PONTIFÍCIA UNIVERSIDADE CATÓLICA
DO RIO DE JANEIRO

Técnicas Inteligentes para Interpretação de Documentos em uma Plataforma de Extração de Dados

Gabriel Augusto Silva de Aquino

PROPOSTA DO PROJETO FINAL DE GRADUAÇÃO

Orientação Prof. Marco A. Casanova

CENTRO TÉCNICO CIENTÍFICO - CTC

DEPARTAMENTO DE INFORMÁTICA

Curso de Graduação em Ciência da Computação

Rio de Janeiro, outubro de 2020

SUMÁRIO

1. INTRODUÇÃO

Danke [1] é uma plataforma para extração de dados e conhecimento, desenvolvida no Instituto Tecgraf [2], PUC-Rio. Ela permite que os usuários realizem busca de palavras-chave sobre bancos de dados relacional e RDF [9]. Danke explora o esquema do banco de dados para compilar uma consulta em SQL ou SPARQL, com o mínimo de cláusulas joins, que retornam dados que melhor se casam com as palavras-chave. A plataforma permite que os usuários explorem os resultados da busca em uma tabela, com os resultados mais relevantes no topo. Os usuários podem adicionar ou remover colunas desta tabela, adequando o resultado de acordo com seu interesse, e também navegar sobre o banco de dados a partir da resposta da consulta, acessando detalhes de cada resultado e seus relacionamentos com outros dados. Para que o Danke possa ser utilizado em um determinado domínio de aplicação, é necessário definir o esquema para o banco de dados de acordo com tal domínio, além de se realizar um processo de preparação do banco de dados, que envolve a ingestão, indexação e enriquecimento dos dados, de forma a responder com maior desempenho e eficácia as buscas do usuário.

Este projeto trata da incorporação de documentos textuais (por exemplo, com extensões .pdf, .txt e .doc) no banco de dados do Danke [1, 8], permitindo que as buscas encontrem palavras-chave no conteúdo destes documentos e que o usuário possa encontrar e navegar sobre os relacionamentos destes documentos com as outras entidades do banco de dados. Para tanto, o projeto desenvolverá técnicas inteligentes para interpretação de documentos, tratando do reconhecimento de entidades nomeadas em documentos (NER), associando-as às entidades já existentes no banco de dados, da extração de relacionamentos (ER), e da indexação dos documentos [6,7].

2. SITUAÇÃO ATUAL

Extração de dados refere-se à tarefa de extrair entidades, relacionamentos e valores de atributos de documentos ou arquivos semiestruturados. Na literatura, a tarefa específica de identificar entidades nomeadas em textos é conhecida por *named entity recognition* (NER) [3], e a tarefa de extrair relacionamentos entre entidades por *relation extraction* (RE) [4].

Extração de dados adquiriu grande importância pois existe um volume considerável de informação disponível sob forma de texto em linguagem natural, que é conveniente para

consumo de usuários humanos, mas não para uso por algoritmos de análise de dados. Este fato exigiu o desenvolvimento de métodos e ferramentas para interpretação, mesmo que superficial, de texto em linguagem natural. Como na tarefa de identificação de entidades, pode-se tratar extração de dados sob duas suposições: “hipótese do mundo fechado”, quando os documentos ou arquivos semiestruturados referem-se a um domínio de aplicação bem conhecido; e “hipótese do mundo aberto”, em caso contrário.

As técnicas mais antigas para extração de dados baseiam-se em regras e usam um dicionário específico para o domínio de aplicação em questão para identificar termos e frases no texto. Em particular, enfoques de linguística computacional baseados em regras exploram a estrutura sintática das sentenças para melhorar o processo de extração. Métodos mais recentes baseiam-se em aprendizagem de máquina [5] e, em particular, deep learning (notadamente recurrent neural networks - RNNs e convolutional neural networks - CNNs). Os métodos para identificar relacionamentos entre entidades mais bem sucedidos aplicam aprendizagem supervisionada para construir classificadores, que utilizam características (features) extraídas de sentenças anotadas manualmente em um corpus de treinamento. Supervisão a distância (distant supervision) endereça o problema de gerar automaticamente exemplos, em número suficiente para treinamento dos algoritmos, com a ajuda de bancos de dados do domínio da aplicação.

Este projeto abordará o problema de reconhecimento de entidades nomeadas (NER) e de extração de relacionamentos (RE) sob a “hipótese do mundo fechado”, assumindo a existência de um banco de dados específico para o domínio com as principais entidades e relacionamentos encontrados nos documentos.

Mais especificamente, este projeto tratará de incorporar técnicas de reconhecimento de entidades nomeadas (NER) nos documentos da plataforma Danke [1], baseado nos dados existentes no banco de dados do Danke. Não é uma interpretação profunda, mas sim o reconhecimento das entidades que um documento se refere com hipótese do mundo fechado, o que significa que o objetivo não é reconhecer qualquer entidade, mas sim reconhecer no documento, as entidades que estão representadas no banco de dados do Danke. Posteriormente, o projeto também tratará de incorporar técnicas de extração de relacionamentos (ER) em documentos na plataforma Danke. São dois problemas que estão interrelacionados, mas em ambos a ideia é utilizar o banco de dados do Danke como referência para as entidades e os relacionamentos.

3. Objetivo e Proposta do Trabalho

A proposta deste projeto final consiste em: (1) fazer um levantamento dos requisitos e as funcionalidades do Danke para incorporar documentos textuais, permitindo que as buscas encontrem palavras-chave no conteúdo destes documentos e que o usuário possa encontrar e navegar sobre os relacionamentos destes documentos com as outras entidades do banco de dados; (2) estudar tecnologias e técnicas já existentes para realizar interpretação de documentos, através de reconhecimento de entidades nomeadas em documentos (NER) e de extração de relacionamentos (RE); (3) propor técnicas já existentes ou criar novas técnicas que sejam mais adequadas para serem incorporadas na plataforma Danke; (4) propor uma evolução da arquitetura do Danke para incorporar estas técnicas; (5) implementar as técnicas propostas; (5) realizar experimentos destas técnicas no Danke com um banco de dados e documentos para algum domínio de aplicação específico.

Espera-se que a nova arquitetura do Danke possa aumentar o poder da busca e do acesso aos dados.

4. Plano de ação

Este projeto divide-se nas seguintes etapas:

1) Fazer um levantamento dos requisitos e funcionalidades do Danke para incorporar documentos textuais, permitindo que as buscas encontrem palavras-chave no conteúdo destes documentos e que o usuário possa encontrar e navegar sobre os relacionamentos destes documentos com as outras entidades do banco de dados.
2) Registrar o levantamento realizado em um documento.
3) Estudar tecnologias e técnicas já existentes para realizar interpretação de documentos, através de reconhecimento de entidades nomeadas em documentos (NER) e de extração de relacionamentos (RE).
4) Registrar o estudo realizado em um documento.
5) Incorporar técnicas NER ao Danke:
    a) Propor ou criar novas técnicas NER que sejam mais adequadas para serem incorporadas na plataforma Danke.

b) De acordo com o nível de complexidade, definir etapas e o que será efetivamente implementado destas técnicas na plataforma Danke durante este projeto final.
c) Propor e documentar uma evolução da arquitetura do Danke para incorporar estas técnicas NER.
d) Implementar as técnicas NER definidas.
e) Realizar experimentos destas técnicas NER no Danke com um banco de dados e documentos para algum domínio de aplicação específico.

6) Incorporar técnicas ER ao Danke:
a) Propor ou criar novas técnicas RE que sejam mais adequadas para serem incorporadas na plataforma Danke;
b) De acordo com o nível de complexidade, definir etapas e o que será efetivamente implementado destas técnicas na plataforma Danke durante este projeto final
c) Propor e documentar uma evolução da arquitetura do Danke para incorporar estas técnicas ER;
d) Implementar as técnicas ER definidas;
e) Realizar experimentos destas técnicas ER no Danke com um banco de dados e documentos para algum domínio de aplicação específico.

7) Criação de um documento detalhado de todo o projeto.

Cronograma

Etapas	Set	Out	Nov	Dez	Jan	Fev	Mar	Abr	Mai
Etapa 1	x	x							
Etapa 2		x							
Etapa 3			x	x					
Etapa 4				x					
Etapa 5				x	x	x			
Etapa 6						x	x	x	
Etapa 7		x	x	x	x	x	x	x	x

5. Referências Bibliográficas

[1] Danke: Data and Knowledge Retrieval. *http://danke.tecgraf.puc-rio.br*. Acesso em Outubro/2020.

[2] Instituto Tecgraf: Instituto de desenvolvimento e pesquisa da PUC-Rio. *http://www.tecgraf.puc-rio.br*. Acesso em Outubro/2020.

[3] Nadeau, David; Sekine, Satoshi. *A survey of named entity recognition and classification*. National Research Council Canada / New York University. Acesso em Outubro/2020

[4] Herman, Andreas. *Different ways of doing Relation Extraction from text.* Acesso em Outubro/2020.

[5] Dong, X.L and Rekatsinas, T. "Data Integration and Machine Learning: A Natural Synergy". VLDB 2018. Acesso em Outubro/ 2020.

[6] Gormley, Clinton; Tong, Zachary. *Elasticsearch: The Definitive Guide*. Acesso em Outubro/2020.

[7] Smiley, David; Pugh, Eric; Parisa, Kranti; Mitchell, Matt. *Apache Solr Enterprise Search Server*. Acesso em Outubro/2020.

[8] Izquierdo, Y. T.; García, G. M.; Cavaliere, M. L.; Novello, A.; Novelli, B. A.; Damasceno, C. O.; Leme, L.A.P.P.; Casanova, M. A. . Comparing and Recommending Conferences. In: Brazilian Symposium on Databases - SBBD, 2020 (To be published). Acesso em Outubro/2020.

[9] García, G.M. A Keyword-based Query Processing Method for Datasets with Schemas. Thesis presented to the Graduate Program in Informatics, PUC-Rio (March 2020). DOI: *https://doi.org/10.17771/PUCRio.acad.48728*. Acesso em Outubro/2020

Rio de Janeiro, 05 de outubro de 2020

Gabriel Augusto Silva de Aquino

Marco Antonio Casanova