Fundamentals of Database Systems – PROJECTS (Modeling)

PGDBA, First Year, 2020–2022

Deadline: February 15, 2021

Total: 10 marks

SUBMISSION INSTRUCTIONS

1. Submit a solution sketch in a single file by the deadline. The solution must be self-explanatory.
2. The solution should include the sections (if applicable): Introduction, Related Work, Terminologies and Definitions, Theory, Methods, Results, Conclusion.
3. Include names and roll numbers of all of your group members (at most 5).
4. Naming convention for your submission file (assuming M is your project number): `projM` (.docx, .doc, .pdf, .tex, etc.).
5. To submit a solution file (say `projM.docx`), ensure that it is not password protected and mail to `<assignisik@gmail.com>` with the subject line as follows: `PGDBA 2020-22 Project M`.

NOTE: The contribution must be novel and non-trivial.

Project 6: **[Crowd Powered Databases]** Crowdsouring is a distributed approach of solving problems online by involving the crowd contributors, either as volunteers or in exchange of payments. In crowdsourcing databases, human operators are embedded into the database engine and they collaborate with the other conventional database operators to process the queries [1]. There are recent advances in developing various types of crowd-powered database functionalities. These include query answering capabilities [2] and normalization of databases [3]. You are required to propose a novel crowd-powered model to support any kind of database management operations.

[1] Sai Wu, Xiaoli Wang, Sheng Wang, Zhenjie Zhang, and Anthony KH Tung. "K-anonymity for crowdsourcing database," *IEEE Transactions on Knowledge and Data Engineering*, 26(9):2207-2221, 2014.

[2] Michael J. Franklin, Donald Kossmann, Tim Kraska, Sukriti Ramesh, and Reynold Xin. "CrowdDB: answering queries with crowdsourcing," In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data (SIGMOD)*, pp. 61-72, 2011. (Link: `http://nike.psu.edu/classes/ist501/2014-fall/ref/crowddb\_sigmod2011.pdf`)

[3] Sairam, G.A., Kolli, P., Immidisetty, A., Kumar, P., Sudhan B, M. and Bhattacharyya, M., 2021. Scalable Database Normalization Powered by the Crowd. In 8th ACM IKDD CODS and 26th COMAD (pp. 213-217). (Link: `https://dl.acm.org/doi/pdf/10.1145/3430984.3431032`)

Project 7: **[Data Visualization]** Visualization of data is considered to be an important area of Data Science. Data visualization symbolizes the efforts that help people in understanding the

significance of data in a better way through representing it in a visual context. The patterns, trends and correlations that might go undetected in raw data can be exposed and recognized effectively with data visualization methods [1]. There are plenty of data visualization approaches in the literature, starting from the basic ones like Venn diagrams, pie charts, boxplots, and scatter plots, and ranging upto the recent ones like alluvial diagrams and sunbursts [2]. You are required to recognize the limitations of the existing visualization approaches and suggest a novel data visualization method. Note that, modeling of your approach is suggested but deployment is not necessary.

[1] Cameron Chapman, "A Complete Overview of the Best Data Visualization Tools," *Toptal*, 2019. (Link: `https://www.toptal.com/designers/data-visualization/data-visualization-tool`

[2] RAWGraphs. (Link: `https://rawgraphs.io`)

Project 8: **[Model for NoSQL Databases]** Unlike relational databases, which are typically driven by the structure of available data, NoSQL data modeling often starts from the application-specific queries [1]. As with NoSQL approaches, we cannot model relations within the data, therefore, sometimes it is hard to visualize the database as structured collection of data.

Propose a novel data model for NoSQL databases for overcoming the said limitations. The model should be created with appropriate characterizations.

[1] NOSQL DATA MODELING TECHNIQUES. (Link: `https://highlyscalable.wordpress.com/2012/03/01/nosql-data-modeling-techniques`)

Project 9: **[Correctness of SQL Queries]** We often test the correctness of SQL queries by executing the query in question on some test database instance and compare its result with that of the correct query [1]. The problem of finding small counterexamples for different classes of queries, including those involving negation and aggregation, is in general known to be NP-hard. There are recent algorithms to address such problems [1]. Building user-friendly tools to learn and debug database queries is an interesting research direction. In particular, building a similar tool with the full functionality of SQL queries is a challenging open problem. Suggest a model for creating such a platform.

[1] Zhengjie Miao, Sudeepa Roy, and Jun Yang. "Explaining Wrong Queries Using Small Examples." In Proceedings of the 2019 International Conference on Management of Data, pp. 503-520. ACM, 2019. (Link: `https://dl.acm.org/citation.cfm?id=3319866`)

Project 10: **[Heterogeneous Information Network]** Existing representations of heterogeneous information networks face several challenges. A recent study proposes a new end-to-end neighborhood-based interaction model for recommendation to address these issues. It first analyzes the significance of learning interactions and then proposes a novel formulation to capture the interactive patterns between each pair of nodes through their metapath-guided neighborhoods.

You are required to propose a model better than the said one with a new perspective and enhanced functionalities.

[1] Jin, J., Qin, J., Fang, Y., Du, K., Zhang, W., Yu, Y., Zhang, Z. and Smola, A.J., 2020, August. An Efficient Neighborhood-based Interaction Model for Recommendation on

Heterogeneous Graph. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery Data Mining (pp. 75-84). (Link: https://dl.acm.org/doi/pdf/10.1145/3394486.3403050)