# Algoritmo para Cálculo da Similaridade Semântica entre Nomes de Conferências Cadastradas no Curriculum Lattes e na Base Qualis

**Rafael Soares**[1], **Rafael Z. Frantz**[1]

[1]Universidade Regional do Noroeste do Estado do Rio Grande do Sul (UNIJUÍ)
Departamento de Ciências Exatas e Engenharias
Rua Lulu Ilgenfritz, 480 — 98700-000 — Ijuí/RS -– Brasil

rafael.dcs@unijui.edu.br, rzfrantz@unijui.edu.br

***Abstract.*** *The Lattes Platform is an environment in which researchers register data related to their academic activities, including publications in journals and events. In Brazil, these journals and events may have an evaluation indicator known as Qualis. Automating the extraction of data from journals published in the Lattes curriculum and obtaining the corresponding Qualis is a trivial task thanks to the use of ISSN as an identification key. However, when it comes to papers published in events, the only identification key available is the name of the event, which makes this search difficult. In this paper, we propose an algorithm for an efficient comparison of sentences and apply it in the comparison of event names so that it is possible to search Qualis for publications in events registered in a curriculum with the utmost accuracy.*

***Resumo.*** *A Plataforma Lattes é um ambiente no qual os pesquisadores registram dados relacionados às suas atividades acadêmicas, dentre elas as publicações em periódicos e eventos. No Brasil, esses periódicos e eventos podem ter um indicador de avaliação conhecido como Qualis. A automatização da extração de dados de publicações de periódicos do currículo Lattes e a obtenção do Qualis correspondente é uma tarefa trivial graças ao uso do ISSN como chave de identificação. No entanto, quando se trata de artigos publicados em eventos, a única chave de identificação disponível é o nome do evento, o que dificulta essa busca. Neste artigo, propomos um algoritmo para uma comparação eficiente de sentenças e o aplicamos na comparação de nomes de eventos para que seja possível buscar com o máximo de assertividade o Qualis para as publicações em eventos registradas em um currículo.*

## 1. Introdução

O Brasil é um dos poucos países no mundo que possui um sistema único e integrado para registro das produções científicas dos pesquisados. Esse sistema, conhecido como Plataforma Lattes, permite que cada pesquisador possa registar seus projetos, orientações, produções, detalhes sobre sua formação acadêmica, colaboração com outros pesquisadores, etc. Pesquisadores que possuem um currículo registrado nessa plataforma podem acrescentar diversos tipos diferentes de produções, tais como artigos em periódicos, artigos em eventos, livros e capítulos de livros. Outro sistema importante para pesquisadores é o sistema Qualis. Desenvolvido inicialmente em 1998 com a intenção de

ser usado como um dos critérios de avaliação de programas de pós-graduação das universidades brasileiras, grande parte das produções científicas são classificadas de acordo com esse sistema. Além de possuir uma base de dados com os mais importantes periódicos e suas classificações, o Qualis também classifica eventos de algumas áreas como a da Ciência Computação. Essa base de periódicos, eventos e classificações é disponibilizada online pela Capes através do Portal Sucupira [Sucupira 2020].

Uma atividade frequente nas instituições de ensino pelos setores administrativos é a construção de uma lista de publicações para um determinado pesquisador ou grupo de pesquisadores, sendo importante constar nesta lista para cada artigo, alguns indicadores de avaliação do foro onde os trabalhos foram publicados, tais como o Qualis, JCR, h-index, Quartil, dentre outros. Embora a extração das publicações dos currículos seja uma tarefa geralmente automatizada, o processo de enriquecimento de informações com os indicadores para cada artigo é feito de forma manual. Para artigos publicados em periódicos a busca dos valores para cada indicador de avaliação pode ser feita por meio do ISSN do periódico, já que as bases que gestionam esses indicadores costumam usá-lo como identificador único. Assim, um sistema automatizado pode encontrar a classificação de um periódico na base Qualis usando o ISSN do periódico como chave da busca. No entanto, essa tarefa não é trivial quando se trata de automatizar a busca dos indicadores para artigos publicados em eventos, uma vez que neste caso o nome do evento passa a ser o identificador único adotado pelas bases. Isso representa um desafio pois o nome de um evento cadastrado pelo pesquisador no seu currículo Lattes pode estar incompleto ou ligeiramente distinto de seu correspondente na base Qualis.

Uma sentença é tecnicamente representada por uma *string*, que, na maior parte das linguagens de programação, nada mais é do que um conjunto de caracteres de 1 byte que formam um texto a ser computado e pode ser utilizado para representar o nome de um evento. Uma comparação simples entre essas *strings*, ainda que utilizando métodos nativos da linguagem de programação, depende que ambas as *strings* comparadas possuam os mesmos caracteres na mesma ordem para que sejam consideradas idênticas. A similaridade semântica entre sentenças é um campo de pesquisa amplamente explorado na área de processamento de linguagem natural e com as mais diversas aplicações. [Ho et al. 2010] argumentam que muitos trabalhos tendem a calcular a similaridade entre sentenças a partir de uma análise individual das palavras que integram a sentença e seus significados próximos, mas que essa abordagem é falha pois nem sempre o significado próximo reflete o real significado da palavra em questão na sentença. No trabalho desses autores, eles propõe comparar duas sentenças a partir do significado real de cada palavra que compõe a sentença. Na proposta os autores mostram e comprovam estatisticamente que sua abordagem é mais eficaz. [Farouk 2018] propõe uma abordagem hibrida que combina o uso de um banco de dados léxico para o idioma inglês com a representação vetorial de palavras. Na sua abordagem o autor ainda propõe considerar a posição das palavras na sentença para resolver conflitos de significado. [Achananuparp et al. 2008] revisaram a literatura e analisaram um conjunto amplo de 14 métricas para calcular a similaridade entre duas sentenças. O trabalho destes autores comparou o uso destas métricas e, embora tenha sido realizado há alguns anos, ainda pode se constituir em um ponto de partida interessante para o estudo e desenvolvimento de novas métricas.

Neste artigo apresentamos um algoritmo que realiza uma comparação morfológica entre sentenças que, a partir da sequência em que as palavras estão dispostas na sentença e da sequência em que as letras estão dispostas nas palavras, utiliza um sistema de pontuação para quantificar a semelhança entre duas sentenças. Para avaliar nosso algoritmo o implementamos em um sistema real que extrai publicações de eventos de um currículo Lattes em formato XML e enriquece os dados de cada artigo com seu Qualis. Avaliamos a assertividade de nosso algoritmo a partir da comparação dos resultados com outra implementação que faz uma comparação de *strings* usando métodos nativos da linguagem de programação. Nosso algoritmo apresenta uma boa assertividade já que desconsidera as pequenas diferenças entre dois nomes do mesmo evento, porém não ao ponto de considerar dois eventos diferentes o mesmo. Trata-se de uma proposta preliminar, que acreditamos mesmo assim, possa ser utilizada no sistema real onde foi implantado.

O restante deste trabalho está organizado da seguinte forma: a Seção 2 detalhamos o algoritmo que desenvolvemos; na Seção 3 introduzimos o sistema real no qual implementamos o algoritmo para poder avaliá-lo e compará-lo com uma implementação simples de comparação de *strings*; na Seção 4 apresentamos nossas conclusões e discutimos os trabalhos futuros que pretendemos desenvolver para evoluir esta proposta.

## 2. Algoritmo Proposto

Nesta seção descreve-se o algoritmo desenvolvido na pesquisa. Uma implementação foi desenvolvida em PHP e pode ser acessada online [1]. As funções externas nativas da linguagem foram utilizadas na construção do algoritmo e estão descritas no pseudocódigo do Algoritmo 1 com assinaturas diferentes do encontrado na linguagem. Esta função recebe três entradas, o nome do evento que se encontra no curriculo Lattes do pesquisador, o nome do evento da Base Qualis e uma pontuação mínima para ajustar o limite entre o que será considerado mesmo evento ou evento diferente. O algoritmo retorna um valor booleano que representa a correspondência da comparação. Será verdadeiro se a pontuação for maior ou igual a pontuação mínima definida, e falso caso a pontuação seja menor. As funções externas utilizadas na construção deste algoritmo são: `explodir()`, que divide uma *string* em um ponto especificado pelo programador e retorna um vetor com as sub *strings*, `contar()`, que devolve um inteiro com a quantidade de itens de um vetor e `ord()`, que devolve o valor ASCII correspondente ao caractere de entrada ou do primeiro caractere de uma *string* de entrada. Nas linhas 3 e 4 a função `explodir()` é usada para separar os títulos nos caracteres de espaçamento, criando um vetor onde cada item é uma palavra.

A partir da linha 7 o algoritmo pode ser visto como três blocos distintos. O primeiro, que vai da linha 7 à linha 16 analisa a ordem das palavras em cada um dos nomes observando apenas a primeira letra de cada palavra. Em um laço de repetição que passa por todas os itens de `vetPalavrasLattes` é feito uma comparação com os itens de `vetPalabrasBase` usando `ord()`, de forma a definir se a primeira letra de cada palavra aparece na mesma ordem em ambos os vetores. Esse bloco tem o propósito de descartar rapidamente os nomes que possuem a menor probabilidade de coincidir, já que a base Qualis conta com 1179 títulos diferentes e uma comparação mais detalhada de cada nome pode tomar muito tempo e recurso de processamento.

[1] www.gca.unijui.edu.br/project/conference.php

**Algoritmo 1:** COMPARADOR DE TÍTULOS

**Entrada:** $conferenciaLattes, conferenciaBase, pontuacaoMinima$
**Saída:** Valor booleano que indica correspondência

```
1  início
2      pontos := 0, inicio := -1, i := 0
3      vetPalavrasLattes := explodir(' ', conferenciaLattes)
4      vetPalavrasBase := explodir(' ', conferenciaBase)
5      para j := 0 até contar(vetPalavrasLattes) faça
6          se ord(vetPalavrasLattes[j]) == ord(vetPalavrasBase[i]) então
7              i++
8              inicio == -1 ? inicio = j : inicio = inicio
9          senão
10             se i > 0 e i < contar(vetPalavrasBase) então
11                 i := 0
12             fim
13         fim
14     fim
15     se i == contar(vetPalavrasBase) então
16         para i := 0 até contar(vetPalavrasBase) faça
17             vetLetrasBase := explodir(vetPalavrasBase[i])
18             vetLetrasLattes := explodir(vetLetrasLattes[inicio])
19             contador := 0
20             para j := 0 até contar(vetLetrasBase) faça
21                 se ord(vetLetrasBase[j]) == ord(vetLetrasLattes[j])
                    então
22                     contador++
23                 fim
24             fim
25             pontos += contador
26             inicio++
27         fim
28     fim
29     pontos := pontos/tamanhoString(conferenciaBase)
30     se pontos >= pontuacaoMinima então
31         retorna TRUE
32     senão
33         retorna FALSE
34     fim
35 fim
```

A variável `i` é o contador que incrementará dependendo de quantas palavras estiverem em ordem. Se `i` tiver o mesmo valor que a quantidade de itens dentro de `vetPalavrasBase`, o segundo bloco, que vai da linha 17 à linha 30, é acionado. Caso contrário, o algoritmo irá considerar nomes diferentes e retornará falso. Dentro do segundo bloco cada palavra em cada vetor é comparada letra por letra com a sua possível correspondente seguindo a ordem definida no primeiro bloco pela variável `inicio`. Um

ponto será somado para cada letra que coincidir em ambos os nomes e esse somatório será levado ao terceiro bloco assim que finalizada esta etapa. No terceiro bloco um índice é calculado dividindo os pontos somados, ou a quantidade de letras que coincidem, pela quantidade de letras no total em `conferenciaBase`. Esse índice será comparado com a pontuação mínima definida em `pontuacaoMinima` pelo programador e retornará o valor booleano correspondente.

## 3. Estudo de Caso

Para demonstrar o algoritmo proposto foi utilizado o currículo Lattes de dez pesquisadores [2]. Os currículos foram lidos por um sistema automatizado criado previamente, que extraiu todos os artigos publicados em eventos por esses pesquisadores. O título de cada evento extraído foi comparado com a base Qualis utilizando tanto um método nativo de comparação quanto o algoritmo proposto. Uma pesquisa manual de cada evento do Lattes dos pesquisadores foi feita para avaliar a eficiência do algoritmo. No total 869 nomes de eventos foram comparados à base Qualis, sendo que 335 possuem uma classificação Qualis. A Tabela 1 demonstra os dados obtidos.

**Tabela 1. Quantidade de identificações de nomes de eventos com Qualis utilizando cada método.**

| Currículos | Publicações em Eventos | Publicações com Qualis | Técnicas de Comparação | |
|---|---|---|---|---|
| | | | Método Nativo | Algoritmo Proposto |
| Currículo 1 | 36 | 6 | 0 | 4 |
| Currículo 2 | 50 | 7 | 0 | 5 |
| Currículo 3 | 62 | 14 | 1 | 8 |
| Currículo 4 | 142 | 23 | 2 | 11 |
| Currículo 5 | 187 | 102 | 5 | 45 |
| Currículo 6 | 91 | 40 | 4 | 28 |
| Currículo 7 | 155 | 83 | 2 | 51 |
| Currículo 8 | 70 | 20 | 2 | 15 |
| Currículo 9 | 47 | 30 | 2 | 15 |
| Currículo 10 | 29 | 10 | 0 | 7 |
| **Total** | 869 | 335 | 18 | 189 |

No currículo 1, de 6 nomes de eventos que possuem Qualis, o algoritmo proposto identificou 4 e o método nativo, nenhum. No currículo 2, o algoritmo identificou 5 de 7 eventos com Qualis, comparado ao método nativo que novamente não encontrou nenhum. No currículo 3, entre 14 eventos com Qualis, o algoritmo identificou 8 e o método nativo, 1. No currículo 4, entre 23 eventos registrados que possuiam Qualis, 11 foram identificados pelo algoritmo proposto em contrapartida a 2 identificados pelo método nativo de comparação. No currículo 5, 187 nomes de eventos foram analisados no total, dentre eles 107 possuíam Qualis, de forma que o algoritmo proposto identificou 45 e o método nativo, 5 eventos. No currículo 6, entre 40 eventos encontrados na base Qualis, o algoritmo proposto identificou 28 e o método nativo identificou 4. No currículo 7 foram analisados 155 nomes de eventos sendo que destes, 83 possuem Qualis, o algoritmo proposto identificou 51 e o método nativo identificou apenas 2. De 20 nomes de eventos que possuem Qualis

[2] Currículos disponíveis em: www.gca.unijui.edu.br/publication/data/curriculos.zip

no currículo 8, o algoritmo proposto encontrou 15 e o método nativo, 2. O currículo 9, que possui 30 nomes de eventos encontrados na base Qualis, teve 15 nomes identificados pelo algoritmo proposto e apenas 2 pelo método nativo. Por fim, no currículo 10 o algoritmo proposto encontrou 7 dos 10 eventos que possuem Qualis e o método nativo não encontrou nenhum.

A pontuação mínima utilizada no algoritmo foi 0.8, ou seja, seriam considerados mesmo evento se 80% das letras em um nome de evento na base Qualis fossem encontradas em um nome de evento registrado em um currículo. Dentre os 335 eventos que possuem Qualis encontrados entre os currículos Lattes desses pesquisadores, considerando eventos duplicados entre currículos, um total de 189 nomes foram identificados pelo algoritmo proposto, representando cerca de 56% de acerto total na comparação de nomes com uma média de acerto de 61,32%. O método nativo de comparação oferecido pela linguagem de programação utilizada identificou um total de 18 nomes de eventos, cerca de 5% de acerto total, tendo uma média de acerto de 4,98%.

## 4. Conclusão

Neste artigo foi apresentado uma proposta preliminar de algoritmo para comparação de nomes de eventos com o intuito de classificar artigos registrados no currículo Lattes de pesquisadores utilizando a base Qualis e outros indicadores de avaliação de forma automatizada. Este algoritmo foi desenvolvido de forma a fazer uma análise morfológica das palavras, ou seja, comparar a estrutura de cada palavra e sua ordem na sentença com seu possível correspondente. Foram utilizados currículos de pesquisadores de diferentes instituições que possuem grande número de trabalhos publicados em eventos voltados à computação pois é visível que nesta área a base Qualis de eventos é mais completa. Estes currículos foram analisados por um sistema automatizado de extração de dados, porém cada nome de evento foi também comparado manualmente com a base Qualis para que fosse possível avaliar a eficiência do algoritmo proposto.

A partir dos resultados obtidos no estudo de caso, é possível observar a grande diferença de eficiência de um método nativo de comparação e o algoritmo desenvolvido nesta pesquisa, considerando que este algoritmo é uma proposta preliminar de solução. Porém, apesar de se mostrar uma solução mais eficiente, a analise dos resultados obtidos mostrou quais são os problemas do algoritmo e as possíveis melhorias que podem ser feitas. Uma das dificuldades apresentadas pelo algoritmo é identificar a sigla dos eventos, sendo que muitas vezes pesquisadores registram apenas isso como nome do evento participado. Outro problema está em um pesquisador registrar o nome de um evento em um idioma diferente do registrado na base Qualis, característica que foi percebida na coleta de dados graças ao registro conjunto da sigla.

Como trabalho futuro deve-se resolver os problemas encontrados na proposta atual do algoritmo, onde um identificador de siglas pode ser adicionado junto com alguns ajustes técnicos necessários. Novas técnicas de comparação que avaliam a similaridade semântica de sentenças podem ser incrementadas ao algoritmo no futuro de forma a aumentar seu índice de acerto e disponibilizar novas métricas para que haja uma avaliação mais precisa de seu funcionamento. Eventualmente pode-se até mesmo desenvolver um método baseado em inteligência artificial, sendo que a utilização de cada nova versão do algoritmo será comparada com a de suas versões anteriores, servindo como guia para

alcançar a maior eficiência.

## Agradecimentos

Este trabalho tem o apoio da Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS) no contexto do Programa Pesquisador Gaúcho, com termo de outorga número 17/2551-0001206-2 e do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pela bolsa de Iniciação Científica.

## Referências

Achananuparp, P., Hu, X., and Shen, X. (2008). The evaluation of sentence similarity measures. In *Data Warehousing and Knowledge Discovery*, pages 305–316.

Farouk, M. (2018). Sentence semantic similarity based on word embedding and wordnet. In *International Conference on Computer Engineering and Systems (ICCES)*, pages 33–37.

Ho, C., Murad, M. A. A., Kadir, R. A., and Doraisamy, S. C. (2010). Word sense disambiguation-based sentence similarity. In *International Conference on Computational Linguistics: Posters (COLING)*, pages 418—-426.

Sucupira (2020). Portal sucupira. https://sucupira.capes.gov.br. Último acesso em 05/03/2020.