Plano de Trabalho para Doutorado

Pós-Graduação em Computação - IC - UFF

Candidato: Henrique Bueno Rodrigues
Orientador: Alexandre Plastino
Coorientadora: Aline Paes
Tema: Explorando Estratégias fracamente supervisionadas de Construção de Consultas em Bases de Dados com Interfaces em Linguagem Natural
Área de Estudo: Engenharia de Sistemas e Informação

Resumo

Processamento de Linguagem Natural é uma área da Ciência da Computação que explora como computadores podem ser usados para entender e manipular textos e discursos em Linguagem Natural (LN) para atender objetivos específicos. Consultas a Bases de Dados através de consultas em LN é um exemplo de aplicação desta área. Nesse contexto, esse plano de trabalho propõe uma avaliação das técnicas atuais que tratam o problema de conversão de consultas em LN para consultas a Bases de Dados, e o desenvolvimento de uma ferramenta que aplique novas estratégias a serem propostas para este problema, com foco em técnicas de aprendizado de máquina fracamente supervisionadas.

1. Caracterização do problema

O volume e a diversidade de dados disponíveis na *web* e nas empresas crescem rapidamente. Nesse contexto, surge o termo *Big Data* [1], que resumidamente se refere a conjuntos de dados que possuem tamanho superior ao tamanho suportado por *softwares* típicos de manipulação de Bancos de Dados. Essa característica implica no aumento da dificuldade de gerenciar informações eficientemente, em especial, de buscar informações relevantes.

Enquanto o problema de busca na *Web* (*Web Search*) possui hoje soluções bem estabelecidas, por exemplo, a máquina de busca Google [2], o problema de Busca Corporativa (*Enterprise Search*) ainda apresenta importantes desafios a serem tratados, por exemplo, a necessidade de acessar dados em diferentes repositórios, as restrições de acesso individuais às informações, a necessidade de indexar documentos de diferentes tipos e idiomas e a necessidade de acessar dados estruturados e não estruturados de forma transparente [3][4].

As exigências às ferramentas de busca corporativas crescem à medida que usuários se acostumaram a utilizar ferramentas de busca eficientes e com alta usabilidade fora das empresas. Por exemplo, oferecer ao usuário uma interface de busca amigável, tipicamente uma caixa de texto capaz de interpretar consultas escritas em linguagem natural, é importante para a aceitação de uma ferramenta de busca corporativa. Nesse contexto, ganha destaque a área de pesquisa chamada Interface em Linguagem Natural, que está na interseção da área Processamento de Linguagem Natural e Interação Humano Computador, e procura fornecer

significado para interações entre humanos e computadores através do uso de linguagem natural [5]. Diversos trabalhos foram propostos com o objetivo de tratar a questão de busca em bases de dados através de consultas escritas em linguagem natural.

Muitos dos dados produzidos pelos processos que formam a cadeia de valor de uma empresa são estruturados e armazenados em Bancos de Dados relacionais, através de sistemas construídos ou comprados de fornecedores. Assim, usuários são limitados às funcionalidades oferecidas por esses sistemas para realizar consultas nas Bases de Dados. Outra opção menos comum é o usuário acessar a base de dados diretamente. Entretanto, esta estratégia possui complicadores, por exemplo, a necessidade do usuário conhecer a linguagem SQL e ter acesso aos repositórios e esquemas.

Algumas ferramentas como VGE [6] e Spotfire [7] se propõem a realizar consultas, por exemplo, em bancos de dados relacionais, de forma flexível e independente do domínio da informação. Entretanto, uma limitação dessas ferramentas é a necessidade de configuração manual prévia de mapeamentos entre os conceitos exibidos e as tabelas do banco de dados.

O uso de técnicas de Aprendizado de Máquina [22], que de forma simplificada podem ser descritas como estratégias para modelagem computacional de processos de aprendizado, podem ser utilizadas para a redução da necessidade de intervenção humana nas ferramentas citadas.

Diversos trabalhos utilizaram técnicas de aprendizado de máquina para execução automática de tarefas de processamento de consultas em linguagem natural.

Uma proposta chamada *Dynamic Memory Network* (DMN) é apresentada em [8], onde é utilizado um *framework* baseado em redes neurais para a execução de tarefas genéricas de *Question Answering* [13]. A entrada da DMN é composta de um conjunto de sentenças, que caracterizam um determinado contexto, mais uma pergunta. Após o processamento das sentenças de entrada, são identificados fatos relevantes que servirão de insumo para um módulo de geração da resposta. Um ponto de atenção dessa estratégia é a necessidade de um conjunto de fatos de entrada para suportar a elaboração da resposta.

Um método baseado em aprendizado de representações, com uso de *Deep Learning*, chamado de *Seq2SQL* é proposto em [9] para transformar perguntas em linguagem natural em SQL e apresenta ótimos resultados principalmente em função do uso de aprendizado por reforço. A entrada da rede é uma pergunta escrita em linguagem natural e o esquema do banco de dados alvo, e a saída é uma consulta SQL. A consulta SQL gerada é executada no Banco de Dados e o resultado atua como uma recompensa para reforçar o treinamento da rede.

Uma estratégia para construir interfaces em linguagem natural para bancos de dados de novos domínios é apresentada em [10]. Neste trabalho, é apresentada uma estratégia de conversão da linguagem natural diretamente para SQL, sem a necessidade de modelos intermediários. Entretanto, um ponto de atenção é a necessidade de publicação *online* dos modelos criados para

a validação de usuários reais. Segundo os autores, a popularidade da linguagem SQL faz com que a comunidade participe ativamente das validações, entretanto, gera uma dependência da intervenção de especialistas ao mesmo tempo que permite que usuários leigos possam dar *feedback*.

Uma abordagem chamada *Schema-Agnostic* é apresentada em [12]. Basicamente, a estratégia é evoluir uma infraestrutura de mapeamento de consultas em linguagem natural para SPARQL [11] sem a necessidade de intervenção de especialistas, para a criação de ontologias que são geradas dinamicamente a partir de processamentos de grandes conjuntos de dados disponíveis na web. Um ponto de atenção é a aplicação desta estratégia em domínios específicos já que o trabalho se concentrou em domínios de dados gerais.

O IBM Watson [14] é um sistema de computação cognitiva que combina as capacidades de processamento de linguagem natural, de geração de hipóteses e avaliação das mesmas, e aprendizagem dinâmica a partir de cada interação e iteração. Apesar de ser um sistema de escopo amplo, no âmbito desse trabalho, um ponto interessante de avaliação do Watson é sua capacidade de considerar o contexto do usuário que está utilizando o sistema no processo de tradução das consultas em linguagem natural para consultas às Bases de Dados. Isso torna os resultados mais relevantes aos consultantes. Por outro lado, ele tem a desvantagem de ser um sistema fechado, ou seja, proprietário.

O projeto Optique [15] utiliza abordagens semânticas para resolver o problema de integração de bases de dados legadas e consulta sobre informações não estruturadas, alcançando bons resultados na indústria de petróleo. Basicamente é criado manualmente um mapeamento de uma ontologia para as diversas fontes de informação, o que permite ao usuário abstrair as fontes originais e ter uma visão semântica das informações. Entretanto, é necessário que um mapeamento prévio seja feito entre a ontologia e as bases de dados.

Considerando os aspectos positivos e negativos dos trabalhos anteriores, este plano de trabalho apresenta uma proposta de construção de um método de construção de consultas em bases de dados, capaz de processar consultas escritas em linguagem natural, sem a necessidade de configurações prévias de mapeamentos entre os conceitos exibidos e as bases de dados consultadas.

2. Objetivo

O principal objetivo a ser alcançado com o doutorado é propor um método de transformação de consultas em linguagem natural para consultas em Banco de Dados, considerando um baixo grau de intervenção, ou seja, métodos fracamente supervisionados. Dessa forma, alterações nas fontes de dados serão consideradas automaticamente pelo novo mecanismo criado, sem a necessidade de muitas intervenções de especialistas.

Para isso, será preciso coletar e transformar conjuntos de dados que serão utilizados para o treinamento do método proposto, e desenvolver uma ferramenta que implemente este método e

funcione como um *front-end* para a realização de uma avaliação qualitativa por um conjunto de usuários.

3. Metodologia

Primeiramente, será feito um levantamento sobre trabalhos relacionados ao tema do doutorado. Após o levantamento, será proposta e implementada uma ferramenta de consulta em bases de dados relacionais baseada em perguntas escritas em linguagem natural.

A implementação será avaliada de forma quantitativa através da comparação da proposta com as estratégias já apresentadas na literatura. Essas comparações utilizarão conjuntos de dados disponíveis na *web* que poderão ser de dois tipos: Dados de domínio aberto ou dados de domínio específico [12].

Dados de domínio aberto consistem em repositórios de dados que cobrem diferentes domínios de conhecimento. Eles cobrem informações menos específicas em diferentes áreas. O usuário comum desse tipo de dados é tipicamente o usuário da *web*. Conhecimento de enciclopédias, definições, dados de filmes e esportes são exemplos desse tipo de domínio, que podem ser encontrados na web em *sites* como DBpedia [16], Freebase [17] e YAGO [18].

Dados de domínio específico descrevem um único domínio, tipicamente técnico e com alta especificidade. O usuário comum desse tipo de dados é o especialista ou analista do domínio. Relatórios financeiros e dados médicos são exemplos de dados de domínio específico, que podem ser encontrados na *web* em PubChem [19], Diseasome [20] e Drugbank [21].

Além da análise quantitativa também será feita uma avaliação qualitativa através de uma pesquisa de satisfação com os usuários da ferramenta que será desenvolvida em um estudo de caso. O grupo de usuários será formado por engenheiros, geólogos e geofísicos que realizam consultas a dados da indústria de Exploração e Produção (E&P) de petróleo no contexto da Petrobras.

4. Cronograma de Trabalho

O cronograma de trabalho está dividido em 4 fases, cada uma correspondendo a um ano de trabalho do doutorado:

Fase 1:

- **Disciplinas**: cursar 6 disciplinas relacionadas ao tema deste plano de trabalho com o objetivo de cumprir os créditos necessários e aprimorar os conhecimentos. No primeiro semestre de 2017 cursei a disciplina “Mineração de dados” como aluno avulso e atualmente (segundo semestre de 2017) estou cursando a disciplina “Tratamento de Incertezas” também como aluno avulso.
- **Revisão bibliográfica**: fazer um levantamento sobre as publicações relacionadas ao tema deste plano de trabalho.

Fase 2:

- **Coleta do conjunto de dados**: selecionar e transformar conjuntos de dados que serão utilizados no treinamento da estratégia proposta.
- **Proposta e implementação de estratégia**: propor e implementar estratégia que atenda aos objetivos apresentados neste plano de trabalho.
- **Realizar avaliação quantitativa**: comparar os resultados da ferramenta implementada com as estratégias já apresentadas na literatura, utilizando os conjuntos de dados apresentados nos objetivos deste plano de trabalho.
- **Redação de um artigo**: escrever e submeter um artigo com os resultados obtidos.

Fase 3:

- **Realizar avaliação qualitativa:** realizar uma pesquisa com usuários em um estudo de caso com o objetivo de obter insumos para ajustes na ferramenta e nos algoritmos implementados.
- **Redação da proposta de tese**.
- **Defesa da proposta de tese**.

Fase 4:

- **Aprimoramento da ferramenta proposta**: implementar melhorias na ferramenta identificadas durante a avaliação qualitativa com os usuários.
- **Redação de um artigo**: escrever e submeter um artigo com os novos ajustes implementados.
- **Redação da tese**.
- **Defesa da tese**.

Referências

[1] Manyika, James, et al. "Big data: The next frontier for innovation, competition, and productivity." McKinsey Global Institute Technical Report (2011).

[2] Brin, Sergey, and Lawrence Page. "Reprint of: The anatomy of a large-scale hypertextual web search engine." Computer networks 56.18 (2012): 3825-3833.

[3] Hawking, David. "Challenges in enterprise search." Proceedings of the 15th Australasian database conference-Volume 27. Australian Computer Society, Inc., 2004. P. 15-24.

[4] Mukherjee, Rajat, and Jianchang Mao. "Enterprise search: Tough stuff." Queue 2.2 (2004): 36.

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

[6] Botelho, Vinicius, et al. "Gerenciamento Empresarial de Conteúdo de Informações de Exploração e Produção da Petrobras com o apoio de um SIG especializado." 11th International Congress of the Brazilian Geophysical Society. 2009: pp. 457-462.

[7] Ahlberg, Christopher. "Spotfire: an information exploration environment." ACM SIGMOD Record 25.4 (1996): 25-29.

[8] Kumar, Ankit, et al. "Ask me anything: Dynamic memory networks for natural language processing." International Conference on Machine Learning. 2016. P. 1378-1387.

[9] Zhong, Victor, Caiming Xiong, and Richard Socher. "Seq2SQL: Generating Structured Queries from Natural Language using Reinforcement Learning." arXiv preprint arXiv:1709.00103 (2017).

[10] Iyer, Srinivasan, et al. "Learning a Neural Semantic Parser from User Feedback." arXiv preprint arXiv:1704.08760 (2017).

[11] https://www.w3.org/TR/rdf-sparql-query acessado em 13/11/2017.

[12] Freitas, André. Schema-agnostic queries for large-schema databases: A distributional semantics approach. PhD Theses, DERI (Digital Enterprise Research Institute), National University of Ireland, Galway. 2015.

[13] Martin, James H., and Daniel Jurafsky. "Speech and language processing." International Edition 710 (2000): 25.

[14] High, Rob. "The era of cognitive systems: An inside look at ibm watson and how it works." IBM Corporation, Redbooks (2012).

[15] Giese, Martin, et al. "Optique: Zooming in on big data." Computer 48.3 (2015): 60-67.

[16] http://datahub.io/dataset/dbpedia acessado em 13/11/2017.

[17] http://datahub.io/dataset/freebase acessado em 13/11/2017.

[18] http://datahub.io/dataset/yago acessado em 13/11/2017.

[19] http://pubchem.ncbi.nlm.nih.gov/ acessado em 13/11/2017.

[20] http://datahub.io/dataset/fu-berlin-diseasome acessado em 13/11/2017.

[21] http://datahub.io/dataset/fu-berlin-drugbank acessado em 13/11/2017.

[22] Michalski, Ryszard S., Jaime G. Carbonell, and Tom M. Mitchell, eds. Machine learning: An artificial intelligence approach. Springer Science & Business Media, 2013.

Niterói, 16 de novembro de 2017

_____

Henrique Bueno Rodrigues – Candidato

_____

Prof. Alexandre Plastino – (Orientador)

_____

Prof. Aline Paes – (Coorientadora)