Parsing techniques for graph analysis

Semyon Grigorev
Saint Petersburg State University
7/9 Universitetskaya nab.
St. Petersburg, 199034 Russia
Semen.Grigorev@jetbrains.com

Ekaterina Verbitskaia
Saint Petersburg State University
7/9 Universitetskaya nab.
St. Petersburg, 199034 Russia
kajigor@gmail.com

Nowadays input data for parsing algorithms are not limited to be linear strings, and context-free grammars are used not only for programming languages specification. One of classical examples is a context-free path querying for graph data bases where input is a graph and path constraints are specified by a grammar. Graph parsing may be applied in different areas, in software engeneering for dynamically generated strings analysis, graph data bases for paths querying, etc. For example, the idea of multiple input GLL parsing which was presented at Parsing@SLE-2016 by Elizabeth Scott and Adrian Johnstone, is an partial case of graph parsing: set of token-with-extent can be treated as directed graph where extents are vertices and tokens are labels of edges. So, it is great connection of different areas: formal languages, parsing algorithms, data bases, graph theory, etc.

There are some open questions. We are working on it. Solutions are CYK-based [**?**]. Effective algorithms and new felds for application.

Our current results: We have some experience in the areas mentioned above [3, 6]. GLL-based context-free path querying algorithm [3] implemented by the authors is faster than solution which was presented at ISWC-2016 [7]. We matrix multiplication [1] is faster than GLL-based, but can not build forest.

Currently we are working on Parser-combinators (meercat based, in progress, GitHub), GLL, (Okhotin-Valiant inspired), conjunctive grammars (in progress), mechanisation in coq (in progress).

We have some ideas of graph parsing applications. For example, context-free pattern search in metagenomical assemblies. , bioinformatics for metagenomic assemblies analysis Secondary structure can be specified in terms of grammar (context-free or conjunctive). Assembley is a graph.

Our plans are All existing applications seem to be special cases of the Bar-Hillel [2] theorem for context-free and regular language intersection, and can be generalized, but today many of them are developed as stand alone solutions. Thus, the goal of our work is to create an abstract framework for parsing based on generalization of GLL parsing algorithm [5] proposed by Elizabeth Scott and Adrian Johnstone.

We want to adopt advanced matrix multiplication techniques, such as approximated matrix multiplication, sparse matrix multiplication, for graph parsing. Also we want to , Is it possible Boolean grammars — problems with non-monotonic. Even for conjunctive grammars we get approhimation of result Anoter research direction is an effective algorithms intersection of other types, and finding of other types of grammars. One of possible start point is non-recursive context-free grammars intersection [8, 9] which can be used in speach recognition or for compressed strings porcrssing. We also want to investigate practical areas of application and to create solutions based on our framework to demonstrate its practical value.

1. REFERENCES

[1] Azimov, Rustam, and Semyon Grigorev. "Graph Parsing by Matrix Multiplication." *arXiv preprint arXiv:1707.01007* (2017).

[2] Bar-Hillel, Yehoshua, Micha Perles, and Eliahu Shamir. "On formal properties of simple phrase structure grammars." *Sprachtypologie und Universalienforschung* 14 (1961): 143-172.

[3] Grigorev, Semyon, and Anastasiya Ragozina. "Context-Free Path Querying with Structural Representation of Result." *arXiv preprint arXiv:1612.08872* (2016).

[4] Hellings, Jelle. "Querying for Paths in Graphs using Context-Free Path Queries." *arXiv preprint arXiv:1502.02242* (2015).

[5] Scott, Elizabeth, and Adrian Johnstone. "GLL parsing.", *Electronic Notes in Theoretical Computer Science*, 253.7 (2010): 177–189.

[6] Verbitskaia, Ekaterina, Semyon Grigorev, and Dmitry Avdyukhin. "Relaxed Parsing of Regular Approximations of String-Embedded Languages." *International Andrei Ershov Memorial Conference on Perspectives of System Informatics.* Springer International Publishing, 2015.

[7] Zhang, Xiaowang, et al. "Context-free path queries on RDF graphs." *International Semantic Web Conference.* Springer International Publishing, 2016. 632–648.

[8] Nederhof, Mark-Jan, and Giorgio Satta. "Parsing non-recursive context-free grammars." *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics.* Association for Computational Linguistics, 2002.

[9] Nederhof, Mark-Jan, and Giorgio Satta. "The language intersection problem for non-recursive context-free grammars." *Information and Computation* 192.2 (2004): 172-184.

[10] Yannakakis, Mihalis. "Graph-theoretic methods in database theory." *Proceedings of the ninth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems.* ACM, 1990.