Proposta de Estudo Orientado

Pós-Graduação em Computação – IC/UFF

Explorando Estratégias de Construção de Consultas em Bases de Dados com Interfaces em Linguagem Natural

Aluno: Henrique Bueno Rodrigues (D022.118.009)
Orientadores: Alexandre Plastino e Aline Paes
Área de Concentração: Engenharia de Sistemas e Informação
Linha de Pesquisa: Inteligência Artificial

1. Descrição do Tema

Apesar do surgimento de diversas tecnologias de bancos de dados não convencionais, por exemplo, NoSQL [32] e Hadoop [33], sistemas de corporações pequenas, médias e grandes ainda possuem enorme acoplamento e dependência de bancos de dados relacionais, tais como Oracle [30] e SQL Server [31].

A contínua valorização das informações no apoio aos processos de tomada de decisão aumenta a demanda pela disponibilidade de dados, a chamada Democratização da Informação [18], e incentiva as áreas de tecnologia da informação das empresas a pensarem e criarem novas estratégias mais simples e ágeis para a interface entre usuários finais e as bases de dados.

Basicamente, existem duas formas para um usuário final acessar uma base de dados. A primeira é a partir de um software, por exemplo, um painel construído com Spotfire [17]. A segunda é através da linguagem SQL, mas para esta é preciso que o profissional desenvolva uma habilidade associada à computação.

Um problema da primeira abordagem é que o profissional que deseja ter novas visualizações sobre os dados precisa demandar a evolução dos painéis ou sistemas que ele utiliza para realizar suas pesquisas. Isso pode ser um inconveniente em relação aos prazos praticados em diferentes indústrias hoje em dia. Da mesma forma, a segunda abordagem também traz problemas, uma vez que é difícil capacitar profissionais de diferentes áreas na linguagem SQL.

O problema conhecido como *Natural Language Interface to Database* (NLIDB) investiga sistemas capazes de permitir que usuários acessem informações armazenadas em um banco de dados por meio de solicitações expressas em alguma linguagem natural, por exemplo, Inglês. Protótipos de sistemas NLIDBs existem desde o final dos anos 60, mas, apesar dos avanços ao longo dos anos, esses sistemas não conseguiram a rápida aceitação comercial como era esperado [1]. O desenvolvimento de alternativas bem-sucedidas aos NLIDBs (em especial a evolução das interfaces gráficas) e as dificuldades inerentes ao processamento de

linguagem natural (ambiguidade, contexto, etc.) são provavelmente as principais razões para a baixa aceitação [35].

Diversos autores têm investigado esse problema ao longo dos anos. Por exemplo, [2], [3], [4], [5], [6], [13], [14], [15] e [16], apresentaram propostas e implementações de soluções de conversão de sentenças em inglês para SQL. Outros trabalhos foram feitos com o objetivo de mapear as diferentes iniciativas no tema, como [7], [8], [9], [10], [11] e [12].

Com o advento das redes neurais profundas [19] e dos modelos *Sequence to Sequence* (SEQ2SEQ) [20], uma linha de soluções para o problema NLIDB passou a aplicar redes neurais profundas em arquiteturas que possuem codificadores para receber as consultas em linguagem natural e decodificadores para a saída das consultas SQL, e a chamar o problema de *Natural Language To SQL* (NL2SQL) [21].

Um exemplo de *dataset* para o problema NLIDB é o WikiSQL [22]. Diferentes autores submeteram resultados para este *dataset*, por exemplo, [23], [24], [25], [26], [27] e [28]. O ranking atualizado do WikiSQL pode ser acessado em [29].

Uma alternativa à abordagem que utiliza modelos SEQ2SEQ é a estratégia baseada em Síntese de Programas ou Indução de Programas [34]. No contexto deste trabalho, os programas a serem gerados nesta nova estratégia são consultas em SQL.

Por fim, um ponto relevante a ser considerado é a carência de material sobre o tema NLIDB para o idioma português. Essa é uma questão importante que será investigada neste estudo orientado e que provavelmente será o maior desafio, visto que existem poucas bases de dados em português para este problema. Assim, o modelo proposto precisará aprender a partir de uma base de dados limitada.

2. Resultados Esperados

Segue a lista de atividades e subatividades previstas, além das expectativas de conclusão.

Atividade 1. Pesquisa Bibliográfica
- Estudar artigos do ranking WikiSQL.
- Estudar outros *datasets* para problemas NLIDB.
- Pesquisar *surveys* e outros artigos que tratam NLIDB
- Propor *survey* para problemas NLIDB.

Atividade 2. Implementação de modelo NLIDB para o *dataset* WikiSQL
- Executar modelos do ranking WikiSQL.
- Propor e implementar um modelo de conversão de consultas em linguagem natural para SQL, e testá-lo no *dataset* WikiSQL. – DESEJÁVEL

Atividade 3. Criação de um *dataset* em português para problemas NLIDB

- Propor estratégia colaborativa para geração do *dataset* em português.
- Implementar estratégia colaborativa. - DESEJÁVEL

3. Critério de Avaliação

O estudo será avaliado com base na realização das atividades descritas no Cronograma de Atividades apresentado na seção a seguir.

4. Cronograma de Atividades

Atividade 1. Pesquisa Bibliográfica (agosto e setembro)
Atividade 2. Implementação de modelo NLIDB para o *dataset* WikiSQL (outubro e novembro)
Atividade 3. Criação de um *dataset* em português para problemas NLIDB (outubro e novembro)

5. Referências

[1] Androutsopoulos, Ion, Graeme D. Ritchie, and Peter Thanisch. "Natural language interfaces to databases–an introduction." Natural language engineering 1.1 (1995): 29-81.

[2] Popescu, Ana-Maria, Oren Etzioni, and Henry Kautz. "Towards a theory of natural language interfaces to databases." Proceedings of the 8th international conference on Intelligent user interfaces. ACM, 2003.

[3] Popescu, Ana-Maria, et al. "Modern natural language interfaces to databases: Composing statistical parsing with semantic tractability." Proceedings of the 20th international conference on Computational Linguistics. Association for Computational Linguistics, 2004.

[4] Minock, Michael. "A phrasal approach to natural language interfaces over databases." International Conference on Application of Natural Language to Information Systems. Springer, Berlin, Heidelberg, 2005.

[5] Chaitali, Kulkarni, et al. "Natural Language Interface To Database." (2016).

[6] Kaufmann, Esther, and Abraham Bernstein. "How useful are natural language interfaces to the semantic web for casual end-users?." The Semantic Web. Springer, Berlin, Heidelberg, 2007. 281-294.

[7] Sujatha, B., Dr S. Vishwanadha Raju, and Humera Shaziya. "A survey of natural language interface to database management system." International Journal of Science and Advanced Technology, ISSN (2012): 2221-8386.

[8] Patel, Jaina, and Jay Dave. "A Survey: Natural Language Interface to Databases." International Journal of Advance Engineering and Research Development (IJAERD) (2015).

[9] Lopez, Vanessa, et al. "Is question answering fit for the semantic web?: a survey." Semantic Web 2.2 (2011): 125-155.

[10] Kolomiyets, Oleksandr, and Marie-Francine Moens. "A survey on question answering technology from an information retrieval perspective." Information Sciences 181.24 (2011): 5412-5434.

[11] Bouziane, Abdelghani, et al. "Question answering systems: survey and trends." Procedia Computer Science 73 (2015): 366-375.

[12] Mishra, Amit, and Sanjay Kumar Jain. "A survey on question answering systems with classification." Journal of King Saud University-Computer and Information Sciences 28.3 (2016): 345-361.

[13] Akula, Arjun R. "A Novel Approach Towards Building a Generic, Portable and Contextual NLIDB System." International Institute of Information Technology Hyderabad(2015).

[14] Nisa, Qamar Un, and Fiaz Majeed. "A Dynamic Form-Based Natural Language Interface to Data Warehouse Question Answering." IJCCER 1.4 (2013): 104-110.

[15] Mony, Manju, Jyothi M. Rao, and Manish M. Potey. "An Overview of NLIDB Approaches and Implementation for Airline Reservation System." International Journal of Computer Applications 107.5 (2014).

[16] Akula, Arjun, Rajeev Sangal, and Radhika Mamidi. "A novel approach towards incorporating context processing capabilities in nlidb system." Proceedings of the Sixth International Joint Conference on Natural Language Processing. 2013.

[17] https://spotfire.tibco.com/ - Acessado em 01/09/2018

[18] Mohanty, Soumendra, Madhu Jagadeesh, and Harsha Srivatsa. Big data imperatives: Enterprise ‘Big Data’warehouse,‘BI’implementations and analytics. Apress, 2013.

[19] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." nature 521.7553 (2015): 436.

[20] Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." Advances in neural information processing systems. 2014.

[21] Xu, Xiaojun, Chang Liu, and Dawn Song. "Sqlnet: Generating structured queries from natural language without reinforcement learning." arXiv preprint arXiv:1711.04436 (2017).

[22] Zhong, Victor, Caiming Xiong, and Richard Socher. "Seq2SQL: Generating structured queries from natural language using reinforcement learning." arXiv preprint arXiv:1709.00103(2017).

[23] Wang, Chenglong, Marc Brockschmidt, and Rishabh Singh. "Pointing Out SQL Queries From Text." (2018).

[24] Huang, Po-Sen, et al. "Natural Language to Structured Query Generation via Meta-Learning." arXiv preprint arXiv:1803.02400 (2018).

[25] Yu, Tao, et al. "TypeSQL: Knowledge-based Type-Aware Neural Text-to-SQL Generation." arXiv preprint arXiv:1804.09769 (2018).

[26] Wang, Chenglong, et al. "Execution-Guided Neural Program Decoding." arXiv preprint arXiv:1807.03100 (2018).

[27] Dong, Li, and Mirella Lapata. "Coarse-to-Fine Decoding for Neural Semantic Parsing." arXiv preprint arXiv:1805.04793(2018).

[28] McCann, Bryan, et al. "The Natural Language Decathlon: Multitask Learning as Question Answering." arXiv preprint arXiv:1806.08730 (2018).

[29] https://github.com/salesforce/WikiSQL - acessado em 01/09/2018.

[30] Oracle. https://www.oracle.com/br/database/index.html - Acessado em 01/09/2018.

[31] SQL Server. https://www.microsoft.com/pt-br/sql-server/sql-server-2017 - Acessado em 01/09/2018.

[32] Han, Jing, et al. "Survey on NoSQL database." Pervasive computing and applications (ICPCA), 2011 6th international conference on. IEEE, 2011.

[33] Hadoop. http://hadoop.apache.org/ - Acessado em 01/09/2018.

[34] Yaghmazadeh, Navid, et al. "Sqlizer: Query synthesis from natural language." Proceedings of the ACM on Programming Languages 1.OOPSLA (2017): 63.

[35] https://www.nexthink.com/blog/natural-language-interfaces-to-databases-nlidb/ - Acessado em 06/09/2018

Niterói, 14 de setembro de 2018

_____

Alexandre Plastino

_____

Aline Paes

_____

Henrique Bueno Rodrigues