Applications of Data Graphs

Theses of the Ph. D. dissertation

Gábor Rácz

Supervisor: Dr. Attila Kiss

Eötvös Loránd University
Faculty of Informatics
Department of Information Systems

Ph. D. School of Computer Science
Head of school: Prof. Erzsébet Csuhaj-Varjú
Ph. D. Program of Information Systems
Head of program: Prof. András Benczúr

Budapest, 2018

Intoduction

Internet became part of our everyday life. We read news, listen music, contact our friends on web pages. Forums, on-line social networks, and web pages are available from anywhere so there is no more physical limit on contacts between people. We can meet new things and new peoples that have impact on our thinking.

Parallel to transformation and growth of Web, science and technology is evolving. More sophisticated and more precise measuring instruments are created by which we can observe more closely the world around us. In addition, computer simulation plays important role in formulating theoretical models and in controlling and evaluating experimental results.

Due to these processes, the amount of available data has drastically increased over a short period of time. Storage and processing of the data are serious challenges for engineers and researchers. As a result, new storage models were developed besides relational databases. Some of the new models limited the available operations to increase efficiency; others are specialized to areas such as storing semi structured documents or graphs.

In this dissertation data graphs are examined from several aspects. Data graphs are directed, labeled graphs whose vertices represent concepts or objects, and edges describe relationships among them. My theses are grouped around three main topics and each of these topics is related in some way to data graphs.

The focus of the first topic is on Semantic Web technologies that are designed to link data available on the Web. For this purpose, a framework was created that allows heterogeneous data from various sources to be integrated and allow us to describe our knowledge about the world. The data sets given in this framework can be represented as directed, labeled graph and can be stored as data graphs. For querying the data sets, a declarative pattern-based language was developed.

Concerning this area we explore how to exploit the benefits of visualization in understanding and in characterizing data sets. We show what options and challenges are offered by the structure of the new framework in displaying the data. In addition, we review the advantages and disadvantages of visual query languages.

The second main topic is the semantic matching in which we have to select the best matching elements among many. Various matching functions are introduced that make it possible to decide how much the elements fit together. For definition of functions, another Semantic Web technology, namely the ontologies, is used. Ontologies provide vocabularies to describe the most important concepts of given domains, and to organize

the concepts into hierarchies and groups. They can be described in the above mentioned framework and so can be represented as graphs. We present probability model based matching functions as well.

The third main topic is related to the analysis of social networks that makes it easier to understand the structure of our society and people's behavior. We examine the most influential set of nodes of the social networks through which the largest part of the network can be accessed and influenced. Greedy algorithms are introduced to find the most influential set of nodes. We also present an extension for RapidMiner data mining software which allows us to use network analyzing algorithms as part of data mining tasks. Moreover, we show how Semantic Web technologies can be used for examining social networks.

Theses

1 Visualization and Semantic Web

The purpose of the Semantic Web [1] is to link data available on the Web to create the *web of data*. Various technologies were developed to achieve this goal. In Resource Description Framework (RDF) data can be described in the form of (subject, predicate, object) or (subject, property, value). Vocabularies and ontologies are available for collecting the main concepts of given domains and for organizing those concepts into hierarchies and groups. SPARQL is a declarative query language for querying data stored in RDF format.

RDF triplets can be represented by directed, labeled edges where subject and object are the start and end points of the edge and the predicate is the label. Thus, RDF data sets form directed, labeled graphs.

Thesis 1.1 ([12]). *A node simulation relation has been introduced, based on which node merging rules have been defined on RDF graphs. Using these rules, the graphs can be simplified in a way that structure of edges belonging to predefined predicates can be preserved, and only the cardinality of the connected nodes are retained. Our experimental results have shown that this method helps detecting failure and deficiency of data sets.*

My part of this work was mainly the elaboration of the node simulation rules, while the implementation of the rules in a distributed computation model was done by my colleagues.

It has been demonstrated how to generate customized content by SPARQL query transformations via prototype of a mobile application. *Tabular views* can be assigned to

users that allow them to browse the data set for different topics. Entities appearing in the rows of the tables can be selected by SPARQL query and the columns contain properties of the entities. When a row is selected then a *detailed view* appears with the details of that particular entity. Content of the detailed view is generated by extending the original query with OPTIONAL blocks that are defined in the view. Additionally, filters can be assigned to columns of tabular views that are converted into FILTER blocks in the query.

SPARQL queries are based on pattern matching where patterns can be given in a form similar to RDF triples but any position of the triples may have variables. Patterns can specify mandatory or optional (OPTIONAL) matching. Results of a query are the matching elements in the form of variable substitutions whose values can further be restricted by FILTER expression. Analogously to the RDF triplets, the patterns can be represented by directed, labeled graph as well. These graphs are usually small therefore they can be easily visualized.

Thesis 1.2 ([**11, 4**])**.** *Benefits and drawbacks of visual languages against textual ones have been collected. It has been shown by the example of VisualQuery, how syntactic errors can be prevented by built-in editing rules in a visual query tool for SPARQL. Then, different aspects have been identified to compare visual query languages for SPARQL.*

2 Semantic matching[1]

In matching tasks, we need to find the most similar entities among many. Entities are described with profiles based on their properties and we look for common elements when pairing. If the properties are chosen from a given semi-structured set of concepts then we are talking about semantic matching.

There are various use cases of matching task, for example, dating applications and on-line product searching tools. We have focused on the human resources area, namely we have been looking for the most suitable candidates for a job advertisement. After selecting the matching function, finding the most suitable candidates can be done with a simple maximum search.

We have presented set-based and probability-model-base approaches to define matching functions. The set-based solutions were inspired by the perfect matching [10] that

[1]The research reported in this section has been supported by the Austrian Ministry for Transport, Innovation and Technology, the Federal Ministry of Science, Research and Economy, and the Province of Upper Austria in the frame of the COMET center SCCH.
This chapter is a result of cooperation of Alfréd Rényi Institute of Mathematics and Software Competence Center Hagenberg.

takes the intersection of the profiles of the job offer and the application and then computes its cardinality normalized by the cardinality of the job offer.

Thesis 2.1 ([**9, 13**])**.** *Additional information has been added to the system to increase the selectivity of the perfect matching model: the set of the properties has been structured. A lattice has been constructed from the properties based on general concept inclusions of field-specific ontologies. Then, the profiles have been extended to filters by adding more general properties to them using the lattice. The perfect matching could be used to the resulted filters. After that, the system has been expanded again with additional information in form of weighted extra edges. As the extra edges may result in circles in the lattice, they have been modeled with directed, labeled graphs. Considering the original and the extra edges, fuzzy filters [8] have been generated from the profiles and the perfect matching has been generalized to fuzzy sets.*

Thesis 2.2 ([**13**])**.** *Extra edges can be considered as a kind of conditional probability between properties, so matching function can be defined using discrete probability models [15]. We have shown how nodes can be converted into probability events and how edges can be converted into probability statements. Discrete probability models have been defined over the probability events which satisfy the probability statements. Then, a maximum entropy problem has been constructed from the statements and the profiles have been transformed into probability events as well. From the profiles of the job offer and the application, a probability query has been generated which gives the conditional probability of the job offer given the application. The result of the probability query has been chosen as the matching value. It can be computed using the result of the maximum entropy problem.*

When the presented matching functions have been compared, we have found that each function can assign different values to the applications and can sort them into different order. Therefore, it depends on the specific task, which function gives the best result.

I have followed Attila Sali's suggestions during construction and comparison of the presented matching functions.

3 Analysis of Social Networks

On-line social network became part of our everyday live: we share pictures, videos, exchange thoughts, make friends through them. We can better understand our society and people's behavior by investigating these networks

Important nodes of social networks have been investigated. A node can be important from several points of view; we have focused on the influence of nodes. From this point of

view, a node is important if the largest part of the network can be influenced through it.

Influence maximization problem [5] is about to find a fixed size subset of nodes which has the maximum influence on the network among the similar size subset of nodes. The spread of influence is simulated by a series of discrete steps according to *Independent Cascade* model, and the influence of a subset of nodes corresponds to the influenced node. The task turned out to be NP-hard, so approximative algorithms were developed to solve it.

NewGreedy [3] is a greedy algorithm that starts from an empty set, and then adds nodes to it iteratively. One node is added to the set in each iteration; the node that maximizes the influence gain of the candidate set. It has been proven in the same paper that the result of the procedure is at least as good as $(1-1/e)$ times the optimal solution.

Thesis 3.1 (**[7]**). *A modified version of the* NewGreedy *has been presented, called* SelingerGreedy*, that uses Selinger-style optimization [16] to improve the accuracy of the algorithm. The basic idea of the optimization is to keep the second, third, ...,* b^{th} *best solution in each iteration. This will increase the search space but it helps avoiding local extreme values. Our experimental results have shown that the fifteen best candidate sets have near the same influence as the one given by* NewGreedy*. Based on the results we have concluded that in practice* NewGreedy *approximates the optimal solution better than the theoretical boundary. We suspect that the algorithm can choose from several nodes with similar properties and the choice has no real effect on the final result. That means that candidate sets given by* SelingerGreedy *are as good as the result of* NewGreedy*.*

Community-based Greedy Algorithm (*CGA*) [17] was proposed to decrease the running time of *NewGreedy*. It is based on the assumption that social networks consist of smaller communities and influence of a node in its community is approximately the same as the node's influence within the whole network. Therefore *CGA* detects the communities as a first step, and then chooses the candidates from the most influential nodes of the communities.

Thesis 3.2 (**[14]**). *We have modified* CGA*, replaced the community detection step with the Louvain method [2]. That is a modularity-based agglomerative community detection method. The modified algorithm is called* LouvainGreedy*. Our experimental results have shown that in some cases* LouvainGreedy *is up to ten time faster than* NewGreedy *and three time faster then* CGA *and the precision loss is less than three percent. The reason of this performance improvement is the differently identified communities. As the edges between communities are not taken into account in the candidate selection phase, they*

have impact on the accuracy. The computation time of the most influential nodes of communities highly depends on the size of the communities.

Network analysis can be combined with the analysis of the user-generated and shared content to get more complete picture about the users themselves and their relationships.

Thesis 3.3 (**[6]**). *We have developed an extension for RapidMiner data mining software. The extension contains important algorithms for network analysis and for node characterization. These algorithms can be used to solve complex data mining problems. As an illustration, we used the extension to compute whether the posts of the most influential Twitter users have negative or positive sentiment.*

I mostly contributed to the analysis of *SelingerGreedy* and *LouvainGreedy*, to the preparation of examples, and to the making of measurments.

Finally, I have collected how the Semantic Web technologies can be used to solve the most common problems of social network analysis.

Related publications of the author

[4] G. Gombos, T. Matuszka, B. Pinczel, G. Rácz, and A. Kiss. "VOSD: A General-Purpose Virtual Observatory over Semantic Databases". In: *Acta Cybernetica* 21.3 (2014), pp. 353–366. DOI: 10.14232/actacyb.21.3.2014.5.

[6] B. Kósa, M. Balassi, P. Englert, G. Rácz, Z. Pusztai, and A. Kiss. "A Basic Network Analytic Package for RapidMiner". In: *Proceedings of the 5th Rapidminer World*. Shaker Verlag GmbH. 2014, pp. 47–59. ISBN: 978-3-8440-2946-8.

[7] B. Kósa, B. Pinczel, G. Rácz, and A. Kiss. "Properties of the most influential social sensors". In: *Cognitive Infocommunications (CogInfoCom), 2013 IEEE 4th International Conference on*. IEEE. 2013, pp. 469–474. DOI: 10.1109/CogInfoCom.2013.6719293.

[9] J. Martinez-Gil, L. Paoletti, G. Rácz, A. Sali, and K.-D. Schewe. "Maintenance of Profile Matchings in Knowledge Bases". In: *International Conference on Model and Data Engineering*. Vol. 9893. LNCS. Springer. 2016, pp. 132–141. DOI: 10.1007/978-3-319-45547-1_11.

[11] G. Rácz. "Vizuális lekérdezésszerkesztő program a SPARQL nyelvhez". MA thesis. Budapest: ELTE, 2012.

[12] G. Rácz, G. Gombos, and A. Kiss. "Visualization of Semantic Data Based on Selected Predicates". In: *Transactions on Computational Collective Intelligence XIV*. Ed. by N. T. Nguyen. Vol. 8615. LNCS. Springer Berlin Heidelberg, 2014, pp. 180–195. DOI: 10.1007/978-3-662-44509-9_9.

[13] G. Rácz, A. Sali, and K.-D. Schewe. "Semantic Matching Strategies for Job Recruitment: A Comparison of New and Known Approaches". In: *International Symposium on Foundations of Information and Knowledge Systems*. Vol. 9616. LNCS. Springer. 2016, pp. 149–168. DOI: 10.1007/978-3-319-30024-5_9.

[14] G. Rácz, Z. Pusztai, B. Kósa, and A. Kiss. "An improved Community-based Greedy algorithm for solving the influence maximization problem in social networks". In: *Annales Mathematicae et Informaticae* 44 (2015), pp. 141–150.

Further references

[1] T. Berners-Lee, J. Hendler, O. Lassila, et al. "The semantic web". In: *Scientific american* 284.5 (2001), pp. 28–37.

[2] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. "Fast unfolding of communities in large networks". In: *Journal of statistical mechanics: theory and experiment* 2008.10 (2008), P10008.

[3] W. Chen, Y. Wang, and S. Yang. "Efficient influence maximization in social networks". In: *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2009, pp. 199–208.

[5] D. Kempe, J. Kleinberg, and É. Tardos. "Maximizing the spread of influence through a social network". In: *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2003, pp. 137–146.

[8] L. Liu and K. Li. "Fuzzy filters of BL-algebras". In: *Information Sciences* 173.1 (2005), pp. 141–154.

[10] N Popov and T Jebelean. *Semantic matching for job search engines: a logical approach*. Tech. rep. Technical report 13–02, Research Institute for Symbolic Computation, JKU Linz, 2013.

[15] M. Schramm and M. Greiner. *Non-monotonic reasoning on probability models: indifference, independence & maxent*. 1995.

[16] P. G. Selinger, M. M. Astrahan, D. D. Chamberlin, R. A. Lorie, and T. G. Price. "Access path selection in a relational database management system". In: *Proceedings of the 1979 ACM SIGMOD international conference on Management of data*. ACM. 1979, pp. 23–34.

[17] Y. Wang, G. Cong, G. Song, and K. Xie. "Community-based greedy algorithm for mining top-k influential nodes in mobile social networks". In: *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2010, pp. 1039–1048.