On Massive JSON Data Model and Schema

Teng Lv¹, Ping Yan², * and Weimin He³

¹School of Information Engineering, Anhui Xinhua University, Hefei 230088, P. R. China
²School of Science, Anhui Agricultural University, Hefei 230036, P. R. China
³Department of Computing and New Media Technologies, University of Wisconsin-Stevens Point, 2100 Main Street, Stevens Point, WI 54481, USA
Email: Lt0410@163.com, want2fly2002@163.com(corresponding), whe@uwsp.edu

Abstract. JSON (JavaScript Object Notation) is a lightweight semi-structured data format based on the data types of programming language JavaScript. It is a popular data exchange format over the World Wide Web and becomes a dominant standard format for sending API (Application Programming Interface) requests and responses in the past few years. Furthermore, JSON has also attracted attention of database community research, especially in data intensive applications. JSON is not only can be integrated in traditional database systems, but also widely used in NoSQL database systems and graph database systems. Compared with XML, JSON document is a set of "key-value" pairs, in which the "value" itself can be a JSON document, which allows arbitrary levels of nesting, so it is more flexible to use and more difficult to process accordingly. JSON data model and schema describe the basic data structures and semantics of the underlying JSON data, so it is the fundamental and key aspects for JSON data format. JSON data model and schema are not only foundations for other data management technologies, such as data indexing, data querying, data searching, data mapping, data integrating, and data mining, but also has important theoretical significance and application prospects to provide theoretical basis and technical means for other related research, such as data integration, data conversion and other semi-structured and unstructured data queries. This paper analyzes the key problems of JSON data model and schema, including what data model should adopted by JSON and the specification and schema outline of JSON model.

1. Introduction

The amount of data generated by human society is increasing exponentially. According to IDC monitoring, the amount of global data doubles every two years, and human society has entered the era of "big data". Big data is an infrastructure carrying economic and social development. Big data is not only large in volume, but also has increasingly complexity in structure. In addition to the traditional structured data, a large number of semi-structured and unstructured data also show explosive growth, such as semi-structured data representation and exchange language XML (Extensible Markup Language) and JSON (JavaScript Object Notation).

JSON [1-3] is a data format based on data types in JavaScript. Because JSON is readable and easy to be understood and processed by computers, JSON is very popular in the Web developer community. It is a main format for information exchange on the Web and becomes a dominant standard format for sending API (Application Programming Interface) requests and responses. In addition, in the field of data management, especially in data-intensive applications, JSON is often used to construct...
NoSQL-based database systems [4-6] and graph database systems [7], and is often integrated into relational database systems. It is a common agreement on the wide acceptance by more and more database products. Compared with XML, JSON document is a set of "key-value" pairs, in which the "value" itself can be a JSON document, which allows arbitrary levels of nesting, so it is more flexible to use and more difficult to process accordingly.

Although the large-scale data represented by JSON provides data resources for data analysis and data mining, which enables us to gain unprecedented insight into data, the cost of processing and querying large-scale JSON data is often very high. Because of the huge amount of data, the traditional query methods often face many challenges, especially the response time of query is too long to meet the needs of users [8]. In many cases, it is often impossible for users to obtain accurate query results in a shorter query response time [9]. Especially in real-time systems or online processing systems, it is more meaningful to get an approximate query result in a short time than to get an accurate result in a longer time [10].

Although JSON has been very popular and widely used in many fields, there is no consensus on many basic scientific issues of JSON data in academic communities, and the research on some issues has just begun. Research on massive JSON data model and schema can effectively improve query effect and user's query experience. It has important theoretical significance and application prospects to provide theoretical basis and technical means for other related research, such as data integration, data conversion and other semi-structured and unstructured data queries.

The rest of the paper is organized as follows. Section 2 discusses what data model should JSON adopt, Section 3 discusses the specification and schema of JSON data, Section 4 discusses the schema summarization of JSON data, and we conclude the paper and point out the future directions of the topic in Section 5.

2. JSON Data Representation and Model
One natural idea is to use the current XML data model to represent JSON data. But compared with XML, JSON has two main differences:

First, JSON allows arbitrary levels of nesting, which results in the rich types of "value" in the key-value pairs, including both simple types and composite types (called objects). This makes the XML data model unable to process JSON nesting efficiently.

Second, JSON is deterministic, that is, for every object in JSON document, all key-value pairs are unique. This suggests that we may design more efficient JSON models and process JSON data according to the characteristics of determinacy.

At present, the representative research of JSON data model is mainly based on tree structure [11, 12]. The main idea is to use edges in the tree to represent the keys of the object, leaf nodes to represent atomic values, subtrees to represent composite types. And for array types, we can use numbers to express the sequentiality of elements in the arrays.

However, the shortcomings of the current research are that they all simplify the actual JSON specification, and consider a set of very limited data types and corresponding modifier keywords. In addition, the comparison of subtrees involved in nesting [13] and the complex relationship with deterministic have not been discussed. Therefore, it is of great theoretical and practical significance to redesign new data models and data type representations specifically for the characteristics of JSON, such as nesting and determinacy. A more detailed analysis can be found in Ref. [14].

3. JSON Schema
The earlier representative JSON schema [15] is based on the tree structure of JSON. It gives the basic grammar specification and the preliminary semantic description, such as satisfiability, correctness verification, etc., but lacks other more complex semantic constraints, such as integrity constraints, and such schema cannot be updated with the dynamic updating of JSON (time-varying characteristic).

Furthermore, for large-scale JSON data in practical applications, the complexity of processing is not only reflected in the large amount of data to be stored, but also in the complexity and time-varying characteristic of data schemas. If such a large-scale JSON data is to be queried, it is necessary to learn and master the whole complex schema beforehand.
In this case, if the system can provide a summarization of schema information, it will be very helpful to the user's query. This is the basic idea of schema summarization, which provides a concise overview of the whole database schema, so that a query only needs to focus on those related partial schemas without paying attention to all the schemas. This leads to the research of schema summarization of massive JSON data.

4. JSON Schema Summarization

For schema summarizations of relational databases and XML data, there have been many studies, such as Ref.[16,17] focused only on the attributes of schemas without considering the distribution of data. Recent studies have taken both into account, such as XML schema summarizations [18, 19], schema summarizations of relational databases [20], and application of schema summarizations in keyword search [21] and recommendation systems [22]. Other research is concentrated on multi-level summarization extraction of large-scale database based on clustering [23].

The above researches only consider the summarization itself or the data distribution, and do not consider the semantic information on the summarization. At present, there is no research on JSON schema summary. The only research work is mostly focused on schema reasoning, including how to deduce its schema [24], the data types [25] and its quantitative relationships [26] from existing JSON data. But these are far from the research goal of schema summarization.

5. Conclusions

This paper presented a survey of large-scale JSON data model, schema, and schema summarization. We proposed related 3 key problems of large-scale JSON data, including what data model should be adopted by JSON, the specification and schema of JSON data, and schema summarization of massive JSON data. We analyzed the advantages and disadvantages of the state of art ideas, methods, and technologies to deal with the above problems. We also gave some suggestions and directions to deal with the above problems, which will be useful to direct future research, such as data integrity, data exchange, database design, data query, etc.

6. Acknowledgments

The work is supported by Talent of Discipline and Specialty in Colleges and Universities of Anhui Province (No.gxbjZD54), Academic Leader Foundation (No.2014XXK06), Anhui Province Quality Engineering (No. 2017ghjc229, No.2015zy073), Introduction of Talents Foundation of Anhui Xinhua University(No.2015kyqd002), and National Nature Science Foundation of China (No.11201002).

7. References

[1] Bray T 2014 The JavaScript Object Notation (JSON) Data Interchange Format. https://datatracker.ietf.org/doc/rfc7159/
[2] ECMA 2013 The JSON Data Interchange Format. http://www.ecma-international.org/publications/standards/ Ecma-404.htm.
[3] Internet Engineering Task Force (IETF) 2104. The JavaScript Object Notation (JSON) Data Interchange Format. https://tools.ietf.org/html/rfc7159
[4] ArangoDB Inc 2016 The ArangoDB database. https://www.arangodb.com/
[5] MongoDB Inc 2015 The MongoDB3.0 Manual. https://docs.mongodb.org/manual/
[6] OrientDB LTD 2016 The OrientDB database. http://orientdb.com/
[7] The Neo4j Team 2016 The Neo4j Manual v3.0. http://neo4j.com/docs/stable/
[8] Fan W and Huai J P 2014 Querying big data: Bridging theory and practice. Journal of Computer Science and technology, 29(5): 849-869
[9] Mozafari B, Ramnarayan J, Menon S, et al 2017 Snappydata: A unified cluster for streaming, transactions, and interactive analytics. CIDR’17, January 8-11, 2017, Chaminade, California, USA.
[10] Mozafari B 2017 Approximate Query Engines: Commercial Challenges and Research Opportunities. SIGMOD’17, ACM, pp.521-524.
[11] Pierre B, and Reutter J L 2017 JSON: Data model, Query languages and Schema specification. *PODS’17*, ACM, pp 123-135
[12] Jan H, Paredaens J, and Jan V D B 2017 J-Logic: Logical Foundations for JSON Querying. *SIGMOD’17*, ACM, pp 137-149
[13] Wong K 2007 Unranked tree automata with sibling equalities and disequalities. *International Conference on Automata, Languages and Programming*, Springer-Verlag, pp 875-887
[14] Lv T, Yan P, and He W 2018 Survey on JSON data modelling. *Journal of Physics: Conference Series (JPCS)*, Volume 1069, pp. 683-687.
[15] Felipe P, Reutter J L, Suarez F, and Vrgoč D 2016 Foundations of JSON Schema. *International World Wide Web Conferences Steering Committee*, pp 263-273
[16] Castano S, Antonellis V D, Fugini M, et al 1998 Conceptual schema analysis: techniques and applications. *TODS’98*, 23(3): 286-332
[17] Moody D and Filtman A 1999 A methodology for clustering entity relationship models - a human information processing approach. *Proc. of ER’99*, pp. 114-130
[18] Yu C, and Jagadish H V 2006 Schema summarization. Proc. VLDB’06, pp. 319-330
[19] Lv T and Yan P 2013 A Framework of Summarizing XML Documents with Schemas. *International Arab Journal of Information Technology*, 10(1): 18-27
[20] Yang X, Procopiuc C M, and Srivastava D 2009 Summarizing relational databases. *PVLDB*, 2 (1): 634–645
[21] Fakas G J 2011 A novel keyword search paradigm in relational databases: Object summaries. *Data & Knowledge Engineering*, 70(2): 208-229
[22] Ragone A, Tomeo P, Magarelli C, et al 2017 Schema-summarization in linked-data-based feature selection for recommender systems. *Proceedings of the Symposium on Applied Computing (SAC’17)*, ACM, pp. 330-335, Marrakech, Morocco, April 3-7, 2017
[23] Wang X, Zhou X, and Wang S 2013 Hybrid Schema Summarization Method of Large Scale Database. *Chinese Journal of Computers*, 36(8): 1616-1625
[24] Baazizi M, Lahmar H B, Colazzo D, et al 2017 Schema Inference for Massive JSON Datasets. *Extending Database Technology (EDBT’17)*, Venice, Italy, Mar 2017
[25] Dario C, Ghelli G, and Sartiani C 2012 Typing Massive JSON Datasets. *International Workshop on Cross-model Language Design and Implementation (XLDI)*
[26] Baazizi M, Colazzo D, Ghelli G, et al 2017 Counting types for massive JSON datasets. *DBPL*, Munich, Germany, 2017