogether, an ontology provides a context that allows the definition of a sound semantics (and possibly pragmatics) of a given text or corpus. Just to make an example, the WordNet graph is a very simple ontology, where the nodes are general words from the English language, their attributes are synonyms and antonyms, and the edges denote semantic contiguity with other words related by meaning.

Usually, ontologies are put together by hand, or by supervised Machine Learning (ML) from a training corpus annotated by hand. The issue is that creating an ontology "by hand" requires too much work for

most specific purposes. On the other hand, an existing low-quality ontology could be corrected (instead of generated) by hand with much less work. It occurred to me that a judicious alternation of unsupervised and supervised ML could be used to obtain a specific ontology automatically directly from a given (non-annotated) corpus. The goal of this project is to test my idea in practice. At most two students can work together on this project, which will be evaluated mostly on the bases of the success of the working software prototype.

References:

[1] G.A. Miller. WordNet: A Lexical Database for English. Communications of the ACM, 1995.

Contacts:

Prof. Leo Liberti (CNRS & Ecole Polytechnique), liberti@lix.polytechnique.fr