O USO DE DISTÂNCIA DE LEVENSHTEIN COM REDES NEURAIS PARA IDENTIFICAÇÃO DE CORRESPONDÊNCIA ENTRE ENTIDADES

IHAN BELMONTE BENDER[1]; RICARDO MATSUMURA DE ARAÚJO[2]

[1]*Universidade Federal de Pelotas – ihanbender@gmail.com*
[2]*Universidade Federal de Pelotas – ricardo@inf.ufpel.edu.br*

1. INTRODUÇÃO

Mesmo sendo muito importante que exista verificação da qualidade da pesquisa desenvolvida em um país, a avaliação da produção científica de programas de pós graduação (PPGs) é uma tarefa muito desafiadora, principalmente em um país de tamanho continental como o Brasil. Além da dificuldade em avaliar quantitativamente o quanto é produzido pelos pesquisadores de cada programa, avaliar qualitativamente o que é produzido nas mais diversas áreas de maneira justa e o menos subjetiva possível é uma tarefa árdua.

Para tentar solucionar, entre outros problemas, o da avaliação qualitativa, nos anos 1990, foi criada a plataforma Lattes, pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e no fim da mesma década, padronizado o currículo Lattes (CNPQ, entre 2005 e 2019). Já para ser possível a avaliar também a qualidade das produções, foi criada uma métrica chamada Qualis, calculada periodicamente por consultores de cada uma das áreas de avaliação. No caso, a métrica consiste em um conceito que é dado a diferentes veículos de publicação (conferências, periódicos e livros, por exemplo). Para avaliar a qualidade de uma produção com o Qualis, verifica-se o conceito do veículo onde foi publicada, e o transfere para a publicação.

Mesmo com essas ferramentas, a avaliação ainda encontra problemas, já que os currículos lattes contam com diversos erros de preenchimento nos campos importantes para a avaliação, principalmente os nomes de conferências, dificultando a qualificação das produções de alguns autores. A preocupação específica em relação às conferências é causada pela não existência de um código identificador único para as mesmas. No caso de periódicos, por exemplo, mesmo que o pesquisador preencha o nome do veículo de forma incorreta, existe um campo para que ele insira o ISSN, um código que permite descobrir qual o periódico em que a produção foi divulgada.

Em algumas áreas, como a computação, grande parte da produção científica é divulgada nesse tipo de meio, sendo importante, então, a avaliação correta. Para poder realizar de maneira automatizada a avaliação de programas de pós-graduação, é necessário um algoritmo que seja capaz de encontrar a correspondência entre nomes de conferências escritos de maneira incorreta em currículos Lattes e nomes de conferências reais com Qualis já atribuído.

Neste trabalho, é criada uma técnica de casamento de entidades, área responsável por identificar correspondências entre diferentes registros que representam a mesma entidade do mundo real (WINKLER, 2014) utilizando uma rede neural artificial que é alimentada com variações da distância de Levenshtein (LEVENSHTEIN, 1966) utilizadas em conjunto com normalização de texto, remoção de *stop words (SWs)* (RAJARAMAN, 2011) e um algoritmo de identificação de siglas.

2. METODOLOGIA

Inicialmente, foi necessária a construção de uma base de dados para realizar o treinamento e o teste do modelo de rede neural desenvolvido. A lista de nomes corretos e com Qualis de conferências da área de ciência da computação foi obtida através da plataforma Sucupira, e possibilitou a construção de uma base de dados. Para obter os dados de treino e avaliação, foram baixados currículos Lattes de pesquisadores participantes de programas de pós-graduação em computação e extraídos os nomes das conferências onde os mesmos afirmaram ter publicado.

A técnica desenvolvida recebe como entrada dois nomes de conferência, um inserido via currículo e outro com nome oficial. Chamaremos estes nomes de Nome A e Nome B, respectivamente. São geradas duas versões de cada um dos nomes, uma normalizada e outra normalizada com remoção de *SWs*, palavras que não oferecem muita semântica ao texto. A normalização é feita da seguinte maneira: Inicialmente, todos os caracteres do texto são transformados para caixa baixa. Em seguida, são removidos símbolos, números (cardinais, ordinais), tanto escritos por extenso quanto em valor numérico, além de algarismos romanos. Por fim, são normalizados os espaços entre diferentes palavras, limitando a apenas um por intervalo.

Ao fim do processo descrito, temos duas versões tanto do Nome A quanto do Nome B. Para verificar a semelhança entre os dois nomes, realizamos um conjunto de operações para cada dupla, sendo a Dupla 1 composta pelos nomes A e B apenas normalizados e a Dupla 2 pelos nomes normalizados com SWs removidas. Cada uma das duplas serve de entrada para quatro algoritmos baseados na Distância de Levenshtein: *Ratio, Partial Ratio, Token Sort Ratio e Token Set Ratio*.

O *Ratio* é o resultado do cálculo da distância de Levenshtein, onde a operação de substituição tem peso dois (a versão original tem peso dois), dividido pela soma dos tamanhos dos dois textos recebidos como entrada.

O *Partial Ratio* é uma métrica que identifica o menor de dois textos e retorna uma porcentagem correspondente à semelhança do menor texto com o pedaço mais parecido do texto maior, utilizando distância Levenshtein.

Tanto o *Token Set Ratio* quanto o *Token Sort Ratio* realizam uma divisão dos textos em *tokens*, que no caso das conferências são as palavras que formam o nome. Após a divisão, verifica-se o quanto tem em comum em relação a esses tokens. A diferença entre ambas é que o *Token Sort Ratio* ordena alfabeticamente as palavras e leva essa ordenação em consideração ao comparar.

Após o cálculo de cada uma das quatro métricas em cada uma das duas duplas, são gerados oito métricas de semelhança de texto, cada uma com suas características. Durante o desenvolvimento do trabalho, notou-se que verificar se as siglas dos Nomes A e B poderia fazer com que houvesse mais uma métrica interessante. Assim, foi desenvolvido um algoritmo que atribui a cada *token* de cada nome uma probabilidade do mesmo ser a sigla da conferência. A distância Levenshtein entre as siglas com maior probabilidade de cada título configuram na nona métrica de semelhança.

Com as métricas calculadas, o desafio é criar uma lógica que atribua um grau de importância para cada. A escolha neste caso foi de criar uma rede neural artificial que recebe um vetor de 9 posições contendo cada uma das métricas calculadas. A rede neural foi definida com duas camadas escondidas, de tamanhos quinze e doze. A saída consiste em um vetor de duas posições, se a primeira for maior que a segunda, o modelo define que são nomes de

conferências diferentes, o caso contrário significa que os nomes passados para o modelo representam a mesma conferência. O processo completo descrito pode ser observado de maneira gráfica da figura 1.

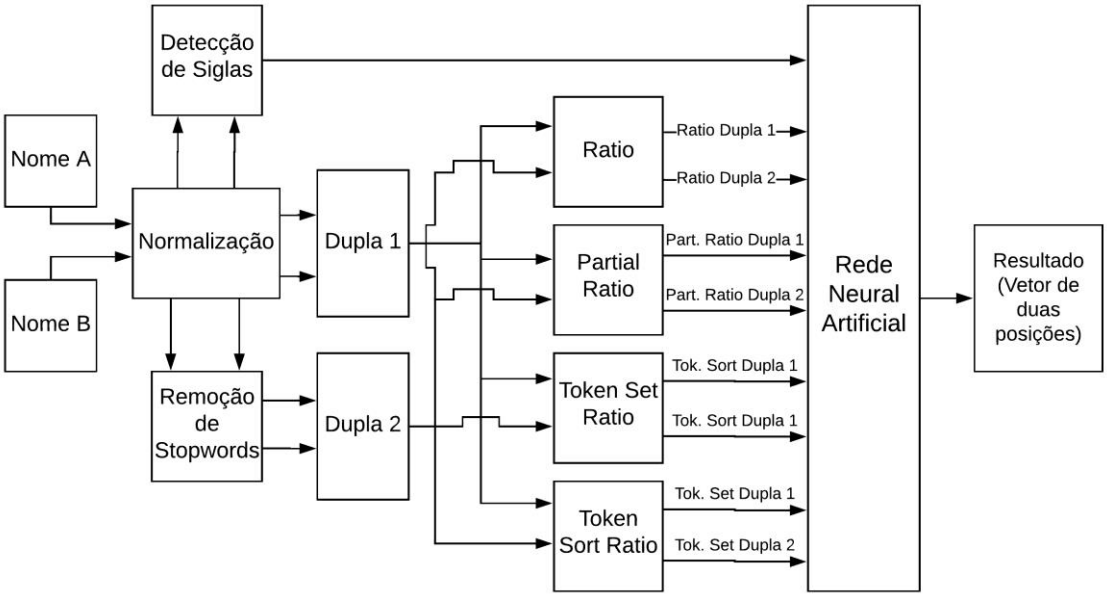

Figura 1. Fluxo dos dados partindo de cada um dos nomes de entrada até o resultado final gerado pela rede neural.

3. RESULTADOS E DISCUSSÃO

Para realizar os testes da rede, foi separado um conjunto de duzentos nomes únicos provenientes de currículos lattes. Destes, cento e cinquenta são nomes de conferências conhecidas e com Qualis atribuído, enquanto os outros cinquenta são de conferências que não são da área, não são qualificadas pela métrica, ou foram escritas incorretamente e não configuram como um nome de conferência. Para cada um desses nomes, a rede neural deveria comparar com cada nome da base de dados de conferências conhecidas com Qualis e verificar se existia correspondência com alguma delas ou não.

Dos cinquenta nomes que não tinham correspondência, a rede não atribuiu o nome a nenhum dos conhecidos da base de dados, gerando 100% de acerto. Já entre as demais, houve erro na avaliação de 27 e acerto em 123 casos, configurando em 82% de acerto.

Dos 27 erros, 20 ocorreram pois o algoritmo não retornou nenhuma conferência como a correta, sendo que deveria ter encontrado. Um dos casos foi o da “TISE – XII Conferência Internacional sobre Informática na Educação”. Nos outros 7, foi encontrada ao menos uma conferência da base de dados como a correta, mas não a correta, como o nome “BCB – ACM International Conference on Bioinformatics and Computational Biology and Health Informatics”, que foi avaliado como “BIOCOMP – International Conference on Bioinformatics and Computational Biology”.

Se levarmos em consideração que cada uma das duzentas avaliações consiste em comparar um nome com mais de dois mil da base de dados, a nível de comparações únicas, o algoritmo teria resultados ainda melhores.

4. CONCLUSÕES

Após observar os resultados obtidos, conclui-se que a técnica construída pode ser de grande ajuda para a avaliação automática ou ao menos auxiliar a avaliação manual, onde um avaliador poderia averiguar os resultados obtidos pelo algoritmo e atribuir conceitos de maneira muito mais rápida.

Quanto a técnica em si, pode ser objeto de estudos futuros o teste da mesma em domínios diferentes, ou a inserção de novos tipos de algoritmo de proximidade ao invés do Levenshtein, ou até mesmo em conjunto.

Também é possível observar que não foram utilizadas técnicas avançadas de processamento de linguagem natural, como redução aos radicais, ou identificação de classes gramaticais. Talvez com essas informações a mais, seja possível construir uma rede neural mais robusta, ou explorar outras técnicas mais simples, como árvores de decisão.

5. REFERÊNCIAS BIBLIOGRÁFICAS

CNPQ. **Histórico**. Plataforma Lattes, Entre 2005 e 2019. Acessado em 14 set. 2019. Online. Disponível em: http://memoria.cnpq.br/web/portal-lattes/historico?fbclid=IwAR1pYhIYDPFPbZcfRSX7lh3hNEQhHDt0kEvfbLvWbuIELaUHhhYMKJEMqU

WINKLER, W, Matching and Record Linkage. **Wiley Interdisciplinary Reviews: Computational Statistics**, v.6, 2014.

LEVENSHTEIN, V.I. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. **Soviet Physics – Doklady**, v.10, n.8, p.707 – 710, 1966.

RAJARAMAN, A; LESKOVEC, J; ULLMAN, J.D. **Mining of Massive Datasets**. Cambrige: Cambridge University Press, 2012. 2v.