UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

ROBSON CARDOSO VIEIRA

Consultas com Palavras-chave em Bancos de Dados Relacionais Descritos por Metadados Multilíngues

Goiânia
2021

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO (TECA) PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TESES

E DISSERTAÇÕES NA BIBLIOTECA DIGITAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio da Biblioteca Digital de Teses e Dissertações (BDTD/UFG), regulamentada pela Resolução CEPEC nº 832/2007, sem ressarcimento dos direitos autorais, de acordo com a Lei 9.610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou download, a título de divulgação da produção científica brasileira, a partir desta data.

O conteúdo das Teses e Dissertações disponibilizado na BDTD/UFG é de responsabilidade exclusiva do autor. Ao encaminhar o produto final, o autor(a) e o(a) orientador(a) firmam o compromisso de que o trabalho não contém nenhuma violação de quaisquer direitos autorais ou outro direito de terceiros.

1. Identificação do material bibliográfico

[x] Dissertação [ ] Tese

2. Nome completo do autor

Robson Cardoso Vieira

3. Título do trabalho

Consultas com Palavras-chave em Bancos de Dados Relacionais Descritos por Metadados Multilíngues

4. Informações de acesso ao documento (este campo deve ser preenchido pelo orientador)

Concorda com a liberação total do documento [x] SIM [ ] NÃO[1]

[1] Neste caso o documento será embargado por até um ano a partir da data de defesa. Após esse período, a possível disponibilização ocorrerá apenas mediante:
a) consulta ao(à) autor(a) e ao(à) orientador(a);
b) novo Termo de Ciência e de Autorização (TECA) assinado e inserido no arquivo da tese ou dissertação.
O documento não será disponibilizado durante o período de embargo.
Casos de embargo:
- Solicitação de registro de patente;
- Submissão de artigo em revista científica;
- Publicação como capítulo de livro;
- Publicação da dissertação/tese em livro.

Obs. Este termo deverá ser assinado no SEI pelo orientador e pelo autor.

Documento assinado eletronicamente por **ROBSON CARDOSO VIEIRA**, **Discente**, em 05/04/2021, às 15:03, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **João Carlos da Silva**, **Usuário Externo**, em 05/04/2021, às 15:28, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **1982121** e o código CRC **EFEC0A95**.

Referência: Processo nº 23070.011847/2021-24 SEI nº 1982121

ROBSON CARDOSO VIEIRA

Consultas com Palavras-chave em Bancos de Dados Relacionais Descritos por Metadados Multilíngues

Dissertação apresentada ao Programa de Pós–Graduação do Instituto de Informática da Universidade Federal de Goiás, como requisito parcial para obtenção do título de Mestre em Ciência da Computação.

Área de concentração: Ciência da Computação.

Orientador: Prof. Dr. João Carlos da Silva

Goiânia
2021

Ficha de identificação da obra elaborada pelo autor, através do
Programa de Geração Automática do Sistema de Bibliotecas da UFG.

Vieira, Robson Cardoso
Consultas com Palavras-chave em Bancos de Dados Relacionais Descritos por Metadados Multilíngues [manuscrito] / Robson Cardoso Vieira. - 2021.
XCVI, 96 f.: il.

Orientador: Prof. Dr. João Carlos da Silva.
Dissertação (Mestrado) - Universidade Federal de Goiás, Instituto de Informática (INF), Programa de Pós-Graduação em Ciência da Computação, Goiânia, 2021.
Bibliografia. Anexos. Apêndice.

1. Metadados Multilíngues. 2. Banco de Dados Relacionais. 3. Consulta com Palavras-chave. I. Silva, João Carlos da , orient. II. Título.

CDU 004

UNIVERSIDADE FEDERAL DE GOIÁS

INSTITUTO DE INFORMÁTICA

ATA DE DEFESA DE DISSERTAÇÃO

Ata nº **06/2021** da sessão de Defesa de Dissertação de **Robson Cardoso Vieira**, que confere o título de Mestre em Ciência da Computação, na área de concentração em Ciência da Computação.

Aos vinte e cinco dias do mês de março de dois mil e vinte e um, a partir das nove horas e trinta minutos, via sistema de webconferência da RNP, realizou-se a sessão pública de Defesa de Dissertação intitulada **"Consultas com Palavras-chave em Bancos de Dados Relacionais Descritos por Metadados Multilíngues"**. Os trabalhos foram instalados pelo Orientador, Professor Doutor João Carlos da Silva (INF/UFG) com a participação dos demais membros da Banca Examinadora: Professor Doutor Plínio de Sá Leitão Júnior (INF/UFG), membro titular interno; Professor Doutor Fábio Nogueira de Lucena (INF/UFG), membro titular externo. A realização da banca ocorreu por meio de videoconferência, em atendimento à recomendação de suspensão das atividades presenciais na UFG emitida pelo Comitê UFG para o Gerenciamento da Crise COVID-19, bem como à recomendação de isolamento social da Organização Mundial de Saúde e do Ministério da Saúde para enfrentamento da emergência de saúde pública decorrente do novo coronavírus. Durante a arguição os membros da banca não fizeram sugestão de alteração do título do trabalho. A Banca Examinadora reuniu-se em sessão secreta a fim de concluir o julgamento da Dissertação, tendo sido o candidato **aprovado** pelos seus membros. Proclamados os resultados pelo Professor Doutor João Carlos da Silva, Presidente da Banca Examinadora, foram encerrados os trabalhos e, para constar, lavrou-se a presente ata que é assinada pelos Membros da Banca Examinadora, aos vinte e cinco dias do mês de março de dois mil e vinte e um.

TÍTULO SUGERIDO PELA BANCA

Documento assinado eletronicamente por **Plinio De Sa Leitão Junior**, **Professor do Magistério Superior**, em 25/03/2021, às 12:21, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Fábio Nogueira De Lucena**, **Professor do Magistério Superior**, em 25/03/2021, às 12:21, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **João Carlos da Silva**, **Usuário Externo**, em 25/03/2021, às 12:22, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **ROBSON CARDOSO VIEIRA**, **Discente**, em 25/03/2021, às 12:36, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Referência: Processo nº 23070.011847/2021-24 SEI nº 1921319

Robson Cardoso Vieira

Possui graduação em Sistemas de Informação na Faculdade Alves Faria (2006). Pós-graduação em Tecnologia em Gestão da Informação na Faculdade Anhanguera de Anápolis, em Administração Pública: Ênfase em Controle Interno e Externo pela Faculdade Alfredo Nasser e MBA em Qualidade e Processos pela Fundação Getúlio Vargas. É Auditor de Controle Externo na área de Informática do Tribunal de Contas dos Municípios do Estado de Goiás atuando com análise e integração de dados e *business intelligence*.

Dedico esse trabalho à minha família e aos colegas que de alguma forma contribuíram com o desenvolvimento desse estudo.

Agradecimentos

A minha esposa pela paciência, dedicação, apoio e por ter me acompanhado lado a lado nessa jornada.

Aos meus pais que foram fundamentais e a base para minha formação.

Aos colegas da turma de mestrado em Ciência da Computação pelo compartilhamento de conhecimento e amizade.

Aos professores do Instituto de Informática por terem fornecido os conhecimentos fundamentais para a minha formação.

Ao meu orientador Professor Doutor João Carlos da Silva pela oportunidade de trabalhar com o tema e pelo empenho em contribuir com o trabalho.

Ao Professor Doutor Sérgio Teixeira de Carvalho que foi meu primeiro contato como aluno de mestrado, ainda como aluno especial, sendo fundamental para motivar a persistir nessa jornada.

A Professora Doutora Márcia Rodrigues Capelle Santana que não mediu esforços para ensinar a disciplina mais temida do curso.

Ao Programa de Pós-Graduação em Ciência da Computação, da Universidade Federal de Goiás, que ofertou o curso de mestrado e disponibilizou os recursos necessários para a formação.

E a todas as pessoas que de forma direta ou indireta contribuíram para a realização desse trabalho.

You can have data without information, but you cannot have information without data.

Daniel Keys Moranr,
Computer programmer and science fiction writer.

Resumo

Vieira, Robson Cardoso. **Consultas com Palavras-chave em Bancos de Dados Relacionais Descritos por Metadados Multilíngues**. Goiânia, 2021. 98p. Dissertação de Mestrado. Instituto de Informática, Universidade Federal de Goiás.

O acesso globalizado à informação demanda soluções que permitam o acesso a fontes de dados descritas em diferentes idiomas. Ademais, grande parte da informação disponível no mundo está armazenada em bancos de dados relacionais. As principais alternativas que permitem consultar essas fontes de informação são monolíngues ou necessitam de interação do usuário. O presente estudo propõe uma solução que possibilita a um usuário, sem conhecimento prévio de esquemas ou linguagem de consulta, acessar dados em bancos de dados relacionais descritos por metadados multilíngues, utilizando uma técnica de consultas por palavras-chave. A solução proposta executa o pré-processamento da consulta expandindo e traduzindo os termos que serão mapeados para os metadados nos idiomas disponíveis, visando preservar a semântica pretendida pelo usuário na consulta. Essa solução apresentou resultados promissores para consultas em bancos de dados descritos por metadados multilíngues, demonstrando a viabilidade de consultar informações em bancos de dados descritos em idioma diferente do utilizado na consulta inicial.

Palavras–chave

Metadados Multilíngues, Banco de Dados Relacionais, Consulta com Palavras-chave.

Abstract

Vieira, Robson Cardoso. **Keywords Query in Relational DataBases Described by Multilingual Metadata**. Goiânia, 2021. 98p. MSc. Dissertation. Instituto de Informática, Universidade Federal de Goiás.

The globalized access to information demands solutions that allow access to data sources described in different languages. Furthermore, much of the information available in the world is stored in relational databases. The main alternatives that allow consulting these sources of information are monolingual or require user interaction. The present study proposes a solution that allows a user without prior knowledge of schemas or query language, to access data in relational databases described by multilingual metadata, using keyword query technique. The proposed solution performs the pre-processing of the query, expanding and translating the terms that will be mapped to the metadata in the available languages, aiming to preserve the semantics intended by the user in the query. This solution presented promising results for queries in databases described by multilingual metadata, demonstrating the feasibility of consulting information in databases described in a language different from that used in the initial consultation.

Keywords

Multilingual Metadata, Relational Databases, Keywords Query.

Sumário

Lista de Figuras **15**

Lista de Tabelas **16**

Lista de Códigos do Programas **17**

1 Introdução **18**
- 1.1 Contexto e Motivação 18
- 1.2 Organização do Texto 20
- 1.3 Principais Problemas 21
- 1.4 Objetivos 21
- 1.5 Metodologia 22
- 1.6 Organização do Texto 22

2 Fundamentação Teórica **23**
- 2.1 Consultas em Bancos de Dados Relacionais 23
  - 2.1.1 Consultas baseadas em Palavras-Chave 23
  - 2.1.2 Consultas baseadas em Linguagem Natural 25
- 2.2 Processamento de Linguagem Natural 26
  - 2.2.1 Identificação da Linguagem 26
  - 2.2.2 Segmentação 27
  - 2.2.3 *Stopwords* 27
  - 2.2.4 Extração do Radical 28
  - 2.2.5 Expansão 29
  - 2.2.6 Tradução 30
- 2.3 Considerações Finais 32

3 Trabalhos Relacionados **34**
- 3.1 Metodologia 34
- 3.2 Condução 35
- 3.3 Análise dos Resultados 36
- 3.4 Considerações Finais 39

4 Solução Proposta **40**
- 4.1 Consultas em Bancos de Dados Relacionais 40
- 4.2 Pré-Processamento da Consulta 41
  - 4.2.1 Segmentação da Consulta 42
  - 4.2.2 Identificação de Funções 44
  - 4.2.3 Remoção de *Stopwords* 45

4.2.4 Extração do Radical 46
4.2.5 Tradução da Consulta 48
4.2.6 Expansão da Consulta 49
4.2.7 Exemplos 50
4.3 Considerações Finais 52

5 Avaliação da Solução Proposta **53**
5.1 Implementação 53
5.2 Base de Dados 58
5.3 Consultas 61
5.4 Métricas 63
5.5 Execução das consultas 64
5.6 Análise dos Resultados 66
5.7 Ameaças à validade 73
5.8 Considerações Finais 73

6 Conclusão **75**
6.1 Contribuições e limitações 75
6.2 Trabalhos Futuros 78

Referências Bibliográficas **79**

Apêndices **85**

A Consultas **85**

B Resultados Esperados **89**

C Resultados Encontrados **91**

Anexos **95**

I Configuração do Banco de Dados **95**

Lista de Figuras

1.1 Processo de consulta multilíngue 20

3.1 Estudos importados [56] 36

4.1 Arquitetura do SQUIRREL [46] 41
4.2 Solução proposta 43
4.3 Nuvem de palavras das consultas iniciais 46
4.4 Técnicas de extração do radical 47

5.1 Interface da solução proposta 54
5.2 Estrutura de dados do IMDB 58
5.3 Diagrama de Entidade Relacionamento do IMDB adaptado de [36] 60
5.4 Ilustração das métricas *precision* e *recall* 64
5.5 Tipo de execução das consultas 67
5.6 Tempo de execução por tipo de consulta 68
5.7 Mapeamentos gerados por tipo de execução 69
5.8 Pré-processamento esperado e realizado 69
5.9 Subconjunto de dados das Métricas 71
5.10 Média das métricas para as 8 consultas com resultados 71

Lista de Tabelas

1.1 Tabela de matrícula 20

3.1 Extração de dados das publicações aceitas [56] 37

4.1 Exemplo de pré-processamento 51

5.1 Conjunto de dados IMDB 59
5.2 Dados do IMDB em Português 60

A.1 Consultas baseadas em Lemos [36] 86
A.1 Consultas baseadas em Lemos [36] 87
A.1 Consultas baseadas em Lemos [36] 88

B.1 Resultados Esperados 89
B.1 Resultados Esperados 90

C.1 Resultados Encontrados 92
C.1 Resultados Encontrados 93
C.1 Resultados Encontrados 94

Lista de Códigos dos Programas

I.1 Código SQL para criação da tabela filme do IMDB 95
I.2 Código SQL para criação da tabela pessoa do IMDB 95
I.3 Código SQL para criação da tabela genero do IMDB 95
I.4 Código SQL para criação da tabela episodio do IMDB 96
I.5 Código SQL para criação da tabela ator_filme do IMDB 96
I.6 Código SQL para criação da tabela diretor_filme do IMDB 96
I.7 Código SQL para criação da tabela genero_filme do IMDB 96
I.8 Código SQL para criação da tabela TME 97
I.9 Código SQL para inserção do IMDB na tabela TME 98

CAPÍTULO 1

Introdução

Este capítulo apresenta o contexto, os problemas, os objetivos, a metodologia e a organização do trabalho. Está organizado em cinco seções, quais sejam: Seção 1.1 apresenta o contexto e a motivação que deu origem a pesquisa; Seção 1.3 destaca os principais problemas abordados no trabalho; Seção 1.4 lista os objetivos gerais e específicos propostos na pesquisa; Seção 1.5 especifica a metodologia utilizada no desenvolvimento do trabalho; Seção 1.6 descreve a estrutura organizacional desse estudo.

1.1 Contexto e Motivação

A globalização do acesso à Internet e o desenvolvimento tecnológico cada vez mais acentuado - provendo dispositivos com inteligência, geração e armazenamento de dados - crescem constantemente o volume de dados armazenados e a necessidade por soluções mais eficientes de recuperação destes dados. Uma parte considerável desse volume de dados gerados globalmente está armazenada em bancos de dados relacionais.

Diante desse contexto, esse trabalho tem como foco a consulta em bancos de dados relacionais, por serem os mais utilizados em sistemas e aplicativos para armazenamento de dados corporativos e institucionais, apesar de haver um crescimento no uso de modelos não relacionais.

Conforme informações disponibilizadas no site *db-engines*[1], em 2 de fevereiro de 2021, dentre os cinco sistemas de gerenciamento de banco de dados (SGBD) mais populares, quatro utilizam o modelo relacional. Desde a última década foi observado um incremento contínuo de dados estruturados disponíveis na Web e fontes de dados estruturados prometem ser o próximo impulsionador de um impacto sócio-econômico significativo para pessoas e empresas [9].

Os dados estruturados armazenados em bancos de dados relacionais tradicionalmente são inacessíveis por usuários externos [9], ou seja, não são indexados por mecanismos de busca. Isso acontece porque o acesso a essas informações exige do usuário

[1] https://db-engines.com

um prévio conhecimento de uma linguagem de consulta estruturada como SQL e das estruturas de dados (nomes de relações, atributos, relacionamentos chave primária ou chave estrangeira, entre outros). Assim, o acesso fica limitado a especialistas que dominam estas tecnologias, tais como programadores e administradores de banco de dados.

A fim de superar essa limitação, surgiram várias propostas que permitem realizar uma consulta em bancos de dados relacionais baseada em palavras-chave [8], [22], [27], [33], [46] e [60] ou baseada em linguagem natural [3], [35], [40] e [52]. Essas propostas possibilitam aos usuários, sem conhecimento específico de uma linguagem de consulta como SQL ou da estrutura de armazenamento dos dados, realizarem consultas utilizando expressões em linguagem natural ou palavras-chave e obter os resultados desejados que estão armazenados.

No entanto, a maioria das propostas que consultam bancos de dados relacionais são monolíngues, ou seja, os metadados consultados se restringem ao idioma utilizado na consulta inicial. Essa situação reduz a possibilidade de obter resultados precisos e que correspondam a intenção do usuário, pois não inclui bancos de dados descritos em idioma diferente do utilizado na consulta.

Em sistemas monolíngues para consultar bancos de dados descritos por metadados em outros idiomas, é necessário reescrever a consulta em diversos idiomas para que o usuário encontre os resultados. Diante disso, se torna evidente a necessidade de transcender as barreiras do idioma de tal forma a possibilitar o acesso a fontes de dados descritas em múltiplos idiomas e não apenas no idioma da consulta.

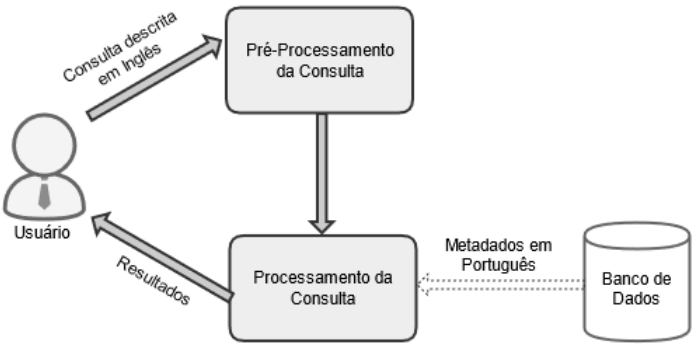
A busca por métodos que realizam consultas em bancos de dados descritos por metadados em múltiplos idiomas - consulta multilíngue - é uma tarefa importante para potencializar o acesso à informação desejada pelo usuário no momento da realização da consulta inicial.

A consulta em linguagem natural potencializa o processo de descrição da consulta pelo usuário. Isso porque habilita o usuário final comunicar com o sistema na sua linguagem nativa [19]. Nesse sentido, a consulta inicial deve passar por uma etapa de pré-processamento com o intuito de normalizar, expandir e traduzir a sequência de palavras utilizadas.

A Figura 1.2 ilustra o processo de consulta multilíngue em que a consulta é descrita em um determinado idioma e um banco de dados descrito por metadados em outro idioma é consultado. Nesse processo, o usuário submete a consulta, o pré-processamento realiza diversos tratamentos nos termos da consulta e identifica o idioma disponível para traduzir os termos da consulta inicial. Em seguida o processamento da consulta realiza o mapeamento entre os termos pré-processados e já traduzidos para os metadados do banco de dados.

Para exemplificar uma consulta a Tabela 1.1, contendo dados fictícios de alunos

Figura 1.1: Processo de consulta multilíngue



1.2 Organização do Texto

matriculados será utilizada como referência. Essa tabela, está descrita por metadados na língua portuguesa e contém dados nesse mesmo idioma. Ao realizar a consulta 'How many students are enrolled in the Administration Course?' um dos possíveis resultados do pré-processamento será formado pelo conjunto de palavras 'estudante, aluno, acadêmico, registro, inscrito, matricula, curso, aula, classe, administração, gerenciamento, gestão'. Ao processar este conjunto de palavras, frente aos metadados disponíveis, deverá retornar o resultado contendo a quantidade de alunos matriculados no curso de Administração.

Tabela 1.1: Tabela de matrícula

matricula	**nome**	**curso**
123456	Miguel Silva	Administração
234567	Arthur dos Santos	Biomedicina
345678	Heitor Oliveira	Ciências da Computação
456789	Bernardo Souza	Direito
567890	Davi Rodrigues	Enfermagem
678901	Gabriel Ferreira	Fisioterapia

Frente à relevância desse tipo de método de consulta, o desenvolvimento de mecanismos de busca que possibilitam consultar fontes de dados descritas em múltiplos idiomas, potencializa a obtenção resultados que estejam relacionados com a intenção do usuário, pois permite expandir as fontes de dados que poderão ser consultadas.

Baseado nesses apontamentos, a proposta que motiva este estudo é a de possibilitar que usuários leigos, ou seja, sem conhecimento prévio da estrutura de dados e da linguagem de consulta formal de bancos de dados relacionais, consigam obter resulta-

dos em múltiplos idiomas utilizando como consulta inicial uma sentença em linguagem natural ou palavras-chave no idioma Inglês.

1.3 Principais Problemas

Os principais trabalhos que abordam as consultas utilizando palavras-chave ou linguagem natural em bancos de dados relacionais são monolíngues [36, 40, 46], ou seja, a consulta fornecida é realizada usando apenas as fontes de dados descritas no idioma utilizado na consulta inicial. Várias dessas abordagens são dependentes de domínio [19, 37, 44], ou seja, o funcionamento é limitado ao contexto da base de dados utilizada na pesquisa dificultando a portabilidade para outros domínios.

Outro problema comum é que parte desses estudos não possuem uma especificação clara e detalhada das etapas e métodos de pré-processamento necessárias para apresentar os resultados ao usuário [19, 50]. Dentre as abordagens multilíngues é predominante a necessidade de interação do usuário com o sistema para realizar o processo de consulta [19, 51].

Ademais, não há proposta de uma solução que aborde o pré-processamento de forma detalhada que realize consultas em metadados multilíngues nos bancos de dados relacionais utilizando expressões em linguagem natural ou palavras-chave em que o usuário informe uma consulta inicial em um idioma e obtenha resultados em outros idiomas automaticamente. Logo, permanece em aberto um problema que diz respeito à consulta em bancos de dados relacionais utilizando linguagem natural ou palavras-chave para a obtenção de resultados multilíngues.

1.4 Objetivos

O objetivo geral dessa pesquisa consiste em desenvolver uma solução que possibilite ao usuário, sem prévio conhecimento da linguagem de consulta estruturada (SQL), dos metadados e do domínio dos dados de banco de dados relacionais, realizar uma consulta inicial na língua inglesa – em linguagem natural ou palavras-chave – e obter resultados correspondentes de bancos de dados relacionais independente da língua usada para descrever os metadados.

Os objetivos específicos, por sua vez, são: identificar técnicas de processamento de linguagem natural para realizar o pré-processamento de consultas multilíngues em bancos de dados relacionais; definir um método para mapeamento da consulta inicial em linguagem de consulta SQL multilíngue; implementar e validar a solução proposta; identificar o estado da arte em relação ao objeto de estudo.

1.5 Metodologia

O desenvolvimento desse trabalho se divide em três fases principais, a saber:

1. **Fundamentação teórica e trabalhos relacionados:** Nessa fase foi realizado o levantamento dos principais conceitos necessários para a compreensão da solução proposta realizando um nivelamento teórico do assunto. Foi reunido um conjunto de referências que tratam do tema da pesquisa e que formam a base teórica do trabalho. Para tanto, foi feita uma revisão da literatura contendo os principais estudos que lidam com consulta multilíngue, consulta baseada em palavras-chave e consulta baseada em linguagem natural utilizando bancos de dados relacionais.
   Todo o processo descrito, resultou na constituição do aporte teórico que forma a base científica dessa pesquisa, de tal modo que as demais fases desse trabalho são pautadas no arcabouço teórico construído nessa etapa.
2. **Proposição da Abordagem:** propõe uma solução com componentes que são utilizados para realizar o pré-processamento de consultas multilíngues em banco de dados relacionais conforme as limitações, características identificadas e teorias descritas na etapa de fundamentação teórica e trabalhos relacionados.
3. **Implementação e Avaliação da Abordagem:** apresenta os detalhes da implementação da solução proposta para realizar consultas em bancos de dados relacionais com metadados multilíngues, a fonte de dados e as consultas utilizadas no experimento, as métricas utilizadas para avaliar os experimentos, bem como a análise dos principais resultados obtidos e as ameaças à validade do trabalho.
4. **Análise de resultados:** A última fase apresenta um resumo dos resultados obtidos, as principais contribuições do estudo, as limitações que foram realizadas nos experimentos para validar a solução proposta e os trabalhos que podem ser realizados futuramente para aperfeiçoar a solução proposta.

1.6 Organização do Texto

O presente trabalho está estruturado em cinco capítulos. Assim, para além deste capítulo introdutório, a pesquisa se divide em: Capítulo 2 que apresenta a fundamentação teórica sobre consultas multilíngues em banco de dados relacionais; Capítulo 3 que apresenta os trabalhos relacionados ao objetivo da pesquisa; Capítulo 4 demonstra a solução proposta e as etapas de pré-processamento das consultas multilíngues; Capítulo 5 aborda a implementação da solução e os resultados obtidos; Capítulo 6 que aponta as conclusões obtidas na investigação, bem como suas contribuições, limitações e trabalhos futuros.

CAPÍTULO 2

Fundamentação Teórica

Este capítulo apresenta as principais teorias utilizadas e necessárias para a compreensão da solução proposta. A referida solução possibilita ao usuário realizar uma consulta utilizando linguagem natural ou palavras-chave que será mapeada automaticamente nos metadados dos bancos de dados relacionais para recuperar resultados em múltiplos idiomas. Para tanto, são utilizadas técnicas de Processamento de Linguagem Natural (PLN) para pré-processar e traduzir a consulta inicial.

2.1 Consultas em Bancos de Dados Relacionais

O acesso às informações armazenadas em banco de dados relacionais é tradicionalmente realizado usando linguagens formais de consulta como SQL [5]. Para disponibilizar o acesso a essas informações para usuários que não possuem conhecimento da linguagem formal de consulta aos bancos de dados nem das estruturas de armazenamento desses dados, geralmente, são utilizadas duas abordagens principais que são as consultas baseadas em palavras-chave e as consultas baseadas em linguagem natural descritas nas Seções 2.1.1 e 2.1.2.

2.1.1 Consultas baseadas em Palavras-Chave

As consultas por palavras-chave em bancos de dados relacionais têm sido largamente aplicadas devido à sua simplicidade e facilidade de uso. Elas são populares porque são intuitivas, fáceis de expressar e permitem o ranqueamento rápido [6]. A consulta baseada em palavras-chave visa a recuperar as tuplas do banco de dados que correspondem às palavras-chave do usuário. O núcleo desses sistemas é a etapa de pesquisa na qual os sistemas tentam corresponder as palavras-chave fornecidas com um índice invertido da base e dos metadados [2].

Segundo [9] as principais técnicas adotadas nas consultas baseadas em palavra-chave são:

- **Baseada em grafos:** a técnica baseada em grafos [1] modela os bancos de dados relacionais como grafos, onde os nós são tuplas e as arestas são relacionamentos entre a chave primária e a chave estrangeira dessas tuplas. Seu principal objetivo é otimizar o cálculo de estruturas específicas sobre grafos para encontrar as tuplas mais relevantes.
- **Baseada em esquema:** a técnica baseada em esquema explora as informações do esquema para formular consultas SQL a partir das consultas por palavras-chave do usuário [30]. Nesse caso, o sistema precisa descobrir as estruturas de dados que contêm as palavras-chave para formular consultas SQL. Essa técnica geralmente é utilizada quando há acesso ao esquema do banco de dados a ser consultado.
- **Baseada em metadados:** a técnica baseada em metadados é útil quando não há acesso prévio ao esquema do banco de dados ou quando atualizações frequentes tornam o processo de criação e atualização de índices muito dispendioso [9]. Um armazém de dados contendo todos metadados disponíveis foi construído em [11]. Uma tabela de metadados para exposição foi usada em [46] para descrever as fontes de dados disponíveis para consultas.

Conforme [10], a consulta por palavras-chave em bancos de dados tornou-se uma área de pesquisa promissora e atraiu pesquisadores das áreas de banco de dados, de recuperação da informação, de teorias, entre outras. Mas, segundo [6] expressar a intenção do usuário na consulta usando poucas palavras restringe a semântica do que pode ser expresso. Por isso, a consulta com palavras-chave em bancos de dados relacionais tem demonstrado ser um problema complexo e desafiador.

A recuperação de informações usando somente palavras-chave não é usualmente muito eficiente [47], pois uma informação pode ser representada com diferentes palavras-chave que podem não coincidir exatamente com os termos informados na consulta pelo usuário devido a ambiguidade. Mas, usando a técnica de expansão da consulta, descrita na Seção 2.2.5, é possível enriquecer e ampliar a consulta adicionando novos termos relacionados para melhorar a efetividade da consulta e obter informações relevantes complementares.

Ademais, há situações em que a consulta baseada em palavras-chave não consegue expressar a verdadeira intenção do usuário. Por exemplo, na consulta *'What is the best movie of each genre?'* uma consulta correspondente baseada em palavras-chave seria *'best movie genre'* que é mais provável de ser interpretada como *'the genre of the best movie'* [2]. Nesses casos, o uso de linguagem natural (ver Seção 2.1.2) para expressar a intenção do usuário na consulta inicial pode mitigar o problema.

2.1.2 Consultas baseadas em Linguagem Natural

As consultas baseadas em linguagem natural permitem que o usuário acesse as informações armazenadas em um banco de dados digitando consultas expressas em alguma linguagem natural [5]. Logo é uma forma mais fácil para os usuários realizarem consultas, uma vez que dispensa o conhecimento prévio sobre padrões específicos de nomenclatura - termos que podem ou não serem utilizados na consulta - para obter os resultados pretendidos na busca.

Os primeiros sistemas de consulta baseada em linguagem natural - *Natural Language Interfaces* (NLI) - para banco de dados relacionais surgiram entre os anos 60 e 70 e o mais popular foi o LUNAR [58]. o LUNAR é um sistema de perguntas e respostas que utiliza um banco de dados contendo informações sobre a análise química das rochas do solo lunar decorrente da missão Apollo.

Em seguida, vários sistemas foram desenvolvidos, tais como: RENDEZVOUS [20], que é baseado em diálogos e; INTELLECT [28], que é uma opção comercial para traduzir a linguagem natural em SQL. Já o uso de métodos de *machine learning* surgiu nos anos 90 com o estudo de [59] representando uma inovação na área.

O estudo [5] apresenta as principais vantagens e desvantagens dos sistemas NLI. Como vantagens são citados os seguintes fatos: o usuário não precisa aprender uma linguagem artificial; as questões que envolvem negação ou quantificação são mais fáceis de expressar usando linguagem natural; as consultas usam expressões anafóricas e elípticas e; elas permitem o uso de perguntas nas quais o significado é complementado pelo contexto do discurso. Como desvantagens são mencionadas as questões a seguir: é difícil para o usuário compreender o que não pode ser feito; quando o sistema não entende a pergunta não é claro para o usuário onde a falha ocorreu; os usuários esperam uma inteligência do sistema que não existe; devido a ambiguidade da linguagem natural esse não é o meio mais adequado para comunicação humano-computador; necessitam de configurações complexas para serem utilizados.

Segundo [35] sistemas NLI para banco de dados podem ser classificados quanto ao domínio da seguinte forma:

- **Sistemas Independentes de Domínio:** são sistemas que não usam informações sobre um banco de dados particular para construção da consulta formal como acontece no PRECISE [42]. Eles usam somente informações linguísticas e traduzem uma consulta em linguagem natural para uma consulta em linguagem formal. Portanto, eles não conseguem encontrar erros conceituais em banco de dados reduzindo a eficiência e a taxa de sucesso do sistema.
- **Sistemas Dependentes de Domínio:** são sistemas desenvolvidos para um domínio de banco de dados particular como acontece no LUNAR [58], que utiliza uma

estrutura de dados sobre as rochas da lua. Esses sistemas necessitam conhecer detalhes internos do banco de dados como o nome e número de atributos, número de tabelas, chave primária, entre outros. Como esses sistemas conhecem o esquema do banco de dados eles fornecem alta eficiência e desempenho.

2.2 Processamento de Linguagem Natural

A linguagem natural é aquela utilizada pelas pessoas no dia a dia - na forma escrita e falada - para se comunicar e envolve palavras e sons. O Processamento de Linguagem Natural (PLN) é uma sub-área da Inteligência Artificial (IA) que apresenta técnicas para analisar computacionalmente a linguagem humana.

O PLN pode realizar os seguintes tipos de análise linguística:

- **Análise morfológica:** a análise morfológica trata a estrutura, forma, flexão e classificação das palavras. Um dos objetivos dessa análise é a identificação dos morfemas, que é a menor unidade linguística que possui algum significado.
- **Análise léxica:** a análise léxica, conforme [6], é o processo de conversão de uma sequência de caracteres em uma sequência de palavras e possui o objetivo de tratar dígitos, hifens, marcas de pontuação e a caixa das letras.
- **Análise sintática:** a análise sintática avalia as possíveis formas de combinar as regras gramaticais. Para isso, mostra a representação da estrutura gramatical das palavras na sentença e os relacionamentos de dependência entre as palavras, ou seja, analisa as relações entre as palavras de uma sentença.
- **Análise semântica:** a análise semântica analisa o sentido das estruturas das palavras usando técnicas de desambiguação semântica para identificar o significado correto de cada palavra considerando o restante da sentença.

2.2.1 Identificação da Linguagem

A identificação da linguagem é responsável por analisar os termos da consulta inicial descrita pelo usuário e identificar o idioma utilizado. Segundo [32], a identificação da linguagem é um caso especial de classificação de texto que o mapeia para um conjunto pré-determinado de classes representando os idiomas.

Um dos métodos utilizados para identificação da linguagem é denominado *short strings* que é um desafio para as técnicas existentes [32] devido a pequena quantidade de termos para serem analisados. Em [55], são realizados testes em textos curtos contendo entre 5 e 21 caracteres, utilizando modelos com *n-grams*, sendo adequado para motores de busca que aceitam poucos termos na descrição da consulta.

Uma das principais aplicações de identificação da linguagem é o uso combinado com a tradução, em que a identificação automática de linguagem é usada como uma etapa de pré-processamento para determinar qual modelo de tradução será aplicado [32]. Nos sistemas de busca multilíngues, em que a consulta pode ser descrita em vários idiomas, a identificação da linguagem utilizada na consulta auxilia no processo de pesquisa, filtrando as fontes de dados que serão consultadas ou restringindo os idiomas que podem ser utilizados para descrever a consulta.

2.2.2 Segmentação

A segmentação de palavras ou *tokenization* é responsável pela análise léxica e possui significados distintos de acordo com o contexto utilizado. Nesse trabalho é adotado o conceito que aborda a *tokenization* como o processo de reconhecer e segmentar as palavras de uma sentença. Segundo [26], esse processo pode ser simples em alguns idiomas como o inglês, no qual é possível separar as palavras por um espaço em branco. Mas, pode ser complexo em idiomas que não possuem limitações explícitas das palavras como acontece com o Chinês. No processo de segmentação, os símbolos que separam as palavras são chamados de delimitadores [6].

Na segmentação das palavras, muitos algoritmos realizam a normalização da sentença padronizando os *tokens* em letras minúsculas, removendo caracteres de pontuação, entre outros. Quando a segmentação é realizada palavra por palavra, os *tokens* são chamados de *unigrams*, quando são extraídas duas palavras por vez são chamados de *bigram* e quando o número de palavras for um número natural n são nomeados de *n-grams* [25].

Segundo [14], geralmente uma simples segmentação das palavras (*tokenization*) funciona igual ou melhor que técnicas de pré-processamento mais complexas como extração do radical (*lemmatization*) nos processos de consultas, com exceção dos dados de domínio específico em que somente a segmentação das palavras gera um desempenho ruim.

2.2.3 *Stopwords*

A eliminação de *stopwords* é o processo de remoção das palavras que não possuem significado em linguagem natural ou não contribuem semanticamente para o significado de outras palavras e por isso podem ser eliminadas sem causar prejuízo da compreensão da sentença. A técnica de *stopwords* é também chamada de dicionário invertido, já que utiliza um dicionário de palavras a serem removidas da sentença.

Segundo [21], o nome *stopwords* se deve ao fato de que, quando as referidas palavras são identificadas elas são eliminadas, aumentando a eficiência e a eficácia da recuperação. As *stopwords* são palavras muito comuns, e devido a isso, não são adequadas

em sistemas de busca, pois geralmente possuem correspondências que não impactam nos resultados das consultas. Assim, as *stopwords* são definidas em [47] como um conjunto de palavras sem relevância para a recuperação da informação e a sua eliminação pode reduzir significativamente o tamanho da estrutura do índice e aumentar a acurácia dos resultados.

Conforme [6], artigos, preposições, conjunções e alguns verbos, advérbios e adjetivos são candidatos naturais para uma lista de remoção de *stopwords*. No entanto, essa eliminação nem sempre é vantajosa porque pode reduzir a revocação (quantidade de resultados que são recuperados). Exemplo: a sentença 'Quando você foi embora?' pode ser reduzida para somente a palavra 'embora' após o processo de eliminação de *stopwords*.

2.2.4 Extração do Radical

Na maioria das linguagens, principalmente as advindas de línguas morfologicamente ricas como o Português do Brasil, uma palavra pode ter variações de acordo com o tempo verbal, flexão de número, gênero e grau. Em sistemas de busca, essas flexões da palavra podem interferir nos resultados necessitando de técnicas de extração do radical como o *stemming* e *lemmatization* para reduzir as palavras a sua forma inflexionada.

O *stemming* é o processo de reduzir as palavras relacionadas para seu *stem*, que é sua forma base ou raíz, por meio da remoção do afixo [47]. Segundo [26], o *stemming* identifica os radicais em comum das variações das palavras sejam elas morfológicas ou sintáticas. Em sistemas multilíngues o *stemming* pode ser realizado antes ou após a tradução da consulta. O *stemming* reduz a flexão das palavras transformando-as em sua forma raíz utilizando heurísticas para remover afixos (prefixo e/ou sufixo) resultando em uma palavra que pode não constar no dicionário linguístico, por exemplo as flexões *study* e *studing* podem ser reduzidas para o termo *stud*.

O principal algoritmo de *stemming* para a língua inglesa é o *Porter Stemmer* [43], que é baseado em algumas regras que são utilizadas para remover o sufixo das palavras, tais como: gerúndios e plurais. Nesses casos, há a substituição das terminações flexionadas. O uso de *stems* melhora a performance da recuperação uma vez que reduz as variações de uma mesma palavra raiz para um conceito comum [6].

Para o Português brasileiro, foi proposto o algoritmo STEMBR [4] que realiza a extração dos radicais das palavras utilizando as regras gramaticais dessa língua. Outra forma muito comum de realizar o *stemming* é utilizando expressões regulares para remover os afixos das palavras. Ademais, conforme [6], o *stemming* permite a recuperação de informações contendo variações sintáticas dos termos da consulta.

Em muitas situações a extração do radical das palavras realizada no processo de *stemming* não é suficiente, pois algumas palavras extraídas não possuem significado.

Nestes casos, a técnica *lemmatization* pode obter melhores resultados ao utilizar um dicionário para extrair a forma base comum da palavra, ou seja, a parte em comum entre todas as variações morfológicas de uma palavra e obter como resultado uma palavra do dicionário linguístico.

Como mencionado, a *lemmatization* pode trazer resultados melhores, porém com um custo maior de processamento e desempenho. Geralmente é utilizado um dicionário de sinônimos ou tesauro, como o WordNet [23], para encontrar a parte comum entre os sinônimos de uma palavra e remover as partes que variam. Segundo [14], o uso de técnicas mais complexas de pré-processamento como *lemmatization*, geralmente, não ajuda nas arquiteturas de rede neural que são capazes de superar a esparsividade pela generalização da palavra, exceto para um conjunto de dados de domínio específico.

Os algoritmos que realizam *stemming* e *lemmatization* são diferentes para cada idioma porque cada idioma possui regras gramaticais diferentes. Um bom algoritmo necessita de um dicionário de palavras e em linguagens aglutinantes como o alemão, finlandês e árabe o processo de *stemming* é ainda mais difícil, bem como no idioma espanhol que possui muitas exceções gramaticais [6].

2.2.5 Expansão

A maioria das buscas realizadas na Internet utilizam poucos termos, ou seja, uma sentença curta que pode não produzir resultados relevantes para o usuário. Isso porque a consulta é muito curta para capturar adequadamente o que usuário pretende buscar. Além disso, o termo inserido pelo usuário pode não conter equivalência na base de dados, mas conter equivalência com um termo similar. Com isso, muitas vezes os usuários precisam reformular suas consultas para obter os resultados que lhes interessam [6]. Diante desse cenário, a técnica de expansão da consulta tem como finalidade mitigar esse problema.

A expansão da consulta (*Query Expansion - QE*) foi inicialmente aplicada em 1960 por [38]. A principal desvantagem dessa técnica é o custo computacional associado a sua aplicação. No caso de pesquisas na Internet - nas quais o tempo de resposta rápido é obrigatório - o custo computacional associado impossibilita seu uso parcial ou total [31]. Porém, com o aperfeiçoamento das técnicas e a melhora do poder computacional, os resultados estão cada vez mais satisfatórios.

Um dicionário é utilizado para realizar a expansão da consulta, o qual inclui informações sobre sinônimos e palavras relacionadas aos termos da consulta. De acordo com [6], palavras relacionadas são geralmente derivadas de um relacionamento de sinonímia. O principal tesauro utilizado para expansão da consulta é a *WordNet* [23], que é um

banco de dados léxico criado na Universidade de *Princeton*[1] que agrupa as palavras em conjuntos denominados de *synsets*. Os *synsets* possuem o *lemma* da palavra, a definição da palavra e exemplos desta palavra em sentenças, atuando como um dicionário.

Em sistemas multilíngues, de acordo com [39], a expansão da consulta pode ser realizada antes ou depois da tradução, sendo que a expansão da consulta após a tradução pode compensar consultas pobres. Foi demonstrado que a expansão da consulta antes da tradução gera melhores resultados em comparação com a tradução realizada após a expansão da consulta, embora a expansão da consulta gere resultados superiores em comparação com não usá-la [7, 39].

Para que o termo expandido não seja considerado com o mesmo grau de relevância que o termo original da consulta, são atribuídos pesos e diferentes técnicas de ranqueamento para os termos expandidos. Em [15], pesos diferentes são atribuídos para as associações um-para-um e um-para-muitos. Na associação um-para-um o termo candidato será adicionado a consulta expandida se estiver correlacionado a um termo da consulta original e na associação um-para-muitos um termo candidato será adicionado a consulta expandida se estiver correlacionado a vários termos da consulta original.

O principal problema da associação individual é que ela pode não demonstrar adequadamente a conectividade entre o termo de expansão e a consulta como um todo. Por exemplo a palavra *'technology'* pode ser expandida para *'information'*, mas se a consulta for *'music technology'* o termo expandido não estará corretamente relacionado. Por isso, a associação múltipla é recomendada para que o contexto em que a palavra está inserido seja considerado na expansão da consulta.

De acordo com os estudos da área, o número de termos ideal a serem expandidos para aprimorar os resultados da consulta sofrem uma grande variação, podendo variar de 1-3 [48] até 350-530 [13]. O uso de expansão da consulta aumenta a taxa de *recall* e *precison* [29]. Alguns estudos recentes mostram que a expansão da consulta melhora a *precison* pela desambiguação da consulta do usuário [61]. A adição desses termos de expansão melhora a efetividade da recuperação entre 7% e 25% [13]. Ademais, segundo [6], o uso da expansão da consulta utilizando um tesauro de similaridade levou a melhorias na qualidade da recuperação (em torno de 20%) em três coleções de dados diferentes.

2.2.6 Tradução

Em contraste com os sistemas monolíngues (*Monolingual Information Retrieval - MIR*), existem diferentes métodos para recuperar informações multilíngues. Entre eles há o *Cross-Language Information Retrieval (CLIR)*, no qual as consultas são realizadas

[1] https://wordnet.princeton.edu/

em um determinado idioma para obter dados em qualquer idioma diferente do idioma de origem [41]. Por meio dele é possível realizar a tradução da consulta inicial para obter resultados em outro idioma conforme foi utilizado nas propostas de [18], [45], [51], [53] e [54]. Conforme [37], o CLIR trata as consultas em um idioma e a recuperação de resultados em outro idioma.

Outro método que pode ser utilizado na recuperação de informações multilíngues é o *Multi-Language Information Retrieval (MLIR)*, que possibilita ao usuário submeter uma consulta ou obter resultados em vários idiomas por meio da tradução. Esse método é abordado em [44], [50] e [57]. De acordo com [37], o MLIR trata consultas em um ou mais idiomas e a recuperação de resultados também em um ou mais idiomas.

A principal diferença entre o CLIR e o MLIR é que no CLIR há restrição de um único idioma para consulta inicial ou obtenção de resultados. De tal modo que o idioma da consulta deve ser diferente do idioma dos resultados. Assim, a consulta é realizada em uma linguagem (por exemplo, Inglês) e os resultados são obtidos em outras linguagens (por exemplo, Português e Espanhol). Já no MLIR a consulta pode ser realizada em vários idiomas e os resultados também podem ser obtidos em vários idiomas inclusive no idioma utilizado na consulta inicial. Além disso, pode ser realizada a tradução dos resultados para o idioma da consulta inicial.

Um dos primeiros sistemas MLIR foi implementado em 1969 por Gerard Salton que recuperava documentos nos idiomas inglês e alemão [49]. O principal responsável pelos avanços em MLIR é o *Cross-Language Education and Function* (CLEF)[2]. Apesar dos recentes avanços, segundo [24], a efetividade da tradução da consulta *cross-lingual* é menor que 60% comparada com a recuperação monolíngue em termos de precisão média.

Esses métodos podem realizar a consulta multilíngue por meio de diferentes tipos de tradução, tais como: da consulta inicial, do resultado da consulta e de ambos. O método de tradução da consulta inicial tem sido o mais utilizado devido a sua simplicidade e, conforme [45], tem atraído muita atenção devido ao seu desempenho.

No caso da tradução da consulta inicial, podem ser utilizados três meios principais:

- **Dicionário:** uso de um dicionário bilíngue para traduzir de uma linguagem L1 para a linguagem L2;
- **Tradução Automática:** uso de um software para traduzir o resultado ou a consulta de uma linguagem L1 para outras linguagens, podendo utilizar um conjunto de dicionários em vários idiomas;
- **Corpora Paralelo:** uso de um corpo de documentos em duas ou mais línguas como referência para realizar novas traduções.

[2] http://www.clef-campaign.org

Geralmente, os usuários não possuem conhecimento de uma variedade de idiomas. Por isso, os sistemas que recuperam resultados multilíngues realizam a tradução automática da consulta inicial. Para traduzir a consulta inicial a abordagem mais simples consiste em usar um sistema de Tradução Automática (TA). O sistema de TA mais comum é o *Google Translator*[3], que aceita mais de 100 idiomas e adota o modelo de tradução *phrase-based* com *Neural Machine Translation* (NMT) para aumentar a fluência e acurácia. Outra forma de realizar a tradução é utilizando um tesauro multilíngue como o *Open Multilingual Wordnet* (OMW) [12], que possui suporte a mais de 57 linguagens.

No processo de tradução dos termos da consulta um problema recorrente é a ambiguidade, haja vista que em idiomas morfologicamente ricos como o Português brasileiro ou Inglês uma mesma palavra pode ter distintos significados e necessita do contexto para identificar o seu real significado. Por exemplo, a palavra 'banco' no idioma Português brasileiro pode se referir a uma instituição financeira ou a um objeto para sentar. Já a palavra correspondente na língua inglesa *bank* pode se referir a instituição financeira ou a margem de um rio. De acordo com [26], a ambiguidade é reconhecida como um dos mais importantes fatores que influenciam a efetividade do CLIR.

Nem todas palavras das consultas podem ser traduzidas na linguagem de destino. Uma questão importante para a tradução da consulta é o manuseio de palavras fora do vocabulário (*out of vocabulary - OOV*). Palavras como nomes próprios ou emprestadas de uma língua estrangeira se enquadram nessa categoria. Essas palavras são diretamente transliteradas para a linguagem de consulta formal.

A transliteração é uma técnica utilizada para resolver o problema de OOV que pode ocorrer no processo de tradução das consultas. No CLIR, a transliteração pode ser realizada por dois métodos [17]:

- **Pivô:** antes de converter as palavras de um idioma de origem para o idioma de destino as palavras são convertidas no símbolo de pronúncia (alfabeto fonético internacional).
- **Direto:** é baseado em *corpus* onde um estado intermediário não é necessário.

2.3 Considerações Finais

As abordagens para consultar bancos de dados relacionais baseada em palavras-chave e linguagem natural são complementares, possibilitando uma maior flexibilidade no formato de descrição da consulta inicial e exigindo menor conhecimento prévio do usuário para obter as informações pretendidas. As técnicas de pré-processamento da consulta

[3] https://translate.google.com

possibilitam a identificação dos termos relevantes para a consulta, o filtro dos termos que não contribuem para a obtenção dos resultados e a expansão e tradução dos termos relevantes para enriquecer a consulta.

As diversas técnicas de processamento de linguagem natural apresentadas tem evoluído constantemente. Elas já são capazes de produzir resultados promissores para pré-processamento da consulta inicial em sistemas de busca nos bancos de dados relacionais com metadados descritos em múltiplos idiomas, mas como as técnicas de pré-processamento apresentadas são dependentes do idioma, para um cenário multilíngue deve ser analisado previamente se um determinado idioma possui suporte nas técnicas utilizadas.

CAPÍTULO 3

Trabalhos Relacionados

A consulta por palavras-chave ou linguagem natural em bancos de dados relacionais é um problema de pesquisa que atrai a atenção de vários grupos de pesquisa [8], [35], [40] e [46], porém ainda há poucos estudos sobre a consulta em bancos de dados relacionais descritos por metadados multilíngues. Para verificar a existência de estudos relacionados a consultas multilíngues em bancos de dados relacionais foi realizada uma revisão sistemática [56] sintetizada a seguir.

3.1 Metodologia

Para a seleção dos estudos relacionados à temática proposta nessa investigação, foi utilizada a seguinte expressão genérica de busca (contendo palavras chave e seus sinônimos) aplicada sobre o título, resumo e palavras-chave: *(('natural language' OR keyword) AND (query OR search OR 'information retrieval') AND (multilingual OR multilanguage OR cross-language OR cross-lingual OR bilingual OR translingual) AND ('relational database' or database))*. Os resultados obtidos foram filtrados para o período compreendido entre os anos de 2015 a 2020, nas seguintes bases de dados: Portal de Periódicos da Capes[1], ACM Digital Library[2], IEEEXplore[3], Science Direct[4], Scopus[5],Spring Link[6], DBLP[7], Google Scholar[8] e Catálogo de Teses da Capes[9].

O período foi selecionado para contemplar os estudos mais recentes devido à constante evolução das técnicas e métodos para realizar consultas em bancos de dados relacionais e de processamento de linguagem natural. O termo *database* foi utilizado para

[1] https://www.periodicos.capes.gov.br
[2] https://dl.acm.org
[3] https://ieeexplore.ieee.org
[4] https://www.sciencedirect.com
[5] https://scopus.com
[6] https://link.springer.com
[7] https://dblp.org/
[8] https://scholar.google.com
[9] https://catalogodeteses.capes.gov.br

aumentar a quantidade de resultados da *string* de busca devido à pequena quantidade de estudos especializados em consultas em bancos de dados relacionais descritos por metadados em vários idiomas.

Ademais, para essa seleção foi definida uma estratégia para determinar os critérios de pesquisa que melhor se encaixam nos objetivos deste trabalho. Assim, foram definidos os seguintes critérios de seleção para classificar os trabalhos analisando o título, resumo e palavras-chave:

- **Critérios de Inclusão:** Os artigos considerados na pesquisa foram publicações na língua inglesa que atendiam ao menos um dos seguintes critérios: 1. Referência bibliográfica de pesquisa obtida pela *string* de busca; 2. Indicação do orientador ou grupo de pesquisa.
- **Critérios de Exclusão:** Os critérios utilizados para excluir trabalhos resultantes da *string* de busca foram: 1. Não publicado; 2. Publicado como pôster ou resumo; 3. Sem acesso público ou institucional ao inteiro teor do documento.

Foi elaborado um formulário para realizar a extração de dados dos artigos selecionados pelos critérios de inclusão e exclusão contendo as seguintes perguntas:

1. Quais técnicas de pré-processamento foram usadas?
2. Quais técnicas de tradução foram usadas?
3. Possui independência de domínio?
4. Possui independência de banco de dados?
5. Necessita de seleção do idioma?
6. Permite apenas instruções de consulta?
7. Retorna resultados em vários idiomas?
8. Permite consultar em mais de um idioma?
9. Possui recurso de autocompletar a consulta?
10. Usa banco de dados relacional?

3.2 Condução

busca genérica foi adaptada para cada uma das bases. Os resultados obtidos foram exportados para o formato .bibtex e importados na ferramenta parsifal[10] para gerenciamento da revisão sistemática. Foram importados 833 estudos distribuídos por base de dados científica conforme apresentado na Figura 3.1.

A seleção dos estudos foi realizada através da leitura do título, resumo e palavras-chave e cada documento, observando os critérios de inclusão e exclusão. Nessa etapa

[10]https://parsif.al

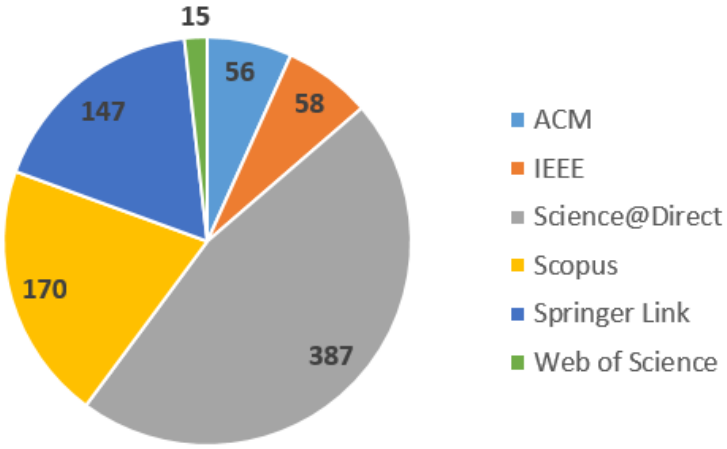


Figura 3.1: Estudos importados [56]

foram identificados 31 trabalhos duplicados que resultaram de bases científicas diferentes e 792 foram rejeitados, restando 10 trabalhos aceitos conforme os critérios definidos.

A extração dos dados foi realizada com a leitura completa dos trabalhos aceitos. Após a leitura, os trabalhos foram classificados de acordo com as questões definidas no formulário de extração de dados. O resultado dessa classificação foi compilado e é apresentado na Tabela 3.1.

3.3 Análise dos Resultados

No trabalho descrito em [19] apresenta-se uma interface bilíngue em linguagem natural para banco de dados que possibilita ao usuário final se comunicar com o sistema em sua linguagem nativa. Neste trabalho o usuário pode criar uma consulta nos idiomas *Hindi* e *Punjabi* e receber a resposta no mesmo idioma utilizado na consulta, sendo que a transformação de linguagem natural para SQL é realizada pelo *Karak Solver*.

Já no estudo de [44], desenvolveu-se o protótipo de uma interface em multi-linguagem natural para fontes de dados estruturados. Nela a conversão da consulta do usuário em linguagem natural utiliza a distância de *Levenshtein* e representa a definição formal de um sistema de diálogo como um processo de decisão de *Markov*. Nesse estudo os métodos para manipulação de linguagem natural são aplicados na tradução automática da consulta, que é realizada no idioma Russo.

A proposta apresentada em [51], utiliza a abordagem de *machine learning* em um sistema Web bilíngue de perguntas e respostas nos idiomas *Hindi* e Inglês. Esse sistema espera uma resposta única e direta em vez de uma lista de resultados. Nessa proposta a consulta é executada em um banco de dados de conhecimento e tem sua similaridade computada.

Tabela 3.1: Extração de dados das publicações aceitas [56]

Referência	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
[18]	-	Google Translate	✓	✓		✓	✓			
[19]	Token, POS e Chunking	-			✓	✓		✓	✓	✓
[37]	-	-				✓				✓
[44]	-	-				✓				✓
[45]	-	Google Translate				✓				
[50]	-	Microsoft Translate				✓	✓			✓
[51]	Stopwords, NER, Lemma e Token	-	✓		✓	✓		✓		
[53]	Lemma, Stopwords, POS e Token	Dicionário bilíngue				✓				
[54]	-	-				✓	✓			✓
[57]	Léxica e sintática	-	✓		✓	✓		✓	✓	✓

Uma nova abordagem de *Cross-Lingual Information Retrieval* (CLIR) para *English-Persian* é apresentada em [45], onde a API do *Google Translate* foi adaptada para traduzir as consultas, sendo utilizadas 50 consultas do conjunto de dados TREC para avaliar esse sistema. Já em [54] é apresentada uma interface em linguagem natural para banco de dados que realiza a conversão da língua *Gujarati* para o Inglês.

Em [57] é descrito um sistema Web avançado bilíngue e com interface independente do domínio para banco de dados usando uma abordagem com *machine learning*. Esse sistema aceita os idiomas de consulta *Punjabi* e *Hindi* e possui a função de autocompletar a consulta do usuário.

As áreas de recuperação da informação em vários idiomas mais importantes são introduzidas em [37], entre elas CLIR e MLIR. Em [53], utiliza-se ontologias multilíngues para reduzir a ambiguidade e os problemas de desempenho de sistemas CLIR. Um módulo de *word sense disambiguation* é utilizado para resolver a ambiguidade na língua *Tamil*.

Uma técnica de expansão da consulta é empregada em [18] para melhorar a efetividade dos resultados do idioma Hindi para o Inglês, no sistema CLIR. Nesse estudo, os resultados são ranqueados usando o método *Okapi-BM25*. Já no estudo discutido em [50], uma abordagem baseada em metadados semânticos é utilizada e a tradução da

consulta do usuário entre a língua nativa e a de destino é realizada utilizando o *Microsoft Translate*.

Com base na análise dos estudos mencionados, podemos fazer as ponderações que serão descritas a seguir. Para realizar o pré-processamento da consulta inicial podem ser utilizadas várias técnicas de processamento de linguagem natural. As principais técnicas utilizadas nos trabalhos identificados foram: *lemmatization* [41, 45], *tokenization* [41, 44, 45, 50], *stopwords* [41, 45] e *PoS Tagging* [41, 44]. Na tradução da consulta inicial foram utilizados os recursos do *Google Translate* [23, 54] e *Microsoft Translate* [37].

A tradução da consulta inicial pode ser realizada antes ou após o pré-processamento. Nos trabalhos [37, 44, 54, 57] a tradução foi realizada antes e nos trabalhos [23, 53] foi realizada após. Dentre os trabalhos que realizam a tradução antes do pré-processamento, a maioria não utiliza técnicas de PLN, mas foram classificados desse modo por traduzirem a consulta original.

A expansão da consulta foi aplicada no trabalho descrito em [23], que realizou 50 (cinquenta) consultas no idioma *Hindi* e as traduziu para o inglês utilizando o *Google Translate*. Os termos em inglês obtidos foram expandidos para produzir resultados mais relevantes. Essa técnica também foi utilizada no trabalho apresentado [41] com apoio da *WordNet* que também foi utilizada em [37, 44].

No trabalho descrito em [45] foi projetada uma arquitetura bilíngue na qual o usuário realiza a escolha do idioma da consulta que seria pré-processada, gerando *tokens* e removendo *stopwords*. Um sistema bilíngue que aceita as línguas *Hindi* e *Punjabi* na consulta é apresentado em [44]. Esse sistema armazena as consultas executadas com sucesso para sugerir novas consultas automaticamente.

Uma arquitetura que identifica o idioma da consulta, extrai os *tokens*, a classe gramatical, realiza análise morfológica, remove as *stopwords* e identifica o domínio da consulta é proposto por [57]. Essa arquitetura permite consultar bancos de dados com metadados em inglês utilizando a distância de *Levenshtein* e o processo de decisão de *Markov*.

Para avaliar a tradução automática, em [54] foi proposto o uso da métrica BLEU para medir a precisão da tradução em comparação com a tradução feita pelo ser humano. O estudo [37] propõe um *framework* que aceita palavras-chave do usuário e traduz no idioma de destino utilizando o *Microsoft Translator*. Esse *framework* recebe a consulta no idioma nativo do usuário sem que o mesmo conheça o idioma em que o banco de dados foi implementado.

3.4 Considerações Finais

Há poucos estudos relacionados à consulta em bancos de dados relacionais descritos por metadados multilíngues, mas esse tipo de consulta tem um grande potencial de aumentar os resultados pretendidos pelo usuário ao consultar uma maior quantidade e diversidade de fontes de dados.

As técnicas de pré-processamento utilizadas são diversificadas e não há comprovação de qual técnica é mais eficiente para a consulta em bancos de dados relacionais. A tradução da consulta nos estudos selecionados é realizada usando ferramentas de tradução automática e há propostas em que a consulta é traduzida, mas também há propostas em que os resultados são traduzidos para o idioma da consulta inicial.

CAPÍTULO 4

Solução Proposta

Neste trabalho, a solução proposta para realizar consultas em banco de dados relacionais descritos por metadados multilíngues é baseada na arquitetura do sistema SQUIRREL [46], apresentada na Figura 4.1. Essa arquitetura realiza consultas por palavras-chave em bancos de dados relacionais descritos por metadados no mesmo idioma da consulta inicial, ou seja, monolíngues.

Para possibilitar a consulta em metadados descritos em vários idiomas, a etapa de pré-processamento da consulta dessa arquitetura foi aprimorada. Isso para possibilitar o uso de técnicas de linguagem natural e a tradução automática da consulta inicial para os idiomas disponíveis nas descrições das fontes de dados.

4.1 Consultas em Bancos de Dados Relacionais

Para mapear palavras-chave em linguagem de consulta estruturada (SQL), a proposta em [46] apresenta a arquitetura do sistema SQUIRREL que realiza o mapeamento semântico da consulta inicial. Os principais aspectos dessa arquitetura consiste na: identificação e seleção do banco de dados exposto na Web contendo informações úteis para a consulta; expansão da consulta adicionando palavras semanticamente relacionadas; análise de funções de ordenação e agregação; análise de palavras compostas.

SQUIRREL [46] identifica as palavras da consulta que não fornecem um significado direto na estrutura do banco de dados, mas outras formas semânticas como o uso de funções de agregação e ordenação. A consulta é segmentada em palavras que, posteriormente, são expandidas com exceção das palavras compostas que estiverem entre aspas simples.

Para a seleção do banco de dados - que contém informações relevantes para a semântica pretendida na consulta inicial - é utilizada a TME que representa uma extensão do catálogo do banco de dados. Os bancos de dados são ranqueados por relevância de acordo com o número de ocorrências das palavras da consulta com os atributos da TME.

No mapeamento da consulta são analisadas as potenciais associações entre os termos do banco de dados relevante e os termos da consulta inicial. As associações são

registradas numa matriz de pesos que consiste numa tabela bidimensional contendo uma linha para cada palavra-chave e uma coluna para cada termo do banco de dados. Os maiores pesos representam os melhores mapeamentos.

Inicialmente são geradas duas matrizes: uma para termos de esquema e outra para termos de valor. Os termos de valor são os termos que não foram mapeados como termos de esquema. Cada combinação dos termos de esquema e de valor gera uma configuração na qual a pontuação é a soma dos pesos de todos elementos. As melhores configurações geram as interpretações que são as consultas em linguagem SQL para serem executadas nos bancos de dados identificados como relevantes.

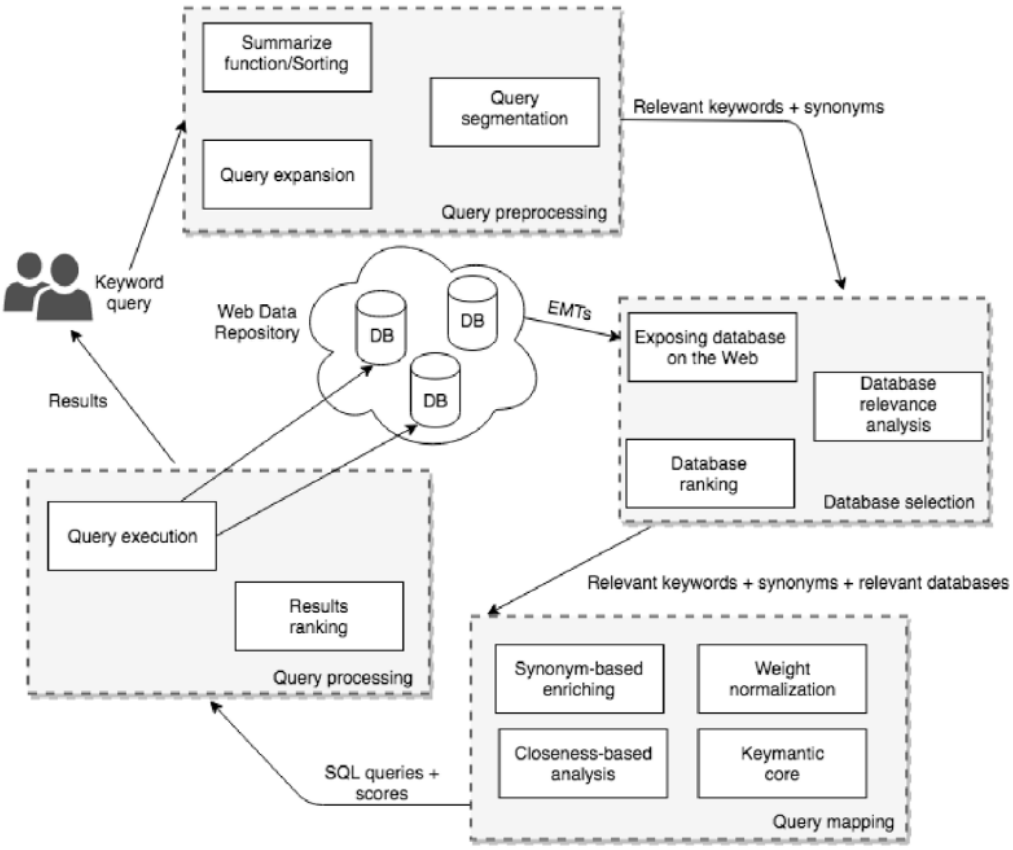


Figura 4.1: Arquitetura do SQUIRREL [46]

4.2 Pré-Processamento da Consulta

O pré-processamento da consulta é responsável por realizar a normalização, a expansão e a tradução dos termos da consulta inicial. As principais operações realizadas são:

- Padronizar todas as letras da consulta inicial para minúsculas;
- Remover caracteres de pontuação;

- Identificar e remover as palavras da consulta inicial que não fornecem um mapeamento direto em estruturas do banco de dados, como funções de agregação ou ordenação;
- Segmentar os termos da consulta;
- Eliminar as palavras muito frequentes que não possuem significado relevante para a sentença;
- Reduzir as flexões das palavras para sua forma base comum do dicionário;
- Expandir as palavras utilizando termos sinônimos para enriquecer a expressão de busca;
- Identificar os idiomas disponíveis na tabela TME e realizar a tradução da consulta inicial para esses idiomas.

Neste trabalho, a consulta inicial pode ser uma sentença em linguagem natural ou uma sequência de palavras-chave que representam a intenção do usuário. O idioma de entrada da consulta inicial é limitado à língua Inglesa, pelos motivos a seguir: os metadados dos bancos de dados utilizados em trabalhos relacionados são comumente criados nesse idioma; as principais técnicas de PLN utilizadas para pré-processamento são mais eficazes nesse idioma e; o foco do estudo é ampliar os resultados da consulta para outros idiomas que possuam informações relevantes através da tradução automática da consulta inicial.

Para a realização da consulta inicial não é esperado que o usuário tenha conhecimento sobre os seguintes quesitos: linguagens formais de consulta como SQL; estrutura dos dados ou metadados e; idiomas disponíveis a serem consultados. Pois, a solução propõe realizar todo esse processo de forma automática e transparente para o usuário. Assim, a consulta inicial poderá conter palavras com variações morfológicas e semânticas sem prejudicar os resultados que serão obtidos.

A Figura 4.2 apresenta as etapas de pré-processamento da consulta que são propostas neste trabalho. Podemos verificar que há uma ordem específica para a realização de cada etapa do pré-processamento. A justificativa para adoção dessa ordem é apresentada no detalhamento de cada etapa.

4.2.1 Segmentação da Consulta

A etapa de segmentação da consulta, também conhecida como *tokenization*, realiza a análise léxica dos termos da consulta e é responsável por separar cada termo da consulta inicial utilizando o caractere 'espaço em branco' como delimitador dos termos. A segmentação da consulta, também adotada nos trabalhos [53, 54, 57], é comumente utilizada no pré-processamento das consultas anteriormente a outras etapas.

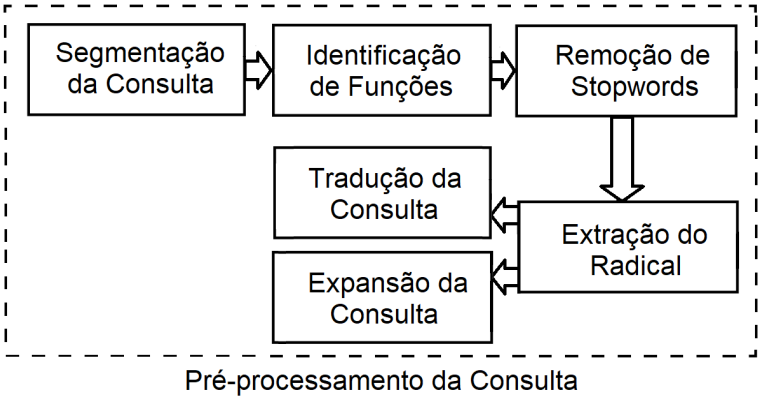


Figura 4.2: Solução proposta

Nessa etapa também é realizada a normalização léxica que padroniza todos os caracteres da consulta em minúsculo para aumentar a precisão das ocorrências com a lista de funções. Além disso, remove os caracteres de pontuação que não contribuem para obtenção dos resultados da consulta, pois se estiverem juntos aos termos da consulta podem reduzir as correspondências com as informações do banco de dados.

Os termos segmentados da consulta inicial são denominados de *tokens* e são armazenados numa estrutura de dados que permite preservar a ordem da consulta original. Essa segmentação é importante para possibilitar o tratamento de cada palavra individualmente, mas a ordem na consulta inicial e os seus relacionamentos devem ser preservados para manter a semântica da consulta.

Termos compostos que não podem ser tratados de forma separada na consulta devem ser informados entre aspas simples para que essa etapa não os segmente. Esses termos, geralmente, pertencem a classe gramatical dos substantivos (nome de pessoas, lugares, entre outros). Os termos compostos que possam indicar o uso de funções de agregação ou ordenação são mapeados na etapa de identificação de funções e, portanto, não podem ser segmentados.

Na proposta descrita nesse trabalho, a segmentação da consulta deve ser realizada antes das etapas de remoção de *stopwords*, extração do radical e tradução e expansão da consulta. Pois, essas técnicas analisam cada *token* individualmente ao invés da sentença completa da consulta.

Um exemplo da etapa de segmentação, pode ser observado a seguir: a consulta [*movie short 'based true story'*] seria segmentada nos *tokens* [*'movie', 'short', 'based true story'*], ou seja, os termos entre aspas simples seriam tratados como um único termo na consulta porque não são segmentados.

4.2.2 Identificação de Funções

Essa etapa é responsável por identificar as palavras da consulta inicial que podem representar uma agregação ou ordenação de resultados. Para realizar essa identificação é utilizada uma lista de palavras reservadas tendo como base o idioma da consulta inicial. Essa lista contêm palavras e seus sinônimos que sugerem o uso de funções de agregação ou ordenação. Ademais, são utilizadas como funções de agregação ou ordenação pelos sistemas gerenciadores de banco de dados. Neste trabalho, serão utilizados como base os sete grupos definidos por [46] que foram expandidos (termos acrescentados em destaque) e são listados a seguir:

- *maximum = {**higher**, highest, maximum, maximal, **larger**, largest, **greater**, greatest, **most** e max}*
- *minimum = {**lower**, lowest, minimal, minimum, **smaller**, smallest, least, **shorter**, shortest e min}*
- *mean = {average, mean e avg}*
- *sum = {total e sum}*
- *count = {quantity, **how many**, amount, **number** e count}*
- *grouping = {for each, **of each**, **aggregate by**, for all, **group** e grouped by}*
- *order = {order by, **ordered by**, descending order by, ascending order by, sorted by, ranked by, classified by, organized by}*

A fase de identificação de funções recebe na entrada a sentença da consulta inicial sem nenhum tratamento ou pré-processamento. Devido a esse fato, o dicionário de palavras reservadas deverá conter todas as possíveis variações morfológicas das palavras que sugerem o uso de funções de agregação ou ordenação. Vale ressaltar que a lista apresentada nesse estudo possui somente uma pequena amostragem dessas variações, portanto, é uma listagem parcial para exemplificar a solução proposta.

As palavras identificadas como funções de agregação ou ordenação são omitidas da consulta inicial nas etapas seguintes. Isso para que não sejam mapeadas para termos de esquema ou termos de valor do banco de dados. Após a identificação das palavras da consulta inicial que sugerem o uso de funções de agregação ou ordenação, é necessário identificar sobre quais termos da consulta essas funções serão aplicadas. Para tanto, a posição de cada palavra na consulta inicial deve ser preservada a fim de possibilitar a identificação da palavra imediatamente seguinte a função como sendo o atributo a ser agregado ou ordenado.

A identificação de funções de agregação ou ordenação deve ser realizada antes da remoção de *stopwords* para que não sejam removidas as palavras que indicam o uso de determinada função. Por exemplo, na consulta *'number of films in each genre'* as palavras

'in each' que poderiam ser mapeadas em uma função de agregação, seriam removidas caso a técnica de *stopwords* fosse aplicada antes da identificação de funções.

4.2.3 Remoção de *Stopwords*

Essa etapa é responsável pela remoção de termos muito frequentes na consulta inicial e também foi utilizada nos estudos [19, 51, 53]. A remoção desses termos, geralmente, não causa prejuízo para obtenção dos resultados e, conforme apresentado na Seção 2.2.3, melhora os resultados. Uma vez que reduz os termos a serem processados eliminando os que não possuem relevância para a consulta. Essa etapa basicamente reduz a consulta inicial em linguagem natural para palavras-chave utilizando uma lista de palavras como *'I', 'My', 'You', 'She', 'Do'* entre outras.

A remoção de *stopwords* é dependente do idioma utilizado na consulta inicial, pois utiliza uma lista de palavras reservadas que são comumente classificadas como *stopwords*. Portanto, não é usado um dicionário restrito ao domínio do banco de dados utilizado devido a diversidade de fontes de informação proposta na solução.

O uso de dicionário genérico, ou seja, de domínio amplo, pode remover palavras da consulta inicial afetando a precisão e a revocação. Como neste trabalho o domínio das fontes de dados disponíveis é desconhecido previamente, o uso de *stopwords* genéricas de uso comum é a forma mais viável para não restringir o domínio das fontes de dados a serem disponibilizadas.

A principal forma de visualizar as palavras mais frequentes em uma sentença é utilizando uma nuvem de palavras que destaca aquelas mais utilizadas com um tamanho maior que as palavras menos usadas. Para identificar visualmente as palavras mais frequentes dentre todas as consultas iniciais apresentadas na Tabela A.1 foi gerada a nuvem de palavras ilustrada na Figura 4.3.

Nessa nuvem de palavras podemos visualizar que o termo mais frequente é *'movies'*, por ser a principal palavra que representa o banco de dados IMDB utilizado como fonte de dados para as consultas. Mas, podemos perceber também uma frequência alta de palavras como *'the'*, *'and'*, *'of'*, *'in'*, *'with'*, *'at'* que podem ser interpretadas como *stopwords* e removidas da consulta.

Para exemplificar a etapa de remoção de *stopwords*, consideramos a consulta segmentada [*'number', 'of', 'movies', 'genres'*]. Após a remoção de *stopwords* essa consulta retornaria a expressão [*'number', 'movies', 'genres'*], pois o termo *'of'* foi identificado no dicionário de *stopwords* e removido da consulta.

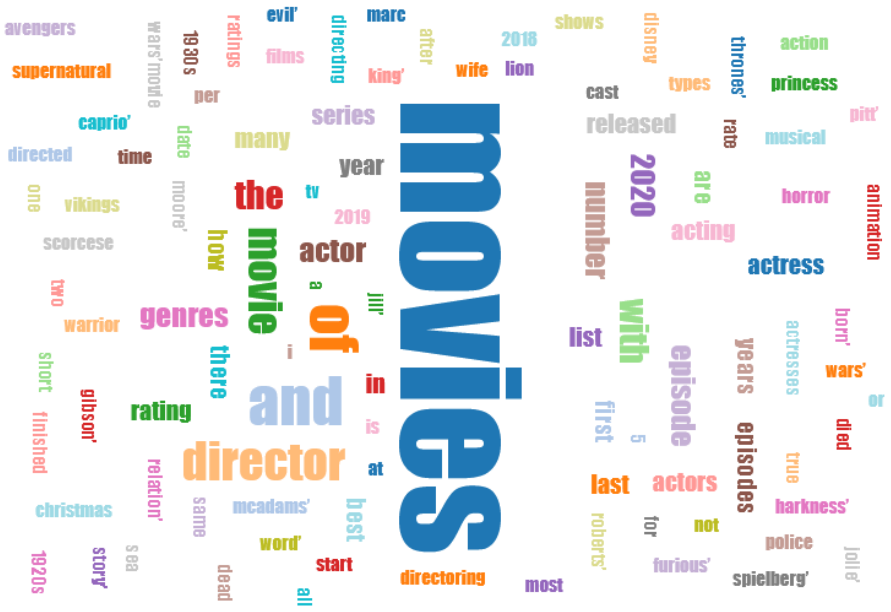


Figura 4.3: Nuvem de palavras das consultas iniciais

4.2.4 Extração do Radical

Um dos principais problemas para obtenção dos resultados esperados em sistemas de consulta é a quantidade de variações morfológicas que uma mesma palavra pode apresentar em um determinado idioma. Por exemplo, como acontece com as variações de plural que prejudicam a identificação de correspondências nas bases de dados consultadas. Haja vista que se na consulta for informado uma variante da palavra que não existe no banco de dados e o sistema utilizar uma busca exata ele não será capaz de consultar as informações.

Para mitigar esse problema, utiliza-se a técnica de remoção dos afixos denominada extração do radical, também aplicada nos trabalhos [19, 51, 53]. Essa técnica realiza a análise morfológica dos termos da consulta e remove prefixos ou sufixos das palavras reduzindo as variações morfológicas e transformando as palavras em uma forma de base comum.

O processo de extração do radical é dependente do idioma utilizado na consulta inicial, já que cada idioma possui regras diferentes para formação morfológica das palavras. Assim, cada idioma necessita de estratégias diferentes para extrair o radical das palavras considerando as regras gramaticais específicas.

Há duas formas principais de realizar a extração do radical: a *lemmatization* e a *stemmization*. A *stemmization* em determinadas situações é mais favorável, já que o *stem* geralmente é mais curto que o *lemma*. Conforme podemos observar na Figura 4.4, o *lemma* de *'changing'* é *'change'*, enquanto o *stem* será *'chang'*. Mas, esse 'truncamento' das palavras que a técnica *stemmization* gera é prejudicial para a etapa de tradução e

de expansão da consulta. Como podemos observar no exemplo já mencionado, o *stem 'chang'* não seria expandido ou traduzido para outros idiomas, pois não encontraria correspondências nos dicionários gerando resultados de busca incompletos.

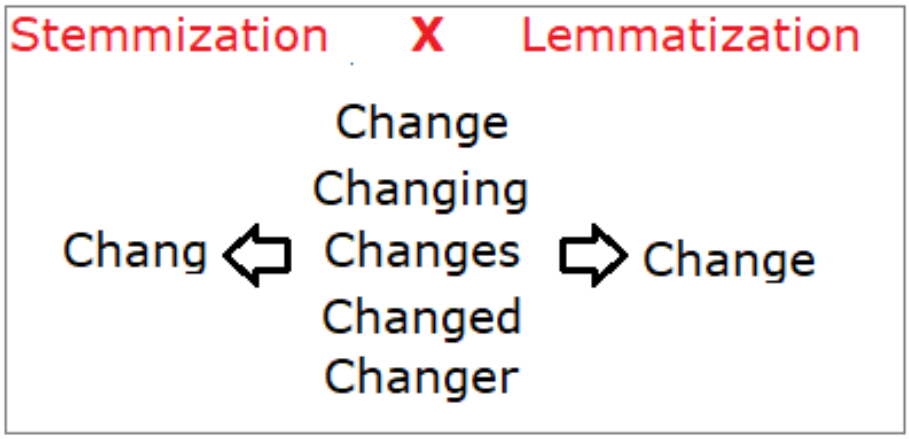


Figura 4.4: Técnicas de extração do radical

A técnica *lemmatization* reduz as variações das palavras para a sua forma base comum do dicionário. Isso é fundamental para que as etapas de tradução e de expansão da consulta recebam como entrada uma palavra compreensível computacionalmente e consiga alcançar os resultados esperados. Outra vantagem da *lemmatization* é a possibilidade de lidar com verbos em suas formas irregulares, ou seja, verbos que não seguem a regra geral de formação do passado e do particípio passado. Por exemplo, como acontece com o verbo 'começar' que tem como forma base no inglês *'begin'*, sua forma no passado *'began'* e sua forma no particípio passado *'begun'*. Nesse trabalho será utilizada a *lemmatization* para extração do radical.

A extração do radical deve ser realizada antes da tradução e da expansão da consulta porque elas utilizam um dicionário linguístico e com os termos reduzidos para sua forma base comum aumenta a possibilidade de correspondências com esse dicionário. Por outro lado, a extração do radical deve ser realizada após a remoção de *stopwords* para que termos considerados como *stopwords* não sejam pré-processados durante a extração radical.

Para compreender com mais clareza a etapa de extração do radical utilizando a técnica *lemmatization*, podemos exemplificar com a consulta pré-processada [*'movies', 'action'*] que após a extração do radical é reduzida para a consulta [*'movie', 'action'*]. Nesse exemplo, a flexão de plural da palavra *'movies'* foi removida para sua forma singular.

4.2.5 Tradução da Consulta

Essa etapa identifica os idiomas distintos disponíveis no atributo *dc_language* da tabela TME e traduz automaticamente os termos da consulta pré-processada para os idiomas disponíveis. Os termos entre aspas simples não devem ser traduzidos porque podem representar termos de valor. Esses termos de valor comumente representam nomes de pessoas, lugares e outros que tem preservada a sua grafia, independentemente, do idioma utilizado. Além disso, podem ser termos que não possuem tradução em outros idiomas.

Neste trabalho, optou-se por transliterar de forma direta os termos identificados entre aspas simples na consulta, ou seja, mantidos com a mesma escrita e grafia do idioma utilizado na consulta inicial. Essa transliteração é importante, já que esses termos geralmente não possuem tradução em outros idiomas.

Para realizar a tradução automática ou *Machine Translation* (MT) da consulta inicial podem ser utilizados dois modelos: o modelo baseado em palavras e o modelo baseado em frases. Ambos utilizam um corpora que é uma coleção de textos traduzidos em diversos idiomas para realizar o treinamento do modelo.

No modelo baseado em palavras a menor unidade que possui significado é a palavra. No entanto, uma única palavra pode conter várias correspondências em outro idioma. Esse modelo utiliza a tradução léxica que é mais adequada no contexto de consultas que possuem poucas palavras. Isso porque traduz as palavras isoladamente através de um dicionário que mapeia as palavras de uma linguagem para outra. Já no modelo baseado em frases, pequenas sequências de palavras são traduzidas simultaneamente alcançando um melhor desempenho [34], pois o melhor significado de cada palavra é aplicado considerando o seu relacionamento com as palavras anteriores e posteriores.

A principal diferença entre os dois modelos é que os sistemas de tradução automática baseados em frases realizam a tradução de todas palavras da consulta simultaneamente retornando a consulta em outro idioma. Enquanto que na tradução léxica baseada em palavras, a tradução é realizada palavra por palavra e cada uma delas pode conter um conjunto de outras palavras no idioma traduzido.

A tradução da consulta pode ser realizada antes, durante ou após a expansão da consulta, conforme podemos verificar nos trabalhos [45, 50, 54]. Se a tradução for realizada antes da expansão da consulta cada palavra poderá conter vários significados no idioma de destino e cada um desses significados (palavras) poderá ser expandido para diversas outras palavras. Com isso, aumentando significativamente a quantidade de termos na consulta dentre os quais muitos podem não ter relação direta com o termo original utilizado. Outro problema que pode ocorrer ao traduzir antes de realizar a expansão é que para expandir os termos traduzidos deverá ser consultado um dicionário de sinônimos para cada idioma identificado. De tal forma que haverá um aumento no número de operações

de pré-processamento e, consequentemente, no tempo da consulta.

No caso de a tradução da consulta inicial ser realizada após a expansão da consulta, serão traduzidos termos expandidos que podem não conter um significado no idioma de destino ou o significado não ser similar ao do termo original. Já quando a tradução acontece durante a expansão da consulta os problemas mencionados nas outras duas formas são minimizados, pois utilizam um dicionário bilíngue para realizar a tradução e pesquisar termos sinônimos. Vale ressaltar que a tradução da consulta inicial e a expansão da consulta dependem da disponibilidade de dicionários multilíngues, portanto são dependentes do idioma em que necessita ser traduzido.

Neste trabalho a tradução da consulta inicial é realizada durante a expansão da consulta, portanto essas etapas são realizadas em paralelo e recebem os mesmos termos pré-processados na fase de extração do radical. A tradução da consulta utiliza o modelo baseado em frases, de modo a preservar o relacionamento semântico entre as palavras da consulta inicial pré-processada.

Para exemplificar a etapa de tradução da consulta consideremos a consulta [*'director', 'movie', 'list'*] que está descrita no idioma Inglês. Essa consulta após ser traduzida para o idioma Português deverá retornar a expressão [*'diretor', 'filme', 'lista'*] como resultado.

4.2.6 Expansão da Consulta

Nas principais línguas, em sua forma falada ou escrita, uma mesma palavra pode conter vários significados e palavras diferentes (sinônimos) podem conter o mesmo significado. Essa diversidade possibilita que um termo da consulta inicial não tenha correspondência em uma determinada base de dados, mas possua um termo similar que represente o mesmo significado.

A etapa de expansão da consulta trata esse problema ao identificar termos sinônimos aos informados na consulta inicial e adicioná-los como termos adicionais. Nos trabalhos [16, 19, 57] a expansão da consulta também foi utilizada com esse propósito. O processo de expansão da consulta proporciona um enriquecimento da consulta, já que possibilita uma maior correspondência de palavras e um maior número de resultados que coincidam com a intenção do usuário.

Essa etapa também pode realizar a desambiguação dos termos da consulta inicial. A desambiguação analisa todos os termos da consulta inicial em conjunto para identificar o melhor significado de cada palavra em separado. Para realizar a expansão da consulta é utilizado o tesauro *Open Multilingual WordNet* (OMW) [12] que fornece sinônimos, hiperônimos, hipônimos, entre outros, em vários idiomas e agrupados numa estrutura

denominada como *synset*. Caso o *lemma* do termo a ser expandido não possua um *synset* na OMW disponível, a consulta prosseguirá sem expandir o termo em questão.

A OMW foi criada utilizando os dados do *Wiktionary* e do *Unicode Common Locale Data Repository*. Ela possui suporte para vários idiomas e contém mais de 2 milhões de sentidos de palavras. A OMW estende a estrutura original da *Princeton Wordnet of English* (PWN) 3.0 e todos os seus componentes são abertos para serem modificados e compartilhados por qualquer pessoa e para qualquer fim [12]. Como a OMW utiliza *synsets* para agrupar palavras que possuem o mesmo significado num mesmo idioma ou em diversos idiomas, a tradução é realizada palavra por palavra, podendo perder o relacionamento semântico entre as palavras da consulta inicial. Como alternativa a este método há várias ferramentas de tradução que aceitam o modelo baseado em frases como o *Google Translate* e *Microsoft Translate*.

Vale destacar que a expansão da consulta deve ser realizada após a extração do radical que retorna um *lemma* que é utilizado para identificar os termos sinônimos correspondentes ao idioma informado. A expansão da consulta é realizada durante a tradução da consulta, pois elas utilizam a mesma consulta pré-processada para realizar a tradução. Ademais, essa etapa aplica o modelo baseado em palavras para realizar a tradução que utiliza a OMW para expandir e traduzir simultaneamente os termos da consulta inicial em cada *synset*.

Na solução proposta, neste trabalho, os três primeiros termos dos dois primeiros *synsets* encontrados no idioma a ser traduzido são escolhidos como os termos da consulta expandida. Como um *synset* pode conter menos que três termos, a expansão da consulta pode adicionar até seis termos sinônimos do termo original.

Para exemplificar a etapa de expansão da consulta no contexto multilíngue apresentado na solução proposta, analisemos o seguinte modelo: o termo *'salary'* de uma consulta no idioma Inglês seria expandido para os termos ['salário', 'ordenado', 'retribuição', 'remuneração'] no idioma Português. Esses termos traduzidos seriam sinônimos do termo em Inglês consultado.

4.2.7 Exemplos

Para ilustrar as etapas de pré-processamento de uma consulta inicial utilizaremos como base as consultas de número 25 e 46, que são apresentadas na Tabela A.1. A consulta inicial e o pré-processamento de cada etapa realizado nessas consultas são descritos na Tabela 4.1. É importante enfatizar que nas consultas de exemplo todas as etapas de pré-processamento foram realizadas, mas que algumas consultas não precisam realizar todas etapas de pré-processamento para serem executadas.

Tabela 4.1: Exemplo de pré-processamento

Etapas	**Consulta 25**	**Consulta 46**
Consulta inicial	*How many smallville episodes are there*	*actors 'The Last Word'*
Segmentação da consulta	*how many, smallville, episodes, are, there*	*actors, 'the last word'*
Identificação de funções	*count*	*'the last word'*
Remoção de *stopwords*	*smallville, episodes*	*actors*
Extração do radical	*smallville, episode*	*actor*
Tradução da consulta	*smallville*, episódio	ator
Expansão da consulta	*smallville*, episódio	ator, atriz, artista

Cabe enfatizar, que na consulta inicial as palavras compostas devem ser informadas entre aspas simples para que sejam consideradas como uma única palavra. O uso de termos entre aspas simples sugere que o termo se refere a um atributo de valor e assim não deve ser traduzido ou expandido.

Na consulta 25, na etapa de segmentação da consulta, a palavra *'how many'* foi segmentada como um único *token* porque foi utilizada a técnica *Multiword Tokenizer* para não segmentar palavras compostas que sugerissem o uso de funções de agregação. Além disso, na etapa de identificação de funções houve o reconhecimento de que o termo *'how many'* era uma função de agregação que devia ser mapeada para a função *count*. Esse termo *'how many'* foi excluído da consulta e o termo seguinte, nesse caso *'smallville'*, mapeado como o termo a ser agregado. Já na etapa de remoção de *stopwords* os termos *['are', 'there']* foram removidos da consulta inicial por serem termos comuns em dicionários de *'stopwords'*.

Na etapa de extração do radical da consulta 25, o termo *'episodes'* foi modificado para remover sua flexão no plural e obter o termo *'episode'*. Para a etapa expansão da consulta chegaram os termos *'smallville'* e *'episode'*. Como *'smallville'* é um nome próprio e não foi informado entre aspas simples ele foi transliterado. Já o termo *'episode'* foi traduzido para 'episódio' no idioma Português, mas não foram encontrados termos sinônimos traduzidos para esse idioma. Portanto, a etapa de expansão da consulta retornou os mesmos termos da etapa de tradução da consulta.

Na consulta 46, o termo *'the last word'* foi informado entre aspas simples e identificado como termo de valor na etapa de identificação de funções, sendo removido das etapas seguintes de pré-processamento. Uma vez que ele foi mapeado como termo de valor e não como termo de esquema. Na etapa de extração do radical o termo *'actors'* teve sua flexão de plural removida, em seguida esse termo foi traduzido e expandido para os termos ['ator', 'atriz', 'artista'].

4.3 Considerações Finais

A solução proposta neste trabalho possibilita que um usuário sem conhecimento especializado - em bancos de dados relacionais, em linguagens de consulta formais como o SQL e em estrutura de armazenamento dos dados - consiga realizar consultas em bancos de dados descritos por metadados multilíngues utilizando na consulta inicial palavras-chave ou linguagem natural.

As etapas de identificação de funções, segmentação da consulta e expansão da consulta foram adaptadas da arquitetura SQUIRREL [46] para comporem a solução proposta nesse trabalho. Enquanto as etapas de remoção de *stopwords*, extração do radical e tradução da consulta foram acrescentadas para possibilitar a descrição da consulta em linguagem natural, a consulta em metadados descritos em múltiplos idiomas e aprimorar a efetividade da solução. Além disso, as principais técnicas de processamento de linguagem natural observadas nos trabalhos relacionados de consultas multilíngues foram utilizadas no pré-processamento da consulta inicial da solução proposta nessa pesquisa.

CAPÍTULO 5

Avaliação da Solução Proposta

Neste capítulo, são apresentados os seguintes elementos: detalhes da implementação da solução proposta; modelo e as fontes de dados utilizadas; consultas iniciais que foram submetidas no idioma Inglês, as respectivas traduções para o idioma Português e a semântica pretendida em cada consulta; métricas utilizadas para avaliação dessas consultas; resultados esperados e; uma análise dos resultados encontrados.

A avaliação de consultas em bancos de dados relacionais é um processo complexo devido à subjetividade que há em analisar se a consulta obteve os resultados esperados pelo usuário. A maioria dos estudos relacionados a avaliação de consultas em bancos de dados utilizam diferentes conjuntos de dados e métricas. Devido a essa heterogeneidade dos modelos de avaliação é difícil comparar diretamente esses sistemas [2]. Frente a essa realidade, para desenvolver a avaliação da solução proposta e facilitar possíveis análises com outros trabalhos utilizou-se um conjunto de dados e métricas comumente empregado em trabalhos relacionados.

5.1 Implementação

Para validar a solução proposta e os objetivos deste trabalho, foi desenvolvida uma implementação utilizando a linguagem *python* 3.8.0[1] e o *framework* Flask 1.1.2[2] para disponibilizar a interface Web apresentada na Figura 5.1. A linguagem *python* foi utilizada porque é uma linguagem simples e abrangente que apresenta alta portabilidade entre sistemas operacionais. Além disso, é uma das principais linguagens utilizadas para processamento de linguagem natural e dispõe de excelentes funcionalidades para processamento de dados linguísticos.

A implementação teve como foco principal a didática e a facilidade de compreensão de cada etapa da consulta em detrimento de boas práticas de otimização e desem-

[1] https://www.python.org
[2] https://flask.palletsprojects.com

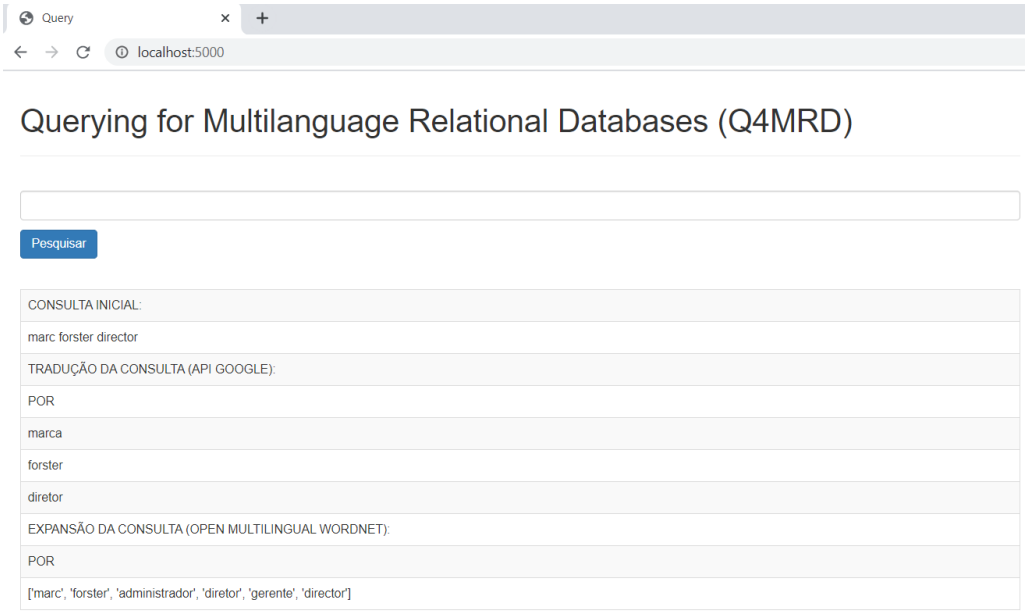


Figura 5.1: Interface da solução proposta

penho. Para realizar o pré-processamento da consulta inicial são utilizados os seguintes pacotes adicionais:

- **NLTK**[3]**:** o *Natural Language Toolkit* é a principal plataforma para construção de aplicações em linguagem *python* que trabalha com dados da linguagem humana. Fornece interface para mais de 50 corporas e recursos lexicais como a *Wordnet* e ainda possui um conjunto de bibliotecas para processamento de textos.
- ***String*:** biblioteca que realiza diversas operações com *strings* e é utilizada para remover as pontuações da consulta inicial.
- ***mysql*:** biblioteca utilizada para acessar o sistema gerenciador de banco dados MySql.
- ***subprocess*:** módulo que possibilita a linguagem *python* executar aplicações desenvolvidas na linguagem Java e obter seu código de retorno.
- ***textblob*:** biblioteca para processamento de dados textuais que oferece, entre vários recursos, a possibilidade de traduzir palavras para vários idiomas.
- ***PyCountry*:** biblioteca que fornece o banco de dados padrão ISO para linguagens (ISO 639-3) e países (ISO 3166), entre outras funcionalidades. É utilizada para converter a sigla do idioma armazenado na tabela TME que se encontra no padrão ISO 639-2 para o padrão ISO 639-1 usado pela biblioteca de tradução do *textblob*.

Os experimentos foram realizados utilizando como equipamento um computador com processador Intel Core i7-8700T 2.40 GHZ de CPU, 32GB de memória RAM, disco

[3] https://www.nltk.org

rígido SSD de 256 GB e sistema operacional Windows 10 de 64 bits. Esse equipamento executa todos serviços necessários para disponibilizar o sistema e as fontes de dados a serem consultadas de forma dedicada, ou seja, não há concorrência.

Na implementação realizada a consulta inicial informada pelo usuário na interface Web é normalizada padronizando todos os caracteres em minúsculos. Essa operação é realizada para aumentar a possibilidade de ocorrências com os dicionários utilizados para processamento de linguagem natural que, geralmente, utilizam como padrão todos os caracteres com letras minúsculas em seus registros.

Após a normalização é realizada uma verificação para identificar se há algum termo entre aspas simples que possa representar uma palavra composta ou termo de valor no banco de dados. Os termos entre aspas simples não são pré-processados sendo preservados em sua forma original e transliterados para o mapeamento da consulta após o pré-processamento.

A segmentação da consulta é a primeira etapa da solução proposta. É utilizada uma técnica para segmentar determinadas palavras compostas para um único *token*. Essas palavras devem estar previamente registradas em uma lista na qual contém as palavras compostas que podem indicar o uso de funções de agregação, agrupamento ou ordenação como, por exemplo: *'order by'*.

Essa técnica altera o caractere delimitador das palavras compostas para o *underline* e na lista de palavras reservadas, que identificam as funções, as palavras são cadastradas usando esse mesmo caractere delimitador para mapear as funções. Um exemplo de registro dessa lista seria mapear *'how_many'* para *'count'*. As palavras são segmentadas utilizando o caractere espaço em branco. No processo de segmentação os caracteres de pontuação da consulta são removidos para melhorar a correspondência dos termos da consulta com dicionários e com a estrutura de dados.

Após a segmentação da consulta é realizada a identificação das funções que possam indicar o uso de agregação, ordenação ou agrupamento. É necessário segmentar previamente a consulta para que seja possível analisar os termos individualmente e, assim, identificar se algum termo da consulta possui correspondência com a lista de palavras reservadas. Essa lista de palavras foi expandida para contemplar termos em linguagem natural que usem palavras compostas e indiquem funções como o *'how many'* que pode sugerir o uso da função *'count'*.

Os termos identificados como funções são removidos da consulta e o termo subsequente é mapeado como o termo ao qual será aplicada a função. A identificação de funções deve ser realizada antes da tradução da consulta, pois em muitos idiomas a tradução pode modificar a posição das palavras na consulta. Isso prejudica a regra em que o termo subsequente é mapeado como o termo em que será aplicada a função.

Por exemplo, na consulta *'Total salary of department project'* a palavra *'Total'*

pode ser mapeada para a função de agregação *'sum'* e a palavra *'salary'* como o termo a ser agregado. Mas, na língua portuguesa a tradução seria 'Salário total do departamento Projeto', com isso o termo 'Departamento' é que seria identificado como o termo a ser agregado.

Depois de identificadas as funções e os termos aos quais elas serão aplicadas, são removidas as *stopwords* da consulta eliminando os termos muito comuns através de uma lista de palavras reservadas. Dos termos resultantes da remoção de *stopwords* são extraídos os radicais de cada palavra para reduzir as variações morfológicas, tais como o plural. Essa ação possibilita obter uma palavra contida no dicionário linguístico facilitando a correspondência com os dicionários utilizados no pré-processamento. A extração do radical é importante para obter a expansão e a tradução dos termos da consulta. Ela é realizada usando a técnica de *lemmatization* da *WordNet* que retorna o *lemma*, a menor unidade de cada palavra.

Por último, é realizada a tradução e a expansão da consulta para os idiomas disponíveis na TME. Para este fim, verifica-se, no momento da execução da consulta, quais são os distintos idiomas que possuem fontes de dados disponíveis na tabela TME. O código usado para criação da estrutura de dados da TME está disponível no Anexo I. Para a validação da solução proposta, foi cadastrada a fonte de dados do IMDB com metadados no idioma português e a consulta inicial sendo realizada na língua inglesa funcionando, assim, como um sistema *cross-lingual*. Caso houvesse fontes de dados descritas em outros idiomas a solução proposta funcionaria como um sistema *multi-lingual*.

A tradução da consulta inicial é realizada usando a biblioteca *textblob* que utiliza a API do *Google Translate*[4]. Essa etapa retorna a melhor tradução para cada termo da consulta inicial processado pelas etapas anteriores considerando o relacionamento semântico entre as palavras. O resultado obtido nessa etapa pode ser comparado com aquele obtido na página do *Google Translate*. Os termos entre aspas simples não são traduzidos, mas transliterados.

A expansão da consulta é realizada utilizando a *Open Multilingual WordNet* (OMW) que encontra os *synsets* correspondentes ao termo da consulta e, posteriormente, retorna as três primeiras palavras sinônimas traduzidas dos dois primeiros *sysnsets* no idioma identificado na tabela TME. Com isso, podem ser retornado até seis termos traduzidos para cada termo da consulta inicial que serão utilizados como termos expandidos. Na expansão da consulta os termos compostos obtidos, separados por *underline*, decorrentes da tradução, são ignorados para não influenciar negativamente nos resultados da consulta. O resultado obtido nessa etapa pode ser comparado ao obtido na página da OMW[5].

[4] https://translate.google.com.br
[5] http://compling.hss.ntu.edu.sg/omw/cgibin/wngridx.cgi

O código do pré-processamento foi adaptado para manter o termo original da consulta (transliterado) quando não for encontrado o termo da consulta inicial no dicionário de expansão da consulta ou de tradução da consulta. Outra adaptação necessária para o contexto multilíngue se refere a quando existe um registro no dicionário do termo utilizado na consulta inicial, mas não existe uma tradução para o idioma identificado na tabela TME. Nesse caso, o termo original também é mantido, ou seja, transliterado. Essa adaptação é útil, por exemplo, para os casos em que na consulta um dos termos não possuem tradução ou termos similares (nome de pessoa, lugar,etc).

Após a efetivação do pré-processamento, conforme a solução proposta, é executada a etapa de mapeamento da consulta SQL. Essa integração é realizada passando parâmetros do código em linguagem Python para o código JAVA do protótipo do SQUIRREL. Os principais parâmetros passados são: a consulta inicial traduzida; a consulta traduzida expandida; a função de agregação, agrupamento ou ordenação; o termo que a função será aplicada e; o termo de valor informado entre aspas simples na consulta inicial. Sendo que esses três últimos parâmetros só são passados quando há ocorrência deles no pré-processamento.

O código do protótipo do SQUIRREL foi modificado minimamente para possibilitar a integração com a solução proposta. Dentre as principais modificações realizadas, podemos citar: 1. As funcionalidades de pré-processamento foram comentadas no código original, pois já receberá a consulta pré-processada; 2. A interface de inserção da consulta foi modificada para que a consulta seja recebida como parâmetro e não inserida pelo usuário; 3. O código foi exportado para um arquivo com extensão .jar executável e adicionado ao projeto da implementação em Python para ser acionado em tempo de execução pelo código Python.

A remoção das funcionalidades de pré-processamento do SQUIRREL foi necessária porque essas etapas do pré-processamento devem ser realizadas antes da tradução da consulta. Uma vez que ela é realizada no pré-processamento da consulta na solução proposta. Ademais, a identificação de funções deve ser realizada antes da tradução da consulta, pois ela utiliza um dicionário de palavras reservadas que é registrado manualmente para cada idioma. Caso a identificação de funções ocorra após a tradução esse dicionário deverá ser traduzido para cada possível idioma da TME.

No processamento da consulta desenvolvido pelo SQUIRREL, primeiramente, é identificado os bancos de dados que são relevantes para a consulta. Essa identificação é realizada verificando se os termos da consulta expandida correspondem aos termos cadastrados na tabela TME para descrever as fontes de dados. Após a identificação dos bancos relevantes é realizado o cálculo dos pesos numa matriz contendo os termos da consulta e os termos de esquema.

Com a matriz de pesos construída são gerados os mapeamentos, as consultas

SQL e os resultados obtidos com a execução da consulta SQL. O processamento da consulta gera um arquivo de *log* contendo as matrizes de peso para termos de valor e termos de esquema. Durante o processo de mapeamento é realizada a inserção das instruções das consultas já mapeadas na tabela resultados que, por sua vez, é criada em tempo de execução.

5.2 Base de Dados

Para avaliar a solução proposta foi utilizado o *Internet Movie Database*[6] (IMDB) que se auto-intitula como a fonte de dados mais popular e confiável do mundo para conteúdo de filmes, séries de televisão (TV) e celebridades. Ela foi projetada para ajudar os fãs a explorar o mundo dos filmes e programas para decidirem o que assistir. Esse conjunto de dados foi escolhido por possuir alta representatividade contendo milhares de registros, além de ser a principal fonte de dados pública utilizada em trabalhos relacionados e *benchmarks*. Essa fonte de dados está descrita originalmente no idioma Inglês e a sua estrutura de dados original é apresentada na Figura 5.2.

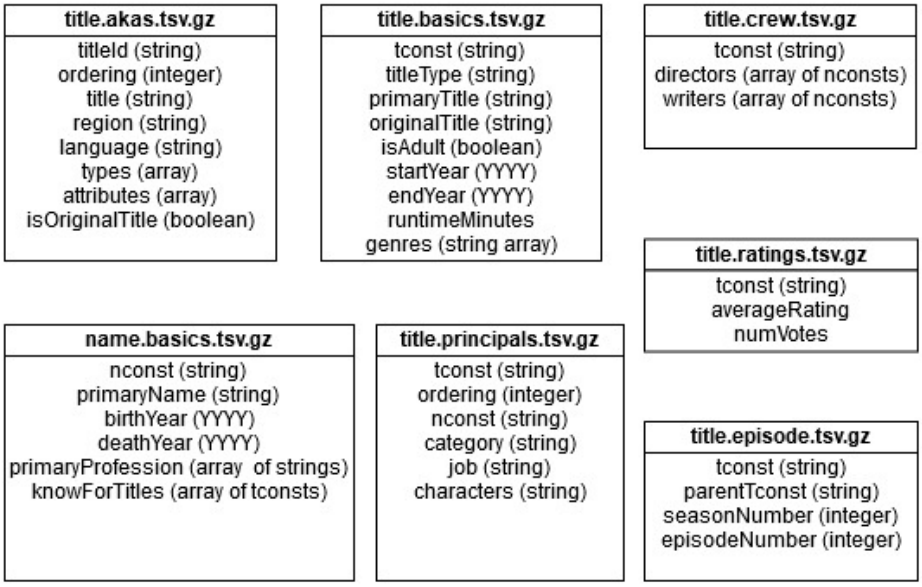


Figura 5.2: Estrutura de dados do IMDB

.

A fonte de dados do IMDB[7] é formada por um conjunto de arquivos compactados no formato GZIP que contém arquivos separados por tabulação no formato TSV. A

[6]Information courtesy of IMDB (http://www.imdb.com). Used with permission.
[7]https://www.imdb.com/interfaces/

Tabela 5.1, apresenta o nome dos arquivos disponibilizados, o tamanho de cada arquivo compactado e o número de registros que cada arquivo possuía em 18/08/2020, data em que o acesso foi realizado.

Tabela 5.1: Conjunto de dados IMDB

Nome do arquivo	**Tamanho (KB)**	**Registros**
title.akas.tsv.gz	196.591	23.206.673
title.basics.tsv.gz	123.535	7.112.660
title.crew.tsv.gz	48.741	7.112.660
title.episode.tsv.gz	27.124	5.103.250
title.principals.tsv.gz	356.091	67.325.316
title.ratings.tsv.gz	5.157	1.067.474
name.basics.tsv.gz	196.491	10.316.291

Os metadados do conjunto de dados do IMDB foram traduzidos para o idioma português para que seja possível realizar a avaliação de consultas em um cenário multilíngue, assim como para se adequar ao objetivo de validação da solução proposta. A organização dos dados e os seus relacionamentos foram baseados no modelo apresentado no *benchmark* [36]. Ele implementa a integridade relacional exigida no processo de mapeamento da consulta e facilita as avaliações da solução proposta nesse trabalho por coadunar com trabalhos semelhantes.

A Figura 5.3 apresenta o Diagrama de Entidade-Relacionamento (DER) do modelo de dados utilizado na validação da solução e a sua estrutura física completa é apresentada nos códigos de A1 a A7, do Anexo I. Esse anexo disponibiliza os códigos de criação da estrutura de dados.

Dentre as principais modificações realizadas na estrutura original do IMDB para o DER proposto em [36], podemos observar que o atributo *'genres'* - que pertencia ao arquivo *title.basics.tsv.gz* como um *array of strings* - foi extraído para uma tabela que possui todos os distintos 'generos'. E, como um filme pode conter vários 'generos' simultaneamente, também foi criada uma tabela que relaciona o 'genero' ao filme. Os atributos *primaryTitle, isAdult, runtimeMinutes* do arquivo *title.basics.tsv.gz* foram removidos por não corresponderem aos termos das consultas utilizadas.

Outra alteração relevante ocorreu no arquivo *name.basics.tsv.gz* que possui os atributos *primaryProfession* do tipo *array of strings* e o atributo *knowForTitles* do tipo *array of tconsts*. Esse arquivo foi extraído para as tabelas pessoa, diretorFilme e atorFilme. Os arquivos *title.ratings.tsv.gz* e *title.basics.tsv.gz* foram extraídos para a tabela filmes e o arquivo *title.episode.tsv.gz* foi extraído para a tabela episódio.

A extração e transformação dos dados do IMDB foram realizadas utilizando a ferramenta *Pentaho Data Integration* (PDI) que é uma ferramenta multiplataforma para

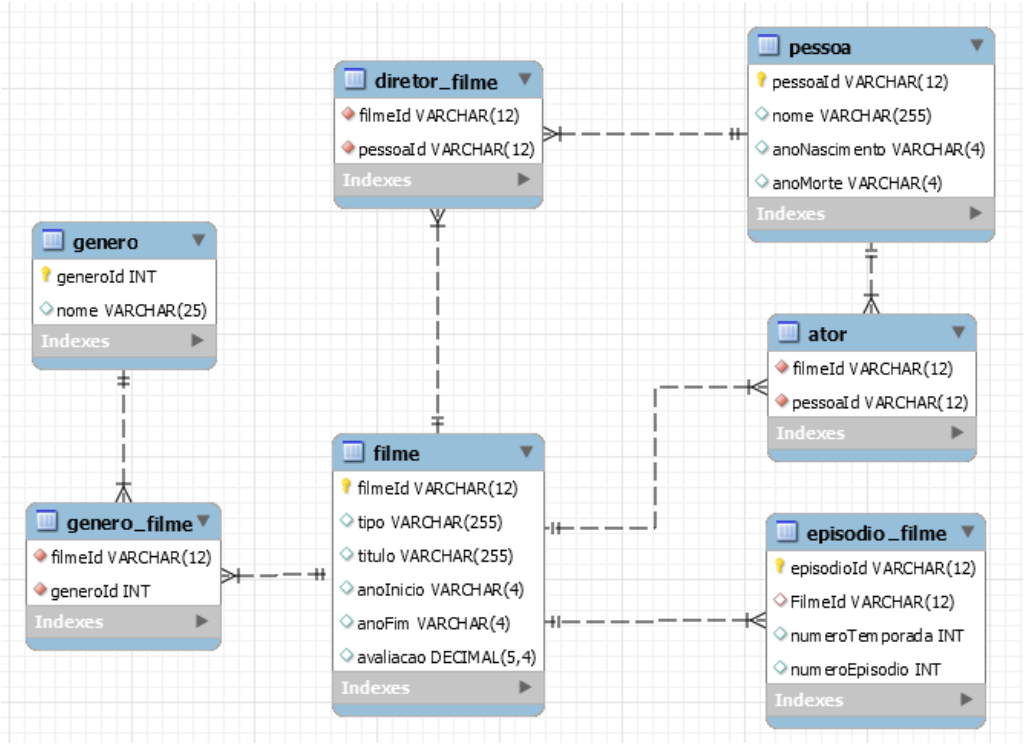


Figura 5.3: Diagrama de Entidade Relacionamento do IMDB adaptado de [36]

.

Extração, Transformação e Carga (ETL) de dados. O PDI possui código aberto, pode ser obtido no repositório do SourceForge[8] e não necessita de instalação para criar ou executar o ETL. O processo de ETL é realizado através de *steps* interligados em um fluxo de dados.

Tabela 5.2: Dados do IMDB em Português

Tabela	**Registros**	**Registros Reduzidos**
filme	7.112.660	488.922
pessoa	10.316.291	562.366
genero	28	28
episodio	5.103.240	4.278.066
diretor_filme	7.768.362	531.099
ator_filme	15.442.651	1.623.938
genero_filme	10.728.136	672.001
Total	**56.471.368**	**8.156.420**

A extração dos dados da fonte original do IMDB e a conversão para o modelo apresentado na Figura 5.3 resultou nas tabelas e registros apresentados na Tabela 5.2.

[8]https://sourceforge.net/projects/pentaho/files/Pentaho%209.0/client-tools/pdi-ce-9.0.0.0-423.zip/download

Para garantir a integridade relacional alguns registros tiveram que ser removidos porque continham quebra de vínculos.

Como o processo de mapeamento de consultas pode demorar várias horas para determinadas consultas, a base de dados do IMDB foi reduzida conforme os critérios adotados pelo *benchmark* [36]. Os critérios utilizados para reduzir a base foram restringir o ano de início dos filmes para após 1980 e o tipo do filme para *movies* e *tvseries*. O quantitativo de registros após a redução pode ser observado na coluna registros reduzidos da Tabela 5.2. Essa redução da fonte de dados visa a aumentar a probabilidade das consultas retornarem resultados em até 60 minutos de execução.

5.3 Consultas

As consultas utilizadas para avaliar a solução proposta na fonte de dados IMDB foram extraídas da proposta de *benchmark* [36]. Essa proposta construiu uma base de consultas de forma semelhante ao que é realizado no cotidiano pelos usuários considerando a inclusão de palavras ambíguas ou que não adicionam sentido à consulta. Esse *benchmark* utilizou o *twiter* para extrair as consultas que são baseadas em tuites e também foi utilizada a ferramenta *Keyword tools*.

Para o conjunto de dados IMDB são utilizadas 50 consultas definidas previamente com suas respectivas semânticas esperadas. Essas consultas, suas respectivas traduções realizadas no pré-processamento da solução proposta e as semânticas pretendidas são apresentadas na Tabela A.1, do Apêndice A.

Todas as consultas foram agrupadas em seis categorias que serão objeto de avaliação, quais sejam: Segmentação da Consulta (SC), Identificação de Funções (IF), Remoção de StopWords (SW), Extração do Radical (ER), Tradução da Consulta (TC) e Expansão da Consulta (EC). Vale ressaltar que uma determinada consulta pode pertencer a mais de uma categoria.

A categoria Segmentação da Consulta avalia se os termos da consulta inicial foram segmentados utilizando o caractere delimitador espaço em branco. Também verifica se os termos compostos inseridos entre aspas simples foram preservados sem a segmentação e transliterados no idioma da consulta inicial.

A categoria Identificação de Funções é responsável por avaliar se a solução proposta realizou a identificação dos termos na consulta inicial que sugerem o uso de funções de agregação, agrupamento e ordenação, mesmo que esses termos sejam palavras compostas. Já a categoria Remoção de *stopwords* verifica se no pré-processamento da consulta é realizada a remoção de palavras que podem ser consideradas *stopwords*, ou seja, palavras que não são comumente utilizadas para descrever os metadados. E, por isso, não encontrarão correspondências no processo de mapeamento da consulta.

A categoria Extração do Radical analisa se as variações ortográficas dos termos da consulta foram removidas. Assim, verifica se houve a extração do radical dos termos e a obtenção da forma canônica para possibilitar um aumento de correspondências entre os termos da consulta e os termos da estrutura de dados. Por sua vez, a categoria Tradução da Consulta investiga se as consultas iniciais no idioma Inglês foram traduzidas para o idioma Português. Para ser considerada como traduzida a consulta deve conter a maioria absoluta dos termos traduzidos desconsiderando, assim, os termos entre aspas simples e nomes próprios que não são traduzidos.

Na fase de Expansão da Consulta é analisado se os termos da consulta inicial foram expandidos com termos sinônimos. Logo, avalia se o pré-processamento da consulta é capaz de adicionar termos sinônimos aos da consulta original e ampliar a possibilidade de correlações.

Na Tabela B.1, do Apêndice B, são apresentados os resultados esperados para todas as consultas. Para cada consulta são apresentadas as categorias que se espera que o pré-processamento da consulta realize, bem como o universo total de registros que serão pesquisados e a quantidade de registros esperados.

Há alguns fatores que podem influenciar no processo de mapeamento da consulta e, consequentemente, na obtenção de resultados, na qualidade dos resultados ou no tempo de processamento da consulta. Os principais fatores identificados são:

- **Padrão de nomenclatura:** o padrão utilizado para nomear as tabelas, atributos e relações podem influenciar no processo de mapeamento. Na base de dados IMDB utilizada nesse trabalho, foi adotado o padrão *SnakeCase* que utiliza o caractere *underline* como separador de espaços para as tabelas e o padrão *lowerCamelCase* para os nomes de atributos. O padrão *lowerCamelCase* não utiliza separador e todas as letras são minúsculas, exceto as iniciais das palavras seguintes à primeira. Como o mapeamento da consulta calcula a similaridade entre os termos da consulta e os termos de esquema da fonte de dados, o uso de padrão de nomenclatura que se aproxime dos termos com que os usuários habitualmente informam na consulta pode aumentar as ocorrências de mapeamento.
- **Relacionamentos:** o uso adequado dos relacionamentos é fundamental para o processo de mapeamento. Portanto, a ausência de chaves estrangeiras prejudica o processo de mapeamento que não encontra os relacionamentos que envolvam várias tabelas.
- **Quantidade de relações e/ou atributos:** quanto maior o número de atributos ou relações, maior será a matriz de pesos criada e maior será o custo computacional para mapear essas consultas. Como os atributos são utilizados no processo de correspondência de similaridade entre os termos da consulta a sua quantidade possui alto impacto no desempenho e tempo de execução da consulta.

- **Quantidade de registros:** quanto maior for o número de registros no banco de dados maior será o custo computacional e o tempo para extrair essas informações do banco de dados. Isso, principalmente, nas consultas em que não são utilizados filtros ou funções de agregação.
- **Quantidade de termos na consulta inicial:** os termos utilizados na consulta inicial são mapeados por similaridade para as tabelas e seus atributos. Assim, quanto maior for o número de termos na consulta maior será a matriz de pesos, o custo computacional e o tempo de execução da consulta.
- **Quantidade de termos expandidos ou traduzidos:** de forma análoga a quantidade de termos na consulta inicial, quanto maior for o número de termos expandidos ou traduzidos, maior será a matriz de pesos, o custo computacional e o tempo de execução da consulta.
- **Tipagem dos dados:** a fonte de dados IMDB utiliza o tipo *string* para todos atributos que são chave primária. Essa prática faz com que o processo de mapeamento realize tentativas de corresponder um termo da consulta inicial do tipo *string* com o atributo chave que geralmente é do tipo inteiro numérico.

5.4 Métricas

O resultado das consultas realizadas com o pré-processamento são ordenados por relevância. Esta é mensurada pela soma da pontuação obtida no cálculo da similaridade entre os termos da consulta inicial e os termos dos metadados. Assim, quanto maior o número de correspondências entre os termos da consulta e os metadados, maior será a relevância dessa consulta.

Para facilitar a compreensão das métricas utilizadas na avaliação da solução proposta, consideremos a Figura 5.4 usada para ilustrar as métricas *precision* e *recall*. Nela, do lado esquerdo temos o conjunto dos resultados relevantes representado pelos subconjuntos A e C. Do lado direito temos o conjunto dos resultados irrelevantes representado pelos subconjuntos B e D. Da mesma forma, temos na parte superior os resultados que foram selecionados na consulta representados pelos subconjuntos A e B. Já na parte inferior temos os resultados que não foram selecionados pela consulta e são representados pelos subconjuntos C e D.

Para avaliar os resultados obtidos com as consultas realizadas foram utilizadas três métricas binárias com valores variando entre 0 (zero) e 1 (um), onde 0 (zero) significa baixa relevância e 1 (um) significa alta relevância. Essas métricas são comumente utilizadas para avaliar consultas como podemos observar no *benchmark* [36]. Desse modo, as métricas utilizadas para avaliar a solução proposta nesse trabalho são:

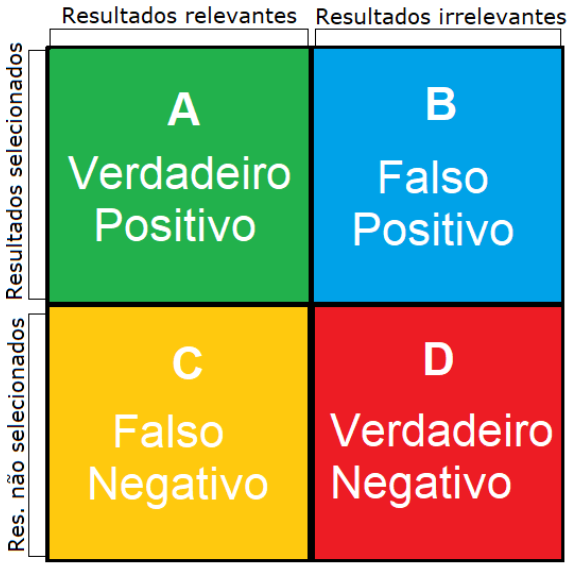


Figura 5.4: Ilustração das métricas *precision* e *recall*

- ***Precision*:** a precisão mede quantos resultados selecionados são relevantes. Ela também pode ser chamada de valor preditivo positivo. Seu cálculo é realizado através da Fórmula 5-1.

$$\textbf{\textit{Precision}} = \frac{Verdadeiro\ positivo}{Verdadeiro\ positivo + Falso\ positivo} = \frac{A}{A+B} \tag{5-1}$$

- ***Recall*:** a revocação mede quantos resultados relevantes foram selecionados. Ela também é conhecida por sensibilidade. Seu cálculo é realizado através da Fórmula 5-2.

$$\textbf{\textit{Recall}} = \frac{Verdadeiro\ positivo}{Verdadeiro\ positivo + Falso\ negativo} = \frac{A}{A+C} \tag{5-2}$$

- ***F-measure*:** é um tipo especial de média (média harmônica ponderada) entre a *precision* e a *recall*. Seu cálculo é realizado através da Fórmula 5-3.

$$\textbf{\textit{F-measure}} = \frac{2 \times (precision \times recall)}{precision + recall} \tag{5-3}$$

As métricas selecionadas avaliam a relevância dos resultados obtidos em relação à semântica esperada pelo usuário no momento da realização da consulta. A semântica pretendida de cada consulta está descrita na Tabela A.1, do Apêndice A.

5.5 Execução das consultas

As consultas apresentadas na Tabela A.1 foram executadas para gerar as informações que serão analisadas. Os resultados da execução dessas consultas foram coletados e

apresentados na Tabela C.1. Para avaliar os resultados da solução proposta são registrados em arquivos de *log* as seguintes informações:

- A consulta inicial informada na interface;
- A consulta segmentada;
- Os termos de valor, quando houver termos na consulta entre aspas simples, ou funções, quando existirem;
- A consulta sem *stopwords*;
- A consulta sem radical;
- A consulta expandida;
- A consulta traduzida.

Nessa etapa, os termos expandidos e traduzidos são utilizados para identificar se há algum banco de dados relevante para a consulta, bem como para verificar se esses termos expandidos possuem correspondências com os campos *dc_title*, *dc_subject*, *dc_description* e *dc_identifier* da tabela TME.

Para cada banco de dados selecionado como relevante é montada uma matriz de pesos para termos de esquema e outra para termos de valor, contendo os termos da consulta inicial e os termos do banco de dados relevante. Se houver algum relacionamento na matriz de esquema são realizados os possíveis mapeamentos para termos de valor, a geração das consultas SQL e a execução no banco de dados para cada mapeamento.

Os experimentos foram realizados limitando em 60 minutos o tempo de execução necessário para iniciar o mapeamento da consulta. Isso considerando os recursos computacionais descritos na Seção 5.1. As consultas que extrapolaram esse limite e não iniciaram o processo de mapeamento foram interrompidas.

Os resultados da execução das consultas foram classificados nas seguintes categorias:

- **Não Existe Banco de Dados Relevante (NEB):** nenhum banco de dados relevante foi encontrado. O uso predominante de termos entre aspas simples e falhas no processo de expansão e tradução da consulta afetaram a identificação de bancos de dados relevantes.
- **Nenhum Resultado Encontrado (NRE):** foram identificados os bancos de dados relevantes, mas nenhum mapeamento foi encontrado para os termos da consulta traduzidos. O uso predominante de termos entre aspas simples e falhas na tradução, na modelagem dos relacionamentos do banco relevante e na nomenclatura dos termos do banco relevante prejudicaram a geração de resultados.
- **Executadas Com Resultados (ECR):** a consulta foi completamente executada, os mapeamentos foram gerados e as consultas SQL geradas foram executadas na base

de dados apresentando resultados relevantes conforme o esperado ou parcialmente conforme o esperado.

- **Executadas Sem Resultados (ESR):** a consulta foi completamente executada, os mapeamentos foram gerados e as consultas SQL geradas foram executadas na base de dados relevante, mas apresentando resultados irrelevantes para a semântica pretendida.
- **Interrompida (INT):** a consulta foi interrompida após 60 minutos de execução sem iniciar a geração de resultados.

Para cada consulta realizada na base de dados IMDB foram identificados os subconjuntos A, B, C e D apresentados na Figura 5.4. Ou seja, foi destacado dentre os resultados da consulta os que são relevantes e os irrelevantes, bem como os resultados que foram selecionados e os que não foram. A identificação desses subconjuntos é necessária para calcular as métricas que foram apresentadas na Seção 5.4: *precision*, *recall* e *f-measure*. Durante a execução das consultas foram analisadas as seis categorias de pré-processamento para identificar se foram realizadas conforme os critérios definidos na Seção 5.3.

O subconjunto A representa os resultados que foram selecionados na consulta e que atendem a semântica pretendida apresentada na Tabela A.1, do Apêndice A, ou seja, os resultados relevantes. O subconjunto B representa os resultados que foram selecionados, mas não atendem à semântica pretendida. O subconjunto C representa a quantidade de resultados que se esperava encontrar (apresentada na Tabela B.1) subtraída do subconjunto A. O subconjunto D representa o universo de resultados (apresentado na Tabela B.1) subtraído dos subconjuntos A, B e C.

As métricas *precision*, *recall* e *f-measure* foram calculadas para as consultas que obtiveram algum valor no subconjunto A utilizando as fórmulas apresentadas na Seção 5.4. Quanto maior o valor de *precision* mais condizentes com a semântica pretendida pela consulta eram os resultados retornados. Já *recall* avalia a proporção dos resultados relevantes retornados com os não retornados e a *f-measure* é uma média entre *precision* e *recall*.

5.6 Análise dos Resultados

Nessa seção é apresentada a análise dos resultados obtidos com a implementação da solução proposta utilizando a fonte de dados IMDB com os metadados descritos no idioma português e as consultas no idioma inglês apresentadas na Tabela A.1. Além disso, é descrita a comparação dos resultados obtidos com os resultados esperados descritos na Tabela B.1 e calculadas as métricas descritas na Seção 5.4 a fim de realizar a análise e descrição dos resultados.

Na Figura 5.5, podemos observar que, das 50 consultas realizadas na base IMDB, 48% (24 consultas) foram executadas completamente (ECR e ESR) em todas as etapas de pré-processamento da consulta. No entanto, dessas 24 consultas executadas por completo, apenas oito retornaram resultados relevantes e 16 consultas não retornaram resultados. Ou seja, realizaram todas etapas e não conseguiram encontrar correspondências entre os termos da consulta e os metadados do banco de dados relevante encontrado.

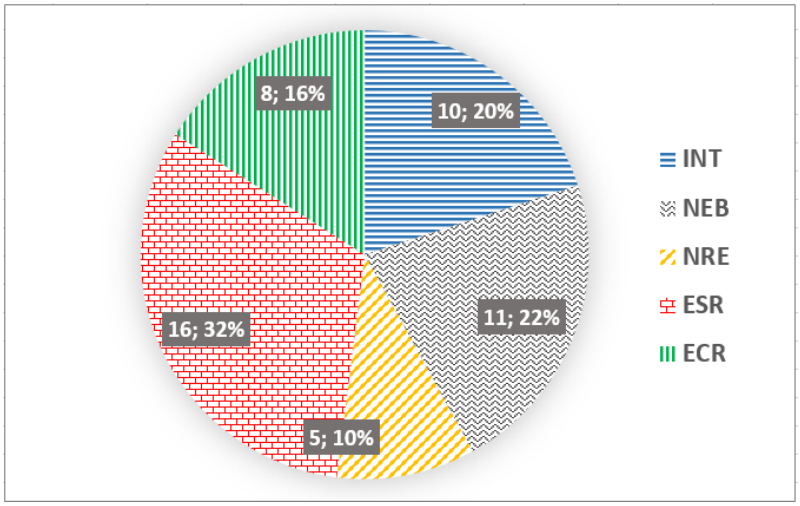


Figura 5.5: Tipo de execução das consultas

Foi identificado também que 11 consultas não encontraram banco de dados relevantes (NEB) com os termos expandidos traduzidos gerados pelo pré-processamento. O principal fator que contribuiu para esse resultado é o uso de palavras na consulta inicial semanticamente distantes das palavras usadas para descrever os dados no banco de dados IMDB, além de falhas na tradução.

Outro aspecto que colaborou para não encontrar bancos de dados relevantes é o fato da tradução dos termos expandidos ser realizada pela OMW, já que ela não retorna um vetor ordenado de possíveis traduções por relevância ou frequência de uso. E como o número de termos traduzidos e expandidos foi limitado nem sempre a melhor tradução ou sinônimo do termo da consulta é retornado.

Foi observado também que 10 consultas tiveram que ser interrompidas (INT), pois após 60 minutos de execução não haviam iniciado a geração de mapeamentos. Duas dessas consultas escolhidas aleatoriamente foram executadas por mais de 12 horas e ainda assim não iniciaram a geração dos mapeamentos. Conjecturando-se, desse modo, que as consultas interrompidas não possuíam potencial de geração de resultados.

As consultas que não obtiveram resultados (NRE), ou seja, tiveram executadas todas as etapas e não encontraram resultados que atendiam a semântica pretendida esperada na consulta, representaram um total de cinco. Nestas consultas, foi observado que a maior parte ou todas as palavras estavam entre aspas simples (lembrando que

essas palavras não são mapeadas para termos de esquema) ou as consultas estavam semanticamente distantes da descrição na fonte de dados.

Quando analisado o tempo de execução das consultas, ilustrado na Figura 5.6, verifica-se que 21 consultas foram executadas em zero segundo sugerindo que houve falhas no processo de execução. As consultas que tiveram esse tempo de execução são todas do tipo INT e NEB, portanto falharam nas etapas iniciais por não encontrarem bancos de dados relevantes (NEB) ou por problemas desconhecidos (INT), conforme apresentado na Figura 5.6.

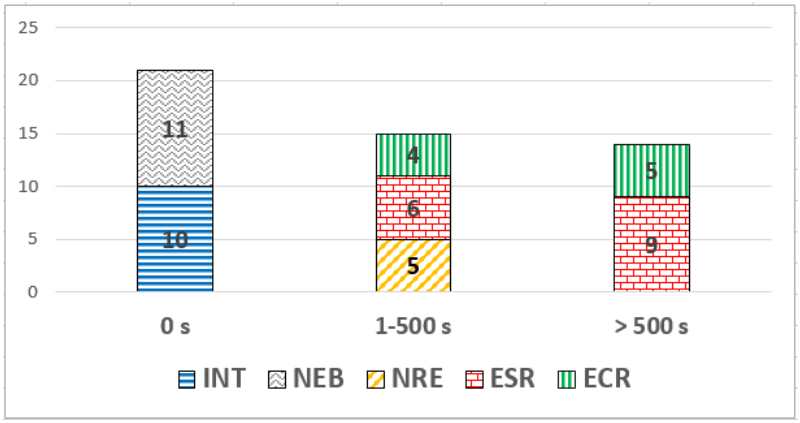


Figura 5.6: Tempo de execução por tipo de consulta

As consultas que gastaram entre 1 e 500 segundos foram as cinco consultas do tipo NRE e as dez consultas que foram completamente executadas. Elas obtiveram até 28 mapeamentos gerados para concluir o processamento da consulta. As 14 consultas restantes gastaram mais de 500 segundos para executar. Essas consultas apresentaram uma quantidade superior de mapeamentos gerados em relação as demais consultas ou tiveram uma quantidade superior de resultados retornados sendo relevantes (ou não). Por isso, elas utilizaram maior tempo de processamento. Dessas 14 consultas mais demoradas, 9 não retornaram resultados relevantes apesar de terem gerado muitos mapeamentos.

Em relação a quantidade de mapeamentos gerados podemos observar, na Figura 5.7, que a maioria absoluta das consultas não geraram mapeamentos e que essas consultas são as do tipo NRE, NEB e INT. Das 14 consultas que obtiveram entre 1 e 500 mapeamentos, seis obtiveram resultados relevantes. Enquanto que das 10 consultas que geraram mais de 500 mapeamentos, apenas duas obtiveram resultados relevantes demonstrando que a quantidade de mapeamentos gerados não é proporcional ao número de resultados relevantes.

A Figura 5.8 compara as etapas do pré-processamento esperado com as etapas do pré-processamento realizado para as consultas apresentadas na Tabela A.1. Pode-se observar que não houve uma variação considerável entre as etapas de pré-processamento

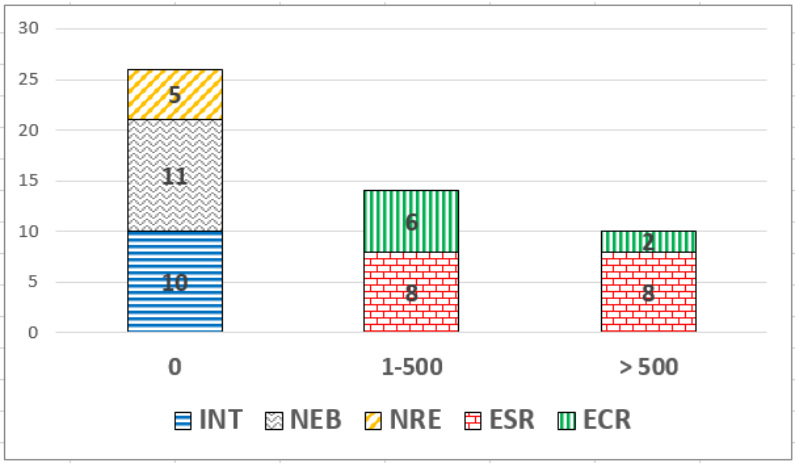


Figura 5.7: Mapeamentos gerados por tipo de execução

realizado e esperado e quando houve divergência o pré-processamento realizado foi inferior ao esperado na maioria dos casos.

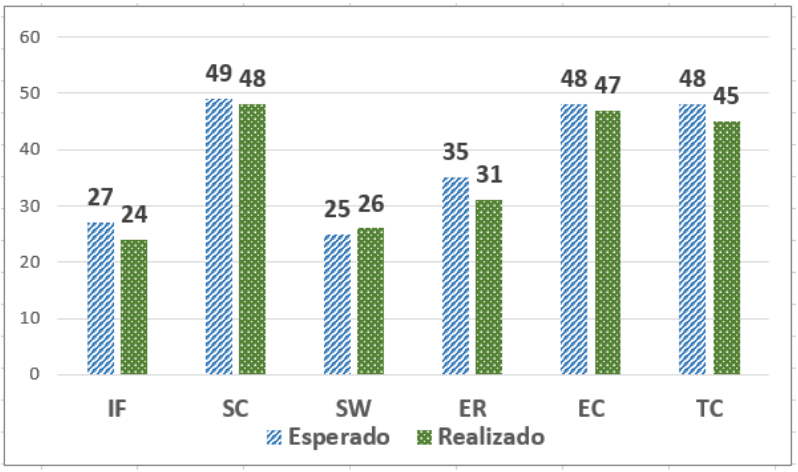


Figura 5.8: Pré-processamento esperado e realizado

As consultas de número 14, 35 e 37 não realizaram a identificação de funções conforme o esperado. As consultas 14 e 35 possuem o termo *'best'* que deveria ser mapeado para a função *'max'*, mas esse termo não possui entrada na lista de palavras reservadas para a identificação de funções. Já a consulta 37 deveria mapear o termo *'for'* para a função de agrupamento *'group by'*, porém esse termo também não possui registro na lista de palavras usadas para identificação de funções.

Quanto à segmentação, a consulta [*'Wayne Roberts' and 'Roger Moore'*] não executou o pré-processamento esperado porque ao identificar o termo entre aspas simples não conseguiu reconhecer que havia dois conjuntos de termos entre aspas. Com isso, identificou toda a consulta inicial como um único termo de valor e não realizou a segmentação da consulta.

A etapa de remoção de *stopwords* (SW) obteve um resultado atípico, pois removeu os termos *after* e *not* que não eram esperados nas consultas [*horror movies after 2018*] e [*movies not released actresses 'rachel mcadams'*]. Além disso, não conseguiu remover a *stopword* esperada da consulta [*'Wayne Roberts' and 'Roger Moore'*], já que o pré-processamento identificou a consulta como um único termo devido ao uso de dois conjuntos de termos entre aspas simples. Nessa etapa foi avaliado apenas se houve a remoção de alguma *stopword* da consulta inicial. Não sendo avaliada a qualidade desse processo de remoção ou se todas as *stopwords* da consulta foram removidas. A principal operação realizada nessa etapa foi a remoção das flexões de plural das palavras.

As consultas 3, 10, 29 e 47 não realizaram a extração do radical conforme o esperado e o principal motivo foi a não extração de palavras no gerúndio. A consulta [*'Steven Spielberg' directoring*] não realizou a expansão da consulta esperada devido à inexistência de gerúndio para a palavra no tesauro utilizado. E as consultas 10, 12 e 46 apresentaram inconsistências na tradução da consulta.

Outra observação relevante é que houve consultas que não foram traduzidas conforme o esperado e a tradução é a etapa mais importante do pré-processamento para sistemas de consulta multilíngue. Nessa etapa não foi avaliado a qualidade da tradução, mas apenas se a tradução foi realizada ou se os termos foram transliterados. Na consulta [*marc forster director*] houve a tradução de um termo que não era esperado porque ele era um nome próprio que não foi informado entre aspas simples.

Um fator que contribuiu negativamente para a tradução foi a submissão de consultas em que todos os termos, ou a maior parte dos termos, foram inseridos entre aspas simples e a solução realizou a transliteração ao invés da tradução. Nesses casos, a consulta deve ser melhor descrita para representar a intenção do usuário colocando entre aspas simples apenas os termos compostos ou que representem um nome próprio.

O uso de termos entre aspas simples também impactou a etapa de segmentação da consulta, pois esses termos não são segmentados. A consulta que não foi segmentada conforme o esperado possuía mais de um conjunto de termos entre aspas simples e o pré-processamento da consulta não foi capaz de lidar com essa situação segmentando esses conjuntos de termos.

A etapa de identificação de funções não reconheceu as funções em três consultas como era esperado. Isso porque essa etapa usa uma lista parcial de palavras que sugerem o uso de funções sendo utilizada para exemplificar a solução proposta. Essa situação indica que a lista precisa ser enriquecida para contemplar todas as possíveis palavras que podem sugerir funções de agregação ou ordenação.

Na Figura 5.9 podemos observar que somente oito consultas obtiveram resultados relevantes (identificadas pelo subconjunto A) e que 13 consultas obtiveram resultados irrelevantes para a semântica pretendida (subconjunto B). Ademais, 35 consultas não

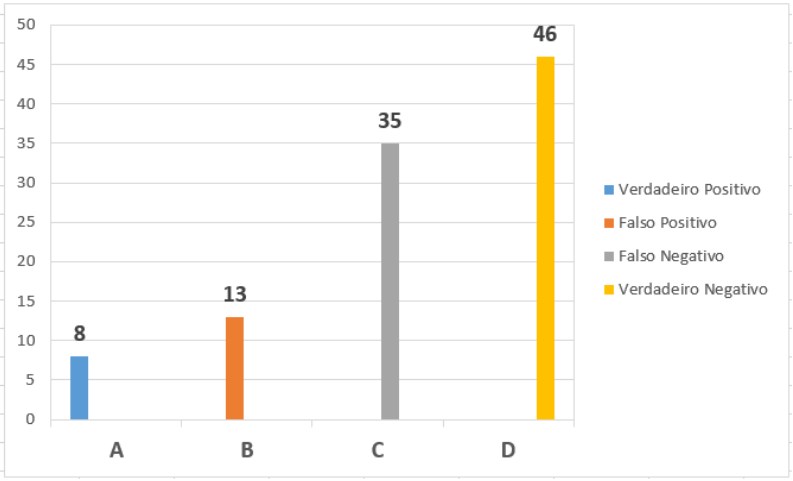


Figura 5.9: Subconjunto de dados das Métricas

retornaram resultados relevantes (subconjunto C) conforme o esperado e 46 consultas tiveram resultados não selecionados (subconjunto D), ou seja, não retornaram resultados relevantes ou retornaram parcialmente os resultados.

Na Figura 5.10 consta a média das três métricas usadas para avaliar a solução proposta. O cálculo considerou somente as oito consultas que obtiveram resultados relevantes retornados. Apesar de apenas 16% das consultas retornarem resultados relevantes, elas obtiveram um valor médio de *precision* acima de 66%.

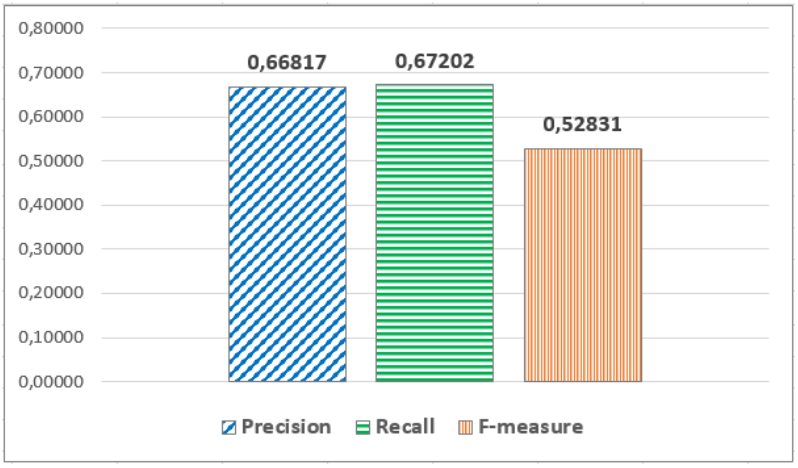


Figura 5.10: Média das métricas para as 8 consultas com resultados

O valor médio do *recall* foi superior a 67% para as consultas que obtiveram resultados relevantes, demonstrando que a maior parte dos resultados relevantes foram recuperados. Assim, apenas 33% dos resultados relevantes não foram recuperados, ou seja, a solução proposta é bem sensível para obtenção de resultados relevantes.

Já a métrica *f-measure* que realiza uma média harmônica ponderada entre a *precision* e a *recall* apresentou valor médio de 52% valor mais baixo que as demais métri-

cas. Considerando o contexto multilíngue da solução proposta, as métricas apresentaram resultados comprovando que a maioria dos resultados relevantes são recuperados nos experimentos realizados.

Na Tabela B.1 podemos observar que 10 consultas (20% das consultas utilizadas nos experimentos) não possuem quantidade de registros esperados, ou seja, os termos utilizados na consulta inicial não possuem correspondência com os termos de esquema ou de valor do banco de dados IMDB. Portanto, essas consultas não têm potencial de gerar resultados condizentes com a semântica pretendida na consulta.

A consulta [*movies princess disney*] não possui registros esperados porque não há na base de dados do IMDB filmes que possuam no título o termo 'princesa', bem como tenham sido dirigidos por alguma pessoa que contenha no nome o termo *'disney'*. Assim, não foram encontrados filmes que contenham os termos 'princesa' e *'disney'* juntos.

Nas consultas de número 11, 12, 16, 21, 34, 45 e 48 não há registros na base de dados IMDB reduzida que correspondam aos termos utilizados na consulta inicial e que satisfaçam a semântica pretendida da consulta. Um dos motivos para não encontrar os resultados esperados nessas consultas foi a redução realizada nessa base. Haja vista que isso restringiu os tipos de filmes e o período em que se realizaram como, por exemplo, a consulta 21 e 16 que são filmes do tipo *'short'*.

A consulta [*movie comedian of the 1920s and 1930s*] também não apresentou resultados esperados, uma vez que a base reduzida não possuía registros de filmes antes da década de 80. A consulta [*movies without director*] não retornou o resultado esperado porque a existência de filmes sem diretor violaria a restrição de integridade relacional do modelo de dados utilizado.

Para realizar um outro nível de validação da solução proposta o SQUIRREL foi usado para executar as mesmas consultas na mesma fonte de dados utilizada na validação da solução proposta. Para essa análise os metadados foram descritos no idioma inglês, pois a versão utilizada do SQUIRREL não atua em contextos multilíngues.

Quando analisados os resultados da execução utilizando o protótipo original do SQUIRREL, com as mesmas consultas e fontes de dados usadas no experimento realizado neste trabalho, foi observado que algumas consultas obtiveram resultados diferentes. Essas diferenças são decorrentes principalmente das etapas de pré-processamento que foram incorporadas na solução proposta.

Dentre as principais diferenças podemos destacar que as cinco consultas que não encontraram resultados (NRE) na solução proposta, com o protótipo do SQUIRREL não identificaram banco de dados relevantes (NEB). Ou seja, não executaram a primeira etapa do processo de mapeamento. Ademais, 24 consultas foram executadas com ou sem resultados (ECR e ESR) na solução proposta. Enquanto com o protótipo do SQUIRREL, dessas 24 consultas, quatro não encontraram banco de dados relevantes (NEB) e quatro

tiveram que ser interrompidas (INT).

Por outro lado, na solução proposta cinco consultas não conseguiram encontrar banco de dados relevante (NEB) enquanto no protótipo do SQUIRREL foi possível encontrar um banco relevante apesar de não encontrar resultados (NRE). O principal motivo identificado é decorrente de falhas no processo de tradução da expansão da consulta na qual o dicionário utilizado não prioriza os termos mais frequentes ou mais próximos semanticamente, bem como devido a descrição parcial ou incompleta dos metadados.

5.7 Ameaças à validade

Durante a execução dos experimentos e análise dos resultados foram identificadas algumas situações que podem ameaçar a validade dos resultados, entre eles: a dependência da qualidade dos dicionários ou tesauros multilíngues que podem apresentar tradução de termos distantes semanticamente do termo original.

A quantidade de resultados obtidos na execução das consultas foi calculada somando os resultados das consultas SQL geradas e executadas sem verificar se havia resultados em duplicidade. Portanto, a quantidade total de resultados selecionados relevantes ou irrelevantes podem conter resultados idênticos contabilizados.

Houve casos em que os resultados considerados relevantes corresponderam aos resultados esperados encontrados e ainda assim a consulta obteve resultados irrelevantes. Nestes casos, os resultados irrelevantes foram desconsiderados para que o valor dos subconjuntos A, B, C e D coincidissem com o valor esperado e com o universo de resultados possíveis.

Os metadados do conjunto de dados do IMDB foram adaptados para o modelo relacional e traduzidos para o idioma português para que seja possível realizar a avaliação de consultas em um cenário multilíngue, já que não foi possível encontrar um fonte de dados relacional pública utilizada em trabalhos relacionados no idioma português.

5.8 Considerações Finais

A implementação da solução proposta é capaz de realizar o pré-processamento da consulta e acionar a implementação do SQUIRREL para realizar o processamento da consulta de forma eficiente. O banco de dados IMDB utilizado nos experimentos teve seus metadados traduzidos para o português e reduzido seu tamanho para agilizar os experimentos. E as consultas utilizadas foram extraídas do *benchmark* [36], bem como a semântica pretendida. As métricas *precision*, *recall* e *f-measure* utilizadas para avaliar

a efetividade das consultas foram calculadas para oito consultas que tiveram resultados relevantes retornados.

Portanto, é possível obter resultados relevantes em bancos de dados descritos por metadados multilíngues utilizando a solução proposta para pré-processamento da consulta. Adequando o formato da consulta e aumentando a diversidade de fontes de dados é possível potencializar ainda mais a quantidade de resultados relevantes retornados com a solução proposta.

CAPÍTULO 6

Conclusão

Conforme apresentado no decorrer desse trabalho, evidenciou-se que as principais abordagens que possibilitam um usuário - sem conhecimento prévio da linguagem de consulta e dos metadados - realizar consultas em bancos de dados relacionais são monolíngues. Neste texto, abordagens monolíngues são aquelas em que as fontes de dados consultadas estão descritas no mesmo idioma da consulta inicial informada pelo usuário. Com isso, mesmo existindo fontes de dados disponíveis descritas em outros idiomas e com potencial para satisfazer a semântica pretendida pelo usuário, essas fontes multilíngues são ignoradas pelo sistema de consulta. Esse fato se constitui em uma problemática porque reduz a capacidade de obtenção de resultados relevantes.

Outra fragilidade diz respeito às propostas que abordam as consultas por palavras-chave ou linguagem natural em bancos de dados relacionais descritos por metadados multilíngues. Nessas propostas não há uma especificação detalhada das etapas de pré-processamento ou dos métodos utilizados. Ademais, uma parte das propostas são dependentes de domínio, ou seja, necessitam ser reconfiguradas para outros domínios de dados.

Para superar tais limitações, neste trabalho é apresentada uma solução de pré-processamento da consulta a fim de possibilitar que um usuário realize consultas utilizando palavras-chave ou linguagem natural em bancos de dados descritos por metadados multilíngues, mesmo que esse usuário não possua prévio conhecimento da linguagem de consulta ou dos metadados dos bancos de dados relacionais. É importante enfatizar que a consulta por palavras-chave ou linguagem natural submetidas a banco de dados relacionais descritos por metadados multilíngues potencializa o acesso às informações, uma vez que possibilita a busca por informações que coincidam com a semântica pretendida pelo usuário no momento da consulta em idiomas diversos ao informado na consulta inicial.

6.1 Contribuições e limitações

A solução proposta nessa pesquisa apresentou resultados promissores para a consulta em bancos de dados relacionais descritos por metadados multilíngues utilizando

linguagem natural ou palavras-chave. Isso porque demonstrou a viabilidade de consultar informações em bancos de dados que podem conter informações relevantes mesmo estando descritos em idioma diverso ao utilizado na consulta inicial.

Na solução proposta o pré-processamento tem papel fundamental na efetividade da consulta, já que remove termos que não contribuem para a obtenção de resultados precisos, além de reduzir as variações morfológicas das palavras, normalizar a sentença da consulta para uma forma comum, expandir a consulta adicionando novos termos sinônimos que enriquecem a consulta inicial aumentando as correspondências no mapeamento e traduzir a consulta para outros idiomas possibilitando a realização de consultas multilíngues. Quando analisado o pré-processamento da consulta, observa-se que os resultados obtidos foram muito próximos do esperado, conforme apresentado na Figura 5.8.

Outro aspecto relevante a se destacar sobre a solução desenvolvida neste trabalho é que ela é portável, ou seja, é independente de domínio. Assim, não necessita de detalhes internos do banco de dados para realizar o pré-processamento da consulta. Inclusive é possível consultar fontes de dados diversas, em domínios variados e descritos em idiomas distintos. Ademais, na avaliação das consultas que retornaram resultados, as métricas *precision* e *recall* demonstraram que a solução proposta consegue obter alta assertividade de resultados relevantes.

Outra contribuição desse estudo foi a definição da ordem de execução para cada etapa de pré-processamento, apresentando os motivos da hierarquia de execução entre cada etapa, bem como os detalhes, métodos e técnicas utilizadas nas atividades de cada etapa de pré-processamento. Foi realizada a classificação e identificação dos resultados esperados e executados em cada etapa de pré-processamento para cada consulta.

No entanto, durante a execução dos experimentos também foram identificadas algumas limitações, quais sejam:

- Algumas consultas (total de 11) não encontraram banco de dados relevantes mesmo pertencendo ao contexto da fonte de dados. A principal justificativa para isso se deve ao uso da OMW no processo de tradução da consulta expandida, pois ela não retornava uma lista ordenada de traduções mais relevantes ou frequentes do termo e possui registros inconsistentes de tradução.
- Houve consultas (total de 10) que foram interrompidas após 60 minutos de execução sem iniciar a geração de algum resultado. Nesses casos, a etapa de pré-processamento da consulta foi realizada, mas o processo de mapeamento das consultas não conseguiu iniciar a obtenção de resultados no tempo decorrido.
- A existência de consultas em que é impossível obter resultados relevantes. Um exemplo é a consulta *movies without director* descrita na Tabela A.1. Conforme o Modelo de Integridade Relacional, apresentado na Figura 5.3, um diretor sempre

está relacionado a algum filme, não podendo haver registros de filmes sem ator ou diretor.

- O mapeamento da consulta não funciona com metadados (nomes de tabelas e atributos) que não possuem semântica como, por exemplo: tab1(col1,col2). Devido a isso, a efetividade na obtenção de resultados é dependente da proximidade semântica entre os termos usados para descrever os metadados do banco de dados e os termos utilizados na consulta.
- A restrição do idioma a língua Inglesa. Essa escolha se deve ao fato de ser o idioma que possui mais recursos de processamento de linguagem natural com qualidade para realizar os experimentos, além de ser o idioma predominante em trabalhos relacionados.
- Os metadados da fonte de dados do IMDB foram traduzidos pelo autor para o idioma Português de forma a viabilizar os experimentos, pois as principais fontes de dados utilizadas em trabalhos relacionados estão descritas no idioma Inglês e não foi encontrada uma fonte de dados com metadados em Português que tivesse quantidade de registros robusta e já tivesse sido utilizada em trabalhos relacionados.
- Para a obtenção dos resultados experimentais foi utilizada apenas uma fonte de dados, restringindo uma análise comparativa com outros esquemas de dados, com diferentes quantidades de tabelas, atributos e relacionamentos. Essa restrição da fonte de dados ocorreu porque não é o objetivo desse trabalho essa comparação, mas sim avaliar o contexto multilíngue de consultas em bancos de dados relacionais.
- As consultas em que todos os termos estavam entre aspas simples (por exemplo as consultas *'good versus evil'* e *'Wayne Roberts' and 'Roger Moore'*) não foram executadas corretamente, pois os termos entre aspas simples são mapeados como um único termo de valor e não há outros termos na consulta que possam ser mapeados para termos de esquema. Esse mapeamento para termos de esquema é necessário para realizar o processo de mapeamento e geração da consulta SQL.
- As listas de palavras reservadas utilizadas no pré-processamento da consulta são parciais ou de domínio geral e amplo. Portanto, a revisão dessas listas para contemplar todos os possíveis termos que são adequados ao contexto utilizado na fonte de dados pode contribuir para melhoria da efetividade das consultas. Por exemplo, a palavra *'of'* está presente na lista de *stopwords*, mas pode representar a intenção de agrupamento se estiver acompanhada da palavra *'each'*. Então no contexto de consultas em bancos de dados a lista de *stopwords* não necessitaria conter a palavra *'of'*.

No entanto, mesmo diante da existência de limitações a solução proposta apresenta muitas contribuições para a área como já mencionado. Logo, a principal contribuição desse trabalho consiste na definição e implementação detalhada das etapas de pré-

processamento necessárias para realizar consultas com palavras-chave em banco de dados descritos por metadados multilíngues, bem como na validação dessa solução proposta comprovando seu funcionamento.

6.2 Trabalhos Futuros

Durante a evolução da pesquisa e o desenvolvimento da solução proposta neste trabalho foram identificadas funcionalidades que poderiam contribuir para melhorias da referida solução. No entanto, elas não estavam previstas no escopo desse estudo ou não poderiam ser utilizadas considerando o cronograma proposto para desenvolvimento do trabalho. Essas funcionalidades podem ser investigadas em outras pesquisas para analisar sua viabilidade de contribuição para a solução proposta, são elas:

- Avaliar o uso da técnica de processamento de linguagem natural *Part of Speech tagging* para classificar gramaticalmente cada termo da consulta e aumentar a acurácia do pré-processamento, passando a classe gramatical dos termos para as etapas de expansão e tradução da consulta de forma que essas operações sejam realizadas somente para termos de mesma classe gramatical.
- Possibilitar que a consulta inicial seja realizada em múltiplos idiomas e a recuperação de resultados também em múltiplos idiomas. Para isso, é preciso fazer o reconhecimento do idioma utilizado na consulta, disponibilizar as fontes de dados descritas por metadados em diversos idiomas, definir consultas para múltiplos idiomas e avaliar se as etapas do pré-processamento terão os recursos necessários para os idiomas utilizados.
- Gerar visualizações dos resultados relevantes encontrados na forma de gráficos (barras, colunas, linha, pizza e etc) automaticamente quando o formato do resultado permitir.
- Identificar possíveis termos de valor (nome de pessoas, lugares, data, números e etc) automaticamente utilizando a técnica de processamento de linguagem natural *Named Entity Recognition* (NER) para que o usuário não precise informar esses termos entre aspas simples.
- Disponibilizar um conjunto de dados representativo com os dados e os metadados descritos no idioma Português viabilizando o processo de consulta multilíngue em fontes de dados nesse idioma.

Desse modo, os resultados demonstram que há pontos de melhoria no processo de consulta que vão desde modernizar a identificação dos bancos de dados relevantes até a otimização do processo de mapeamento para que seja possível iniciar o mapeamento de todas consultas.

Referências Bibliográficas

[1] ADITYA, B.; BHALOTIA, G.; CHAKRABARTI, S.; HULGERI, A.; NAKHE, C.; PARAG.; SUDARSHANXE, S. **Banks: Browsing and keyword searching in relational databases**. In: *VLDB '02: Proceedings of the 28th International Conference on Very Large Databases*, p. 1083 – 1086. Morgan Kaufmann, San Francisco, 2002.

[2] AFFOLTER, K.; STOCKINGER, K.; BERNSTEIN, A. **A comparative survey of recent natural language interfaces for databases**. *The VLDB Journal*, 28(5):793–819, Aug 2019.

[3] ALEXANDER, R.; RUKSHAN, P.; MAHESAN, S. **Natural language web interface for database (nlwidb)**. *CoRR*, 07 2013.

[4] ALVARES, R. V.; GARCIA, A. C. B.; FERRAZ, I. **STEMBR: A stemming algorithm for the brazilian portuguese language**. In: *Progress in Artificial Intelligence*, p. 693–701. Springer Berlin Heidelberg, 2005.

[5] ANDROUTSOPOULOS, I.; RITCHIE, G.; THANISCH, P. **Natural language interfaces to databases – an introduction**. *Natural Language Engineering*, 1(1):29–81, 1995.

[6] BAEZA-YATES, R.; RIBEIRO-NETO, B. **Recuperação de Informação: Conceitos e Tecnologia das Máquinas de Busca**. Bookman, Porto Alegre, 2 edition, 2013.

[7] BALLESTEROS, L.; CROFT, W. B. **Resolving ambiguity for cross-language retrieval**. In: *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, p. 64–71, New York, NY, USA, 1998. Association for Computing Machinery.

[8] BERGAMASCHI, S.; DOMNORI, E.; GUERRA, F.; ORSINI, M.; LADO, R. T.; VELEGRAKIS, Y. **Keymantic: Semantic keyword-based searching in data integration systems**. *Proceedings of the VLDB Endowment*, 3(1-2):1637–1640, Sept. 2010.

[9] BERGAMASCHI, S.; FERRO, N.; GUERRA, F.; SILVELLO, G. **Keyword-Based Search Over Databases: A Roadmap for a Reference Architecture Paired with an Evaluation Framework**, p. 1–20. Springer Berlin Heidelberg, Berlin, Heidelberg, 2016.

[10] BERGAMASCHI, S.; GUERRA, F.; SIMONINI, G. **Keyword Search over Relational Databases: Issues, Approaches and Open Challenges**, p. 54–73. Springer Berlin Heidelberg, Berlin, Heidelberg, 2014.

[11] BLUNSCHI, L.; JOSSEN, C.; KOSSMAN, D.; MORI, M.; STOCKINGER, K. **Soda: Generating sql for business users**, 2012.

[12] BOND, F.; FOSTER, R. **Linking and extending an open multilingual Wordnet**. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 1352–1362, Sofia, Bulgaria, Aug. 2013. Association for Computational Linguistics.

[13] BUCKLEY, C.; SALTON, G.; ALLAN, J.; SINGHAL, A. **Automatic query expansion using smart: Trec 3**. In: *TREC*, 1994.

[14] CAMACHO-COLLADOS, J.; PILEVAR, M. T. **On the role of text preprocessing in neural network architectures: An evaluation study on text categorization and sentiment analysis**. *Workshop on Analyzing and interpreting neural networks for NLP*, p. 1–7, 08 2018.

[15] CARPINETO, C.; ROMANO, G. **A survey of automatic query expansion in information retrieval**. *ACM Comput. Surv.*, 44(1), Jan. 2012.

[16] CHANDRA, G.; DWIVEDI, S. K. **A literature survey on various approaches of word sense disambiguation**. In: *2014 2nd International Symposium on Computational and Business Intelligence*, p. 106–109, Dec 2014.

[17] CHANDRA, G.; DWIVEDI, S. **Assessing query translation quality using back translation in hindi-english clir**. *International Journal of Intelligent Systems and Applications*, 9:51–59, 03 2017.

[18] CHANDRA, G.; DWIVEDI, S. K. **Query expansion for effective retrieval results of hindi–english cross-lingual IR**. *Applied Artificial Intelligence*, 33(7):567–593, Apr. 2019.

[19] CHOUDHARY, M.; DUA, M.; VIRK, Z. S. **A web-based bilingual natural language interface to database**. In: *2015 Third International Conference on Image Information Processing (ICIIP)*. IEEE, Dec. 2015.

[20] CODD, E. F. **Seven steps to rendezvous with the casual user.** In: *IFIP Working Conference Data Base Management*, p. 179–200, January 1974. IBM Research Report RJ 1333, San Jose, California.

[21] CROFT, W. B.; METZLER, D.; STROHMAN, T. **Search engines: Information retrieval in practice**. Addison-Wesley, 01 2015.

[22] FAKHRAEE, S.; FOTOUHI, F. **Dbsemsxplorer: Semantic-based keyword search system over relational databases for knowledge discovery**. In: *Proceedings of the Third International Workshop on Keyword Search on Structured Data*, KEYS '12, p. 54–62, New York, NY, USA, 2012. ACM.

[23] Fellbaum, C., editor. **WordNet: An Electronic Lexical Database**. Language, Speech, and Communication. MIT Press, Cambridge, MA, 1998.

[24] GAO, J.; NIE, J.-Y.; XUN, E.; ZHANG, J.; ZHOU, M.; HUANG, C. **Improving query translation for cross-language information retrieval using statistical models**. In: *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, p. 96–104, New York, NY, USA, 2001. Association for Computing Machinery.

[25] GHOSH, S.; GUNNING, D. **Natural Language Processing Fundamentals**. Packt Publishing, Birmingham B3 2PB, UK, 2019.

[26] Goker, A.; Davies, J., editors. **Information Retrieval**. John Wiley and Sons, Ltd, Oct. 2009.

[27] HAAM, D.; LEE, K. Y.; KIM, M. H. **Keyword search on relational databases using keyword query interpretation**. In: *5th International Conference on Computer Sciences and Convergence Information Technology*. IEEE, Nov. 2010.

[28] HARRIS, L. R. **Experience with intellect: Artificial intelligence technology transfer**. *AI Magazine*, 5(2):43, Jun. 1984.

[29] HE, B.; OUNIS, I. **Studying query expansion effectiveness**. In: Boughanem, M.; Berrut, C.; Mothe, J.; Soule-Dupuy, C., editors, *Advances in Information Retrieval*, p. 611–619, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[30] HRISTIDIS, V.; PAPAKONSTANTINOU, Y. **Discover: Keyword search in relational databases**. In: *VLDB*, 2002.

[31] IMRAN, H.; SHARAN, A. **Selecting effective expansion terms for better information retrieval**. *International Journal of Computer Science & Applications*, 7, 12 2010.

[32] JAUHIAINEN, T.; LUI, M.; ZAMPIERI, M.; BALDWIN, T.; LINDÉN, K. **Automatic language identification in texts: A survey**. *J. Artif. Int. Res.*, 65(1):675–782, May 2019.

[33] KARGAR, M.; AN, A.; CERCONE, N.; GODFREY, P.; SZLICHTA, J.; YU, X. **Meanks: Meaningful keyword search in relational databases with complex schema**. In: *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, SIGMOD '14, p. 905–908, New York, NY, USA, 2014. ACM.

[34] KOEHN, P. **Statistical Machine Translation**. Cambridge University Press, USA, 1st edition, 2010.

[35] KUMAR, R.; DUA, M.; JINDAL, S. **D-hird: Domain-independent hindi language interface to relational database**. In: *2014 International Conference on Computation of Power, Energy, Information and Communication (ICCPEIC)*, p. 81–86, April 2014.

[36] LEMOS, A. D. **Avaliação semântica de consultas por palavras-chave a bancos de dados relacionais**. Master's thesis, Programa de Pós-graduação em Ciência da Computação, 2020. Instituto de Informática - INF/UFG.

[37] MADANKAR, M.; CHANDAK, M.; CHAVHAN, N. **Information retrieval system and machine translation: A review**. *Procedia Computer Science*, 78:845–850, 2016.

[38] MARON, M. E.; KUHNS, J. L. **On relevance, probabilistic indexing and information retrieval**. *J. ACM*, 7(3):216–244, July 1960.

[39] MCNAMEE, P.; MAYFIELD, J. **Comparing cross-language query expansion techniques by degrading translation resources**. In: *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR 02*. ACM Press, 2002.

[40] PAGRUT, A.; PAKMODE, I.; KARIYA, S.; KAMBLE, V.; HARIBHAKTA, Y. **Automated sql query generator by understanding a natural language statement**. *International Journal on Natural Language Computing*, 7:01–11, 06 2018.

[41] POIBEAU, T. **Machine Translation**. The MIT Press Essential Knowledge Series, London, England, 1 edition, 2017.

[42] POPESCU, A.-M.; ETZIONI, O.; KAUTZ, H. **Towards a theory of natural language interfaces to databases**. In: *Proceedings of the 8th International Conference on Intelligent User Interfaces*, IUI '03, p. 149–157, New York, NY, USA, 2003. Association for Computing Machinery.

[43] PORTER, M. **An algorithm for suffix stripping**. *Program*, 14(3):130–137, Mar. 1980.

[44] POSEVKIN, R.; BESSMERTNY, I. **Multilanguage natural user interface to database**. In: *2016 IEEE 10th International Conference on Application of Information and Communication Technologies (AICT)*. IEEE, Oct. 2016.

[45] RAHMANI, A. **Adapting google translate for english-persian cross-lingual information retrieval in medical domain**. In: *2017 Artificial Intelligence and Signal Processing Conference (AISP)*. IEEE, Oct. 2017.

[46] RAMADA, M. S.; DA SILVA, J. C.; DE SÁ LEITÃO-JÚNIOR, P. **From keywords to relational database content: A semantic mapping method**. *Information Systems*, 88:101460, Feb. 2020.

[47] RIVAS, A.; IGLESIAS, E.; BORRAJO, M. **Study of query expansion techniques and their application in the biomedical information retrieval**. *The Scientific World Journal*, 2014:1–10, 03 2014.

[48] ROBERTSON, A. M.; WILLET, P. **A Comparison of Spelling-Correction Methods for the Identification of Word Forms in Historical Text Databases***. *Literary and Linguistic Computing*, 8(3):143–152, 01 1993.

[49] SALTON, G. **Automatic processing of foreign language documents**. In: *Proceedings of the 1969 Conference on Computational Linguistics*, COLING '69, p. 1–28, USA, 1969. Association for Computational Linguistics.

[50] SATYAMURTY, C. V. S.; MURTHY, J. V. R.; RAGHAVA, M. **Metadata-based semantic query in multilingual databases**. In: *Advances in Intelligent Systems and Computing*, p. 249–255. Springer Singapore, 2018.

[51] SINGLA, K.; DUA, M.; NANDA, G. **A language based comparison of different similarity functions and classifiers using web based bilingual question answering system developed using machine learning approach**. In: *Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies - ICTCS 16*. ACM Press, 2016.

[52] TEMPLETON, M.; BURGER, J. **Problems in natural-language interface to dbms with examples from eufid**. In: *Proceedings of the First Conference on Applied Natural Language Processing*, ANLC '83, p. 3–16, USA, 1983. Association for Computational Linguistics.

[53] THENMOZHI, D.; ARAVINDAN, C. **Ontology-based tamil-english cross-lingual information retrieval system**. *Sādhanā*, 43(10), Aug. 2018.

[54] VALIVETI, S.; TRIPATHI, K.; RAVAL, G. **Natural language interface for multilingual database**. In: *Information and Communication Technology for Intelligent Systems (ICTIS 2017) - Volume 2*, p. 113–120. Springer International Publishing, Aug. 2017.

[55] VATANEN, T.; VÄYRYNEN, J. J.; VIRPIOJA, S. **Language identification of short text segments with n-gram models**. In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010. European Language Resources Association (ELRA).

[56] VIEIRA, R. C.; SILVA, J. C. **Consulta multilíngue em banco de dados relacionais: Uma revisão sistemática**. In: *Anais da VII Edição da Escola Regional de Informática de Goiás*, p. 21–32, Brasil, 2019.

[57] VIRK, Z.; DUA, M. **An advanced web-based bilingual domain independent interface to database using machine learning approach**. In: *Advances in Intelligent Systems and Computing*, p. 581–589. Springer Singapore, 2017.

[58] WOODS, W.; KAPLAN, R.; WEBBER, B. **The lunar sciences natural language information system**. 07 1972.

[59] ZELLE, J. M.; MOONEY, R. J. **Learning semantic grammars with constructive inductive logic programming**. In: *Proceedings of the 11th National Conference on Artificial Intelligence*, p. 817–822. Menlo Park, CA: AAAI Press, 1993.

[60] ZENG, Z.; BAO, Z.; LE, T. N.; LEE, M. L.; LING, W. T. **Expressq: Identifying keyword context and search target in relational keyword queries**. In: *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, p. 31–40, New York, NY, USA, 2014. ACM.

[61] ZHONG, Z.; NG, H. T. **Word sense disambiguation improves information retrieval**. In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 273–282, Jeju Island, Korea, July 2012. Association for Computational Linguistics.

APÊNDICE A

Consultas

Tabela A.1: Consultas baseadas em Lemos [36]

	Consulta inicial	Consulta traduzida	Semântica pretendida
1	marc forster director	marca forster diretor	Informações sobre o diretor Marc Forster
2	director movie list	diretor filme lista	Lista de diretores
3	finished animation series	acabado animaçao series	Séries do gênero animação que terminaram
4	movies action	filme açao	Lista com todos os filmes de ação
5	movies princess disney	filme princesa disney	Lista com nome dos filmes de princesa da Disney
6	movies years 2020	filme ano 2020	Filmes cujo ano de lançamento seja 2020
7	supernatural episodes list	sobrenatural episódio lista	Lista de episódios da série Supernatural
8	'game of thrones' actors	ator 'game of thrones'	Atores da Série Game of Thrones
9	christmas tv shows	natal televisão mostrar	Shows de TV de Natal
10	'Steven Spielberg' directoring	direção 'steven spielberg'	Todos os filmes dirigidos por Steven Spielberg
11	movies 'James Harkness' acting	filme atuação 'james harkness'	Filmes que o ator James Harkness atuou
12	actor dead 'Fast and Furious'	ator morto 'fast and furious'	Atores do filme Velozes e Furiosos que estão mortos
13	movies with most actress	filme atriz	Filmes com muitos atores
14	best movies ratings 2020	melhor filme avaliação 2020	Filmes melhor classificados e lançados em 2020
15	actor died in 2020	ator morreu 2020	Atores que morreram em 2020
16	movie short 'based true story'	filme curto 'based true story'	Filmes curtos com título based true story
17	movies episode in the sea	filme episódio mar	Episódios de série com título sea
18	avengers movies all actress name	vingador filme atriz nome	Todos os atores e atrizes do filme Avengers
19	episode of 'Yes, Prime Minister'	episódio 'yes, prime minister'	Episódios da série Yes, Prime Minister
20	movies of drama and romance	filme drama romance	Filmes que possuem como gênero drama e romance

Tabela A.1: Consultas baseadas em Lemos [36]

	Consulta inicial	Consulta traduzida	Semântica pretendida
21	'good versus evil'	'good versus evil'	Filmes de nome Good versus evil
22	comedy or documentary	comédia documentário	Informações do gênero comédia e documentário
23	movies about police	filme polícia	O nome de todos filmes sobre policiais
24	number of movies genres	filme gênero	O número de filmes por gênero
25	how many smallville episodes are there	smallville episódio	A quantidade de episódios de Smallville
26	movies with more than one director	filme 1 diretor	Filmes que possuem mais de um diretor
27	movies with 'angelina jolie' and 'brad pitt'	filme 'angelina jolie' and 'brad pitt'	Filmes com atores Angelina Jolie e Brad Pitt
28	movie comedian of the 1920s and 1930s	filme comediante década de 1920 década de 1930	Filmes de comédia lançados entre 1920 e 1930
29	directed a 'star is born'	dirigido 'star is born'	Diretores do filme Star is born
30	i rate 'The Lion King'	taxa 'the lion king'	Classificação do filme The Lion King
31	First episode of Warrior 2019	primeiro episódio guerreiro 2019	Primeiro episódio de Warrior em 2019
32	'Wayne Roberts' and 'Roger Moore'	'wayne roberts' and 'roger moore'	Registros da pessoa Wayne Roberts e Roger Moore
33	cast of 'Jack and Jill'	fundida 'jack and jill'	Todos os participantes do filme Jack and Jill
34	movie director Scorcese and actor 'Di Caprio'	filme director sorcese ator 'di caprio'	Filme do diretor Scorcese e Ator Di Caprio
35	movies best rating	filme melhor avaliação	Filmes melhores classificados
36	year first 'star wars'movie	ano primeiro filme 'star wars'	Primeiro filme da Star Wars
37	rating movies for genres	avaliação filme gênero	Classificação dos filmes por gênero
38	number of films released per year	filme liberado para ano	O número de filmes lançados por ano
39	movies with two genres	filme dois gênero	Filmes que possuem dois gêneros

Tabela A.1: Consultas baseadas em Lemos [36]

	Consulta inicial	Consulta traduzida	Semântica pretendida
40	vikings series start date	viking series começar encontro	Ano de inicio da série Viking
41	number movies types	filme tipo	Número de filmes por tipo
42	how many 'star wars' movies are there	filme 'star wars'	A quantidade de filmes da saga Star Wars
43	movie musical the last 5 years	filme musical último 5 ano	Os filmes do gênero musical dos últimos 5 anos
44	horror movies after 2018	horror filme 2018	Filmes do gênero 'horror' e lançados a partir de 2018
45	movies 'husband wife relation'	filme 'husband wife relation'	Informações sobre o filme Husband and Wife Relation
46	actors 'The Last Word'	ator 'the last word'	Todos os atores do filme The last Word
47	acting and directing at the same time	atuação direção tempo	Atores que atuaram e dirigiram o mesmo filme
48	movies not released actresses 'rachel mcadams'	filme liberado atriz 'rachel mcadams'	Filmes não lançados da atriz Rachel Mcadams
49	movies director 'mel gibson'	filme diretor 'mel gibson'	Filmes do diretor Mel Gilbson
50	movies without director	filme sem diretor	Filmes sem diretor

APÊNDICE B

Resultados Esperados

Tabela B.1: Resultados Esperados

	Pré-processamento						Quantidade de Registros	
	IF	SC	SW	ER	EC	TC	Universo	Esperado
1		✓			✓	✓	531.099	19
2		✓			✓	✓	531.099	531.099
3		✓		✓	✓	✓	488.922	3.217
4		✓		✓	✓	✓	672.001	33.506
5		✓		✓	✓	✓	672.001	0
6		✓		✓	✓	✓	488.922	18.358
7		✓		✓	✓	✓	4.278.066	1.309
8	✓	✓		✓	✓	✓	1.623.938	7
9		✓		✓	✓	✓	488.922	83
10	✓	✓		✓	✓	✓	531.099	53
11	✓	✓		✓	✓	✓	1.623.938	0
12	✓	✓			✓	✓	488.922	0
13	✓	✓	✓	✓	✓	✓	1.623.938	436
14	✓	✓		✓	✓	✓	488.922	23
15		✓	✓		✓	✓	1.623.938	5.368
16	✓	✓			✓	✓	488.922	0
17		✓	✓	✓	✓	✓	4.278.066	9.791
18		✓	✓	✓	✓	✓	1.623.938	203
19	✓	✓	✓		✓	✓	4.278.066	22
20		✓	✓	✓	✓	✓	488.922	166.236
21	✓						488.922	0
22		✓	✓		✓	✓	28	2
23		✓	✓	✓	✓	✓	488.922	299
24	✓	✓	✓	✓	✓	✓	28	27

Tabela B.1: Resultados Esperados

	Pré-processamento						Quantidade de Registros	
	IF	SC	SW	ER	EC	TC	Universo	Esperado
25	✓	✓	✓	✓	✓	✓	4.278.066	229
26		✓	✓	✓	✓	✓	531.099	195.573
27	✓	✓	✓	✓	✓	✓	1.623.938	82
28		✓	✓		✓	✓	672.001	0
29	✓	✓	✓	✓	✓	✓	531.099	3
30	✓	✓	✓		✓	✓	488.922	5
31		✓	✓		✓	✓	4.278.066	10
32	✓	✓	✓				562.366	6
33	✓	✓	✓		✓	✓	2.155.037	11
34	✓	✓	✓		✓	✓	2.155.037	0
35	✓	✓		✓	✓	✓	488.922	66
36	✓	✓			✓	✓	488.922	1
37	✓	✓	✓	✓	✓	✓	672.001	27
38	✓	✓	✓	✓	✓	✓	48	48
39		✓	✓	✓	✓	✓	672.001	85.657
40		✓		✓	✓	✓	488.922	1
41	✓	✓		✓	✓	✓	2	2
42	✓	✓	✓	✓	✓	✓	488.922	92
43		✓	✓	✓	✓	✓	672.001	1.176
44		✓		✓	✓	✓	672.001	4.712
45	✓	✓		✓	✓	✓	488.922	0
46	✓	✓		✓	✓	✓	1.623.938	17
47		✓	✓	✓	✓	✓	1.778.104	8
48	✓	✓		✓	✓	✓	488.922	0
49	✓	✓		✓	✓	✓	531.099	8
50		✓		✓	✓	✓	531.099	0

APÊNDICE C

Resultados Encontrados

Tabela C.1: Resultados Encontrados

	Pré-Processamento Realizado	Execução		Quantidade de Mapeamentos	Selecionados		Não Selecionados		*precision*	*recall*	*f-measure*
		Tipo	Tempo(s)		A	B	C	D			
1	SC, EC, TC	NRE	1,1	0	0	0	19	531.080	-	-	-
2	SC, EC, TC	ESR	3.108	3.128	0	0	531.099	0	-	-	-
3	SC, EC, TC	NRE	1,2	0	0	0	3.217	485.705	-	-	-
4	SC, ER, EC, TC	ESR	189	14	0	150	33.506	638.345	-	-	-
5	SC, ER, EC, TC	ESR	2.210	1.803	0	124	0	671.877	-	-	-
6	SC, ER, EC, TC	ECR	2.285	2.644	110	17.960	18.248	452.604	0,00608	0,00599	0,02396
7	SC, ER, EC, TC	ESR	1.123	1.567	0	0	1.309	4.276.757	-	-	-
8	IF, SC, ER, EC, TC	NEB	0	0	0	0	7	1.623.931	-	-	-
9	SC, ER, EC, TC	NEB	0	0	0	0	83	488.839	-	-	-
10	IF, SC	NEB	0	0	0	0	53	531.046	-	-	-
11	IF, SC, ER, EC, TC	ESR	2.265	1.803	0	0	0	1.623.938	-	-	-
12	IF, SC, EC	NEB	0	0	0	0	0	488.922	-	-	-
13	IF, SC, SW, ER, EC, TC	ESR	207	14	0	20	436	1.623.482	-	-	-
14	SC, ER, EC, TC	ESR	33.556	4.559	0	17	23	488.882	-	-	-
15	SC, SW, EC, TC	NEB	0	0	0	0	5.368	1.618.570	-	-	-
16	IF, SC, EC, TC	ESR	2.363	1.803	0	6	0	488.916	-	-	-
17	SC, SW, ER, EC, TC	ESR	2.127	58	0	1.037.456	9.791	3.230.819	-	-	-
18	SC, SW, ER, EC, TC	INT	0	0	0	0	203	1.623.735	-	-	-
19	IF, SC, SW, EC, TC	ECR	84	23	22	0	0	4.278.044	1	1	1

Tabela C.1: Resultados Encontrados

	Pré-Processamento Realizado	Execução		Quantidade de Mapeamentos	Selecionados		Não Selecionados		*precision*	*recall*	*f-measure*
		Tipo	Tempo(s)		A	B	C	D			
20	SC, SW, ER, EC, TC	ESR	151	28	0	830	166.236	321.856	-	-	-
21	IF	NRE	1	0	0	0	0	488.922	-	-	-
22	SC, SW, EC, TC	NEB	0	0	0	0	2	26	-	-	-
23	SC,SW, ER, EC, TC	ECR	182	14	35	0	264	488.623	1	0,11705	0,20956
24	IF, SC, SW, ER, EC, TC,	INT	0	0	0	0	27	1	-	-	-
25	IF, SC, SW, ER, EC, TC	ECR	51	23	229	0	0	4.277.837	1	1	1
26	SC, SW, ER, EC, TC	INT	0	0	0	0	195.573	335.526	-	-	-
27	IF, SC, SW, ER, EC, TC	ESR	183	14	0	0	82	1.623.856	-	-	-
28	SC, SW, EC, TC	INT	0	0	0	0	0	672.001	-	-	-
29	IF, SC, SW, EC, TC	NEB	0	0	0	0	3	531.096	-	-	-
30	IF, SC, SW, EC, TC	NEB	0	0	0	0	5	488.917	-	-	-
31	SC, SW, EC, TC	INT	0	0	0	0	10	4.278.056	-	-	-
32	IF	NRE	1	0	0	0	6	562.360	-	-	-
33	IF, SC, SW, EC, TC	NEB	0	0	0	0	11	2.155.026	-	-	-
34	IF, SC, SW, EC, TC	INT	0	0	0	0	0	2.155.037	-	-	-
35	SC, ER, EC, TC	ESR	130	14	0	7	66	488.849	-	-	-
36	IF, SC, EC, TC	INT	0	0	0	0	1	488.921	-	-	-
37	SC, SW, ER, EC, TC	ECR	4.613	8	27	671.974	0	0	0,00004	1	0,00003
38	IF, SC, SW, ER, EC, TC	INT	0	0	0	0	48	0	-	-	-

Tabela C.1: Resultados Encontrados

	Pré-Processamento Realizado	Execução		Quantidade de Mapeamentos	Selecionados		Não Selecionados		*precision*	*recall*	*f-measure*
		Tipo	Tempo(s)		A	B	C	D			
39	SC, SW, ER, EC, TC	ESR	705	54	0	59	85.657	586.285	-	-	-
40	SC, ER, EC, TC	NRE	1	0	0	0	1	488.921	-	-	-
41	IF, SC, ER, EC, TC	ECR	9.359	8	2	0	0	0	1	1	1
42	IF, SC, SW, ER, EC, TC	ECR	195	14	92	0	0	488.827	1	1	1
43	SC, SW, ER, EC, TC	INT	0	0	0	0	1.176	670.825	-	-	-
44	SC, SW, ER, EC, TC	ECR	5.143	4.559	1.569	3.056	3.143	664.233	0,33924	0,25310	0,01448
45	IF, SC, ER, EC, TC	ESR	211	14	0	0	0	488.922	-	-	-
46	IF, SC, ER, EC	NEB	0	0	0	0	17	1.623.921	-	-	-
47	SC, SW, EC, TC	NEB	0	0	0	0	8	1.778.096	-	-	-
48	IF, SC, SW, ER, EC, TC	INT	0	0	0	0	0	488.922	-	-	-
49	IF, SC, ER, EC, TC	ESR	2.860	1.803	0	0	8	531.091	-	-	-
50	SC, ER, EC, TC	ESR	2.492	1803	0	4.636	0	526.463	-	-	-

ANEXO I

Configuração do Banco de Dados

Código I.1 Código SQL para criação da tabela filme do IMDB

```
CREATE DATABASE imdb_pt;
CREATE TABLE IF NOT EXISTS imdb_pt.filme (
  filmeId VARCHAR(12) NOT NULL,
  tipo VARCHAR(255) NULL,
  titulo VARCHAR(255) NULL,
  anoInicio VARCHAR(4) NULL,
  anoFim VARCHAR(4) NULL,
  avaliacao DECIMAL(5,4) NULL,
  PRIMARY KEY (filmeId));
```

Código I.2 Código SQL para criação da tabela pessoa do IMDB

```
CREATE TABLE IF NOT EXISTS imdb_pt.pessoa (
  pessoaId VARCHAR(12) NOT NULL,
  nome VARCHAR(255) NULL,
  anoNascimento VARCHAR(4) NULL,
  anoMorte VARCHAR(4) NULL,
  PRIMARY KEY (pessoaId));
```

Código I.3 Código SQL para criação da tabela genero do IMDB

```
CREATE TABLE IF NOT EXISTS imdb_pt.genero (
  generoId INT NOT NULL,
  nome VARCHAR(25) NULL,
  PRIMARY KEY (generoId));
```

Código I.4 Código SQL para criação da tabela episodio do IMDB

```
CREATE TABLE IF NOT EXISTS imdb_pt.episodio (
  episodioId VARCHAR(12) NOT NULL,
  FilmeId VARCHAR(12) NULL,
  numeroTemporada INT NULL,
  numeroEpisodio INT NULL,
  PRIMARY KEY (episodioId),
  CONSTRAINT fk_episodio_filme
    FOREIGN KEY (FilmeId) REFERENCES imdb_pt.filme (filmeId));
```

Código I.5 Código SQL para criação da tabela ator_filme do IMDB

```
CREATE TABLE IF NOT EXISTS imdb_pt.ator_filme (
  filmeId VARCHAR(12) NOT NULL,
  pessoaId VARCHAR(12) NOT NULL,
  CONSTRAINT fk_atorFilme_pessoa
    FOREIGN KEY (pessoaId) REFERENCES imdb_pt.pessoa (pessoaId),
  CONSTRAINT fk_atorFilme_filme
    FOREIGN KEY (filmeId) REFERENCES imdb_pt.filme (filmeId));
```

Código I.6 Código SQL para criação da tabela diretor_filme do IMDB

```
CREATE TABLE IF NOT EXISTS imdb_pt.diretor_filme (
  filmeId VARCHAR(12) NOT NULL,
  pessoaId VARCHAR(12) NOT NULL,
  CONSTRAINT fk_diretorFilme_filme
    FOREIGN KEY (filmeId) REFERENCES imdb_pt.filme (filmeId),
  CONSTRAINT fk_diretorFilme_pessoa
    FOREIGN KEY (pessoaId) REFERENCES imdb_pt.pessoa (pessoaId));
```

Código I.7 Código SQL para criação da tabela genero_filme do IMDB

```
CREATE TABLE IF NOT EXISTS imdb_pt.genero_filme (
  filmeId VARCHAR(12) NOT NULL,
  generoId INT NOT NULL,
  CONSTRAINT fk_generoFilme_filme
    FOREIGN KEY (filmeId) REFERENCES imdb_pt.filme (filmeId),
  CONSTRAINT fk_generoFilme_genero
    FOREIGN KEY (generoId) REFERENCES imdb_pt.genero (generoId));
```

Código I.8 Código SQL para criação da tabela TME

```
CREATE DATABASE repositorio;
CREATE TABLE repositorio.tme (
  serial int NOT NULL AUTO_INCREMENT,
  provider varchar(255) DEFAULT NULL,
  url varchar(255) DEFAULT NULL,
  email varchar(255) DEFAULT NULL,
  oai_set varchar(255) DEFAULT NULL,
  dispquery enum('false','true') NOT NULL,
  dc_title varchar(255) DEFAULT NULL,
  dc_creator text,
  dc_subject varchar(255) DEFAULT NULL,
  dc_description text,
  dc_contributor varchar(255) DEFAULT NULL,
  dc_publisher varchar(255) DEFAULT NULL,
  dc_date date DEFAULT NULL,
  dc_type varchar(255) DEFAULT NULL,
  dc_format varchar(255) DEFAULT NULL,
  dc_identifier varchar(255) DEFAULT NULL,
  dc_source varchar(255) DEFAULT NULL,
  dc_language varchar(255) DEFAULT NULL,
  dc_relation varchar(255) DEFAULT NULL,
  dc_coverage varchar(255) DEFAULT NULL,
  dc_rights varchar(255) DEFAULT NULL,
  loginbd varchar(15) NOT NULL,
  passwordbd varchar(15) NOT NULL,
  PRIMARY KEY (serial)
```

Código I.9 Código SQL para inserção do IMDB na tabela TME

```
INSERT INTO repositorio.tme VALUES
(1,
'http://localhost/website/system/database/oai',
'http://localhost/website/system/',
'name@domain.com',
'http://localhost/website/system/systemname',
'true',
'imdb_pt',
'IMDB Inc',
'filmes, séries, atores, papéis, diretores, gêneros e episódio.',
'Contem registros dos filmes, séries de TV e episódios, os atores e diretores
de cada filme, os diversos gêneros de cada filme bem como sua avaliação.',
' ',
'Amazon',
'2020-08-28',
'Database- MySQL',
'SGBDR',
'imdb_pt',
' ',
'por',
' ',
' ',
' ',
'root',
'1234');
```