UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

ALINE DAYANY DE LEMOS

Avaliação Semântica de Consultas por Palavras-Chave a Bancos de Dados Relacionais

Goiânia
2020

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

Aline Dayany de Lemos

3. Título do trabalho

Avaliação Semântica de Consultas por Palavras-chave a Bancos de Dados Relacionais

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 **João Carlos da Silva**, **Usuário Externo**, em 06/10/2020, às 13:44, 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 **ALINE DAYANY DE LEMOS**, **Discente**, em 06/10/2020, às 14:31, 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 **1599032** e o código CRC **10391FF6**.

Referência: Processo nº 23070.032939/2020-67 SEI nº 1599032

ALINE DAYANY DE LEMOS

Avaliação Semântica de Consultas por Palavras-Chave a Bancos de Dados Relacionais

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 Programa de Pós-Graduação em Ciência da Computação.

Área de concentração: Ciência da Computação.

Orientador: Prof. Dr. João Carlos da Silva

Goiânia
2020

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

Lemos, Aline Dayany de
Avaliação Semântica de Consultas por Palavras-Chave a Bancos de Dados Relacionais [manuscrito] / Aline Dayany de Lemos. - 2020.
CXLIII, 143 f.

Orientador: Prof. Dr. João Calos da Silva.
Dissertação (Mestrado) - Universidade Federal de Goiás, Instituto de Informática (INF), Programa de Pós-Graduação em Ciência da Computação, Goiânia, 2020.
Bibliografia. Anexos. Apêndice.
Inclui tabelas, algoritmos, lista de figuras, lista de tabelas.

1. consultas por palavras-chave. 2. bancos de dados relacionais. 3. Avaliação padronizada. I. da Silva, João Calos, orient. II. Título.

CDU 004

UNIVERSIDADE FEDERAL DE GOIÁS

INSTITUTO DE INFORMÁTICA

ATA DE DEFESA DE DISSERTAÇÃO

Ata nº **19/2020** da sessão de Defesa de Dissertação de **Aline Dayany de Lemos**, que confere o título de Mestra em Ciência da Computação, na área de concentração em Ciência da Computação.

Aos vinte e quatro dias do mês de agosto de dois mil e vinte, a partir das dez horas, via sistema de webconferência da RNP, realizou-se a sessão pública de Defesa de Dissertação intitulada **"**Avaliação Semântica de Consultas por Palavras-chave a Bancos de Dados Relacionais**"**. 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 Fábio Nogueira de Lucena (INF/UFG), membro titular externo; e Professor Doutor Leonardo Andrade Ribeiro (INF/UFG), membro titular interno. Durante a arguição os membros da banca não fizeram sugestão de alteração do título do trabalho. 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. A Banca Examinadora reuniu-se em sessão secreta a fim de concluir o julgamento da Dissertação, tendo sido a candidata **aprovada** 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 quatro dias do mês de agosto de dois mil e vinte.

TÍTULO SUGERIDO PELA BANCA

Documento assinado eletronicamente por **João Carlos da Silva**, **Usuário Externo**, em 24/08/2020, às 13:14, 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 **ALINE DAYANY DE LEMOS**, **Usuário Externo**, em 24/08/2020, às 13:59, 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 24/08/2020, às 16:18, 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 **Leonardo Andrade Ribeiro**, **Professor do Magistério Superior**, em 24/08/2020, às 16:31, 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 **1458493** e o código CRC **4D818893**.

Referência: Processo nº 23070.032939/2020-67 SEI nº 1458493

Aline Dayany de Lemos

Graduou-se em Ciência da Computação na FAA - Faculdade Anhanguera de Anápolis, possui Pós Graduação em Desenvolvimento de Sistemas Web pela FAA. Atualmente é docente no Centro Universitário de Anápolis - Uni-Evangélica onde trabalha com pesquisas de Bancos de Dados voltados a área de mineração de dados, também atua com desenvolvimento em *Python*.

Dedico este trabalho aos meus pais Antônio e Maria e a minha filha Melissa.

Agradecimentos

Agradeço a Deus em primeiro lugar, por tudo.

Aos meus pais Antônio e Maria pelo incentivo constante aos estudos e suporte, por cuidarem da Melissa melhor que eu. Por vocês fui incentivada e orientada desde cedo que a educação seria o único meio de ter sucesso na vida. Além disso, por repetirem constantemente aquele ditado popular "educação é a única coisa que ninguém te tira". E obrigado por não facilitarem nada durante o percurso. Aos meus irmãos Luiz e Viviane, cunhado Francisco e cunhada Eliane, meus sobrinhos Luiza, Henrique, Rafael, Marcos e Mariana que souberam entender a minha ausência. A minha vó materna D. Lourdes e madrinha Marly, as orações de vocês me sustentaram durante a caminhada, obrigado pelo privilégio de conviver e aprender tanto. Agradeço pela sabedoria e paciência da minha florzinha, minha filha linda Melissa, que apesar de externar a insatisfação pela minha ausência e por eu estar sempre "fazendo a dissertação"me incentivava a estudar e a finalizar o processo. Melissa quero ser exemplo pra você.

Ao irmão que tive a oportunidade de escolher, sem laços sanguíneos especial Luiz Fernando Borges, o amigo para todas as horas. Ao Osilmar Mendonça incentivador inicial e que por motivos diversos não comemora comigo hoje, mas essa vitória também é sua. Aos colegas do laboratório 254, em especial Altino Dantas pelo suporte e auxílio de todas as horas, as correções dos documentos e por conseguir me dar clareza e entendimento do que precisava ser feito. Aos meus alunos - não vou citar nomes para não ficarem com ciúmes, que no fundo todo conhecimento será redirecionado para vocês.

Por fim, ao Instituto de Informática da UFG e aos diversos professores que compartilharam comigo seu conhecimento e em especial ao Professor João Carlos que sempre foi paciente e atencioso comigo.

Nossa maior fraqueza está em desistir. O caminho mais certo de vencer é tentar mais de uma vez

Thomas Edson.

Resumo

Lemos, Aline Dayany de. **Avaliação Semântica de Consultas por Palavras-Chave a Bancos de Dados Relacionais**. Goiânia, 2020. 143p. Dissertação de Mestrado. Instituto de Informática, Universidade Federal de Goiás.

As técnicas de consulta por palavras-chave a bancos de dados relacionais têm sido desenvolvidas e aprimoradas nos últimos anos e para tanto é fundamental existir meios de avaliação destes métodos. O presente estudo tem como proposta, além de mensurar a eficácia das técnicas, verificar o processamento semântico realizado em relação às consultas durante sua submissão nas técnicas de pesquisa. Para isso, este estudo considera as características individuais de cada técnica para que a avaliação da eficácia não seja prejudicada, e propõe grupos de consultas para a avaliação de diferentes características semânticas. O conjunto de palavras-chave proposto é construído pela composição provenientes de *sites* específicos e *tuítes* que simulam consultas comuns às bases de dados contempladas. Espera-se compreender o processamento das palavras-chave, e se a eficácia do resultado é prejudicado em razão do processamento ou da falta dele.

Palavras–chave

Consulta por palavras-chave, semântica, avaliação padronizada

Abstract

Lemos, Aline Dayany de. **Semantic Evaluation Keyword Search to Relational Databases**. Goiânia, 2020. 143p. MSc. Dissertation. Instituto de Informática, Universidade Federal de Goiás.

The querying techniques by keywords to relational databases have been developed and improved in recent years and it is therefore essential to have means of evaluating these methods. This study proposes to, in addition to measuring the effectiveness of the techniques, verify the semantic processing performed in relation to the consultations during their submission in the research techniques. For this, the proposal considers the individual characteristics of each technique so that the evaluation of effectiveness is not impaired, and proposes groups of consultations for the evaluation of different semantic characteristics. The proposed set of keywords is built by the composition of keywords from specific sites and tweets that simulate common queries to the contemplated databases. It is expected to understand the processing of keywords and whether the result's effectiveness is impaired due to the processing or lack of it.

Keywords

Keyword search, semantic, standard evaluation

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Códigos de Programas

CAPÍTULO 1

Introdução

A internet se consolidou como principal meio de difusão de conhecimento, e, associada a sua utilização, uma elevada quantidade de dados é gerada diariamente [35, 38, 26]. Esses dados são armazenados em vários formatos em ambientes estruturados, semi-estruturados e não estruturados e uma grande quantidade desses dados são armazenados em bancos de dados relacionais[16]. Os mesmos usuários que geram e armazenam estes dados também aspiram acessá-los ou que outras pessoas os acessem.

Os usuários habitualmente não sabem onde ou como encontrar os dados dispostos na *web*. Logo, na tentativa de encontrá-los, motores de busca são utilizados, tais como *Google*[1], *Bing*[2] ou *Yahoo*[3], que buscam em textos ou títulos de arquivos. Apesar de tais motores apresentarem retorno significativo para algumas consultas, boa parte dos dados não ficam visível na *web*, pois podem estar dispostos em ambientes que esses motores de busca não são capazes de encontrá-los. Alguns desses dados que são disponibilizados em bancos de dados relacionais e necessitam de *interfaces* e conhecimentos específicos para a visualização destas informações [35, 38, 31, 34].

Para visualizar estes bancos de dados é necessário um software capaz de acessar e apresentar estes dados, conforme solicitado. Se forem utilizadas interfaces de conexão com o banco de dados será necessário por parte do usuário conhecimento em Linguagem de consulta estruturada - *SQL*. E este conhecimento será para além da linguagem, mas também da construção estrutural e relacional do banco de dados [35, 31].

Para reduzir a complexidade da operação e, com o passar dos anos, interfaces de técnicas de pesquisa por palavras-chave a bancos de dados relacionais foram desenvolvidos. Essas técnicas realizam buscas no banco de dados indicado e apresenta os resultados encontrados, conforme as palavras-chave submetidas. O banco de dados é mapeada como um grafo, de dados ou de esquema, e depois que a consulta é submetida a consulta os devidos processamentos são realizado e as consultas ranqueadas. Para os registros encon-

[1] www.google.com
[2] www.bing.com
[3] https://br.yahoo.com/

trados, dados ou comando em *SQL* são apresentados ao usuário a fim de demonstrar os resultados encontrados [21, 35].

Os estudos de cada técnica apresentam um experimento sobre seu funcionamento em relação a um banco de dados previamente definido. Nestes experimentos o banco de dados e as palavras-chave são escolhidas e submetidas em forma de consulta para validar o estudo e demostrar a técnica apresentada. Um estudo deveria selecionar o banco de dados e as palavras-chave baseados em critérios científicos e que fosse possível, posteriormente, replicar os experimentos e que fosse possível encontrar os mesmos resultados. A Tabela 1.1 lista as técnicas e as bancos de dados utilizadas para realizar os experimentos.

Tabela 1.1: *Técnicas e bancos de dados utilizadas.*

Técnica	Bancos de dados
DISCOVER [25]	*TPC-H*
BANKS [1]	*DBLP*
DBXplorer [3]	*TPC-H* e *IMDB*
Efficient [24]	*DBLP* e *IMDB*
Banks-II [27]	*DBLP* e *IMDB*
DPBF [18]	*DBLP* e *Mondial*
BLINKS [23]	*DBLP* e *IMDB*
SPARK [30]	*DBLP*, *IMDB* e *Mondial*
FIRSK [33]	*IMDB* e *Northwid*
Keymantic [5]	*IMDB* e *Tourist Database*
DBSemSXplorer [20]	*IMDB*
QUEST [7]	*DBLP*, *IMDB* e *Mondial*
MeanKs [28]	*TPC-E*
Ghanbarpour [21]	*DBLP*, *IMDB* e *Mondial*
Ramada et al.[35]	*Company Database*

Percebe-se que os bancos de dados selecionados para o experimento são similares, mas as palavras-chave são criadas pelos próprios autores, exceto o estudo proposto por *Ghanbarpour* [21] que utiliza uma avaliação padronizada. Ferramentas de avaliação padronizada, ou *benchmarks* poderia ser utilizado nos estudos. A utilização de um *benchmark* existente em estudos científicos por demonstrar o rigor cientifico. Em se tratando de pesquisa por palavras-chave a banco de dados relacionais pode ser observado que, apesar de várias pesquisas explorando diversas técnicas diferentes, não há avanços significativos. Além disso, a ausência do emprego comercial das técnicas pesquisa por palavras-chave a bancos de dados relacionais é outro indicativo sobre o nível de maturidade alcançado até o presente momento [6].

Quando algumas destas técnicas foram executadas em um ambiente controlado e previamente definido, com dados e recursos computacionais disponíveis de forma igualitária, os resultados não refletiram nas avaliações de seus autores. As avaliações

existentes [14, 16, 32] demonstram que, apesar dos esforços, as técnicas de pesquisa por palavras-chave a bancos de dados relacionais ainda são insatisfatórias para o que se propõe.

Tratando-se dos *benchmarks* existente e que os bancos de dados *IMDB*, *Mondial* ou *DBLP*, em combinação ou individual estão presente em 12 das 15 técnicas. As bancos de dados utilizadas no *benchmark* [14, 16] foram escolhidas por serem populares, pequenas em relação ao número de registros e pela possibilidade de serem frequentemente utilizadas em sistemas de avaliação. Enquanto no *benchmark* [32], é ressaltada a importância de uma heterogeneidade das bancos seja por quantidade de tabelas, relações ou dados [14, 16, 32].

Ao realizar o processamento da consulta submetida, um grafo de dados ou grafo de esquema é gerado, e apresenta os resultados encontrados. E uma possível comparação, seja, do processamento ou da apresentação de resultados de técnicas que implementam grafo diferentes, é arbitrário. As necessidades particulares de cada técnica precisam ser consideradas em uma avaliação de eficácia e/ou eficiência [14, 16].

A eficácia e a eficiência são termos constantemente citados e, para dirimir quaisquer dúvidas, serão consideradas no seguinte contexto: A eficácia é a capacidade da técnica responder com precisão a consulta submetida. Já eficiência trata do tempo gasto para a execução da consulta, desde a submissão até a apresentação do resultado final [15, 16, 6, 32].

Nas técnicas avaliadas neste trabalho as consultas são apresentadas como conjunto de palavras-chave para cada banco de dados e estas consultas foram definidas por seus respectivos autores, mas não há definição técnica ou científica clara que justifique a escolha. Após o retorno da submissão da consulta, para cada consulta a relevância é definida através da percepção pessoal e unilateral dos pesquisadores [14, 16, 32].

Uma estudo recente [6] discute direções significativas que relaciona a área de consulta por palavras-chave a bancos de dados relacionais à outras áreas de pesquisas relacionadas a bancos de dados. O estudo salienta a importância de avaliar detalhes das técnicas de consulta e não somente a eficiência ou eficácia. Ou seja, não somente se ao submeter uma consulta, os resultados retornados serão exatamente os resultados esperados. Por fim, as atuais abordagens de avaliação de técnicas de pesquisa por palavras-chave a bancos de dados relacionais apenas verificam a eficiência ou eficácia dos resultados, verificando se os dados retornados são aqueles estabelecidos como esperados, e mensurando o quão assertivo é esse conjunto.

Mas é preciso abranger as verificações e avaliações destes sistemas de pesquisa, e tentar descobrir quais palavras ou sub palavras afetou o desempenho da técnica ou se a técnica não foi projetada para uma busca mais refinada. E se a técnica não foi projetada para uma busca refinada estabelecer a partir de então, vertentes que podem e precisam ser

exploradas para que consigamos chegar ao estado da arte.

1.1 Objetivos

Este trabalho tem como objetivo principal a proposição de um método para avaliar as técnicas de pesquisa por palavras-chave a bancos de dados relacionais em relação à eficácia e o tratamento semântico das palavras-chave.

Para atingir o objetivo supracitado, os seguintes objetivos específicos foram definidos:

- Identificar e selecionar as técnicas de pesquisa com palavras-chave a bancos de dados relacionais;
- Definir conjuntos de consultas que apresentem impacto semântico na submissão de consultas por palavras-chave a bancos de dados relacionais, assim como a semântica das consultas e os resultados esperados;
- Selecionar métricas para definição de critérios de avaliação das consultas a partir das palavras-chave;
- Configurar um ambiente de experimentação, realizar as avaliações propostas e apresentar os resultados.

1.2 Metodologia

Para conseguir atingir os objetivos, este estudo apresenta uma pesquisa bibliográfica para delinear todos os conteúdos teóricos a serem verificados, e trata também de uma pesquisa exploratória que busca elucidar e analisar os resultados construídos [42]. Ademais, os objetivos específicos forão atingidos da seguinte forma:

Identificar e selecionar as técnicas de pesquisa com palavras-chaves a bancos de dados relacionais: estudos bibliográficos que discriminar as técnicas e identificar características relevantes. As técnicas foram separadas quanto ao tipo de grafo implementado. Após a identificação das técnicas, buscou-se implementações disponíveis para execução.

Definir conjuntos de consultas que apresentem impacto semântico na submissão de consultas por palavras-chave a bancos de dados relacionais, assim como a semântica das consultas e os resultados esperados: O conjunto de palavras-chave foi construído baseado em históricos de pesquisas de sites que tratem do contexto da banco de dados ou mensagens de tuítes com a utilização de *hastags* vinculadas ao contexto. Para cada consulta foi definido uma semântica, isto é, o que se espera como resultado de cada

técnica em relação à consulta submetida. E foi construído os conjuntos de resultados esperados através de análises e buscas manuais nas bancos de dados escolhidas levando em consideração a semântica previamente definida.

Selecionar métricas para definição de critérios de avaliação das consultas a partir das palavras-chave: Identificar as métricas utilizadas nas avaliações atuais de pesquisa por palavras-chave a bancos de dados relacionais.

Configurar um ambiente de experimentação, realizar as avaliações propostas e apresentar os resultados: O ambiente foi configurado para que atendesse as necessidades de cada metodologia para a correta execução. Após a imputação do esquema e do dados nos sistemas gerenciadores de bancos de dados, as execuções de cada conjunto de palavras-chave foi iniciado, e comparado o conjunto de resultados retornados e o conjunto de resultados esperados. Aplicar as métricas selecionadas a cada conjunto de palavras-chave analisando os resultados retornados em comparação aos resultados esperados.

1.3 Organização do Texto

Este documento segue organizado da seguinte forma:

No Capítulo 2, os conceitos teóricos referente as técnicas de pesquisa por palavras-chave a bancos de dados relacionais e aos *benchmarks* são apresentados. Bem como as especificações para um *benchmark* e as métricas mais utilizadas na área de consulta por palavras-chave.

No Capítulo 3, serão apresentados os trabalhos relacionados a esta pesquisa os *benchmarks* existentes e atuais, e suas características individuais.

O Capítulo 4 apresenta o detalhamento do *benchmark* proposto, quais seus componentes e a justificativa da escolha de cada um deles e traça um paralelo geral apresentado as características dos *benchmarks* atuais.

O Capítulo 5, é apresentado inicialmente a metodologia utilizada para a construção das consultas e a definição da semântica. Posteriormente é abordado o contexto dos experimentos em relação às técnicas e aos bancos de dados utilizadas, e como sucedeu a verificação e processamento das características semânticas desejadas para cada consulta. Este Capítulo apresenta ainda uma análise dos resultados obtidos e tabulados em perspectivas diferente - (1) na perspectiva das técnicas, (2) na perspectiva dos bancos de dados e por fim (3) na perspectiva das características semânticas.

O Capítulo 6 apresenta as observações gerais do trabalho, e lista as contribuições, limitações e trabalhos futuros.

CAPÍTULO 2

Fundamentação Teórica

Este Capítulo discute os conceitos teóricos principais que darão fundamentação para da seções subsequentes. A Seção 2.1 apresenta as definições das técnicas de consulta por palavras-chave a banco de dados relacionais e apresenta um exemplo sobre o mapeamento interno das técnicas em relação ao grafo gerado. Na Seção 2.2 é apresentado um breve resumo histórico dos *benchmarks* para bancos de dados relacionais e uma definição teórica é construída. A Seção 2.3 apresenta uma proposta de arquitetura dos elementos que precisam ser considerados para a construção de um *benchmark* de consulta por palavra-chaves e utilizando *benchmarks* já existentes evidencia a proposta construída. E por fim a Seção 2.4 as métricas que são comumente utilizadas e que serão utilizadas neste estudo.

2.1 Técnicas para Consultas por Palavras-chave a Bancos de Dados Relacionais

As técnicas de pesquisa com palavras-chave a bancos de dados relacionais possuem um núcleo principal que as segrega, que trata de como cada consulta manipula a entrada e apresenta o resultado. Após a submissão da consulta a técnica realiza o processamento e apresenta o resultado em formato de grafo, podendo ser grafo de dados ou esquema. Grafo é uma estrutura que demonstra os objetos de um conjunto visualmente como um nó, e a relação entre esses objetos são denominadas arestas. As técnicas que modelam os grafos baseado nos metadados do banco de dados modelam utilizando uma *Candidate Network* e as técnicas que modelam os grafos baseado nos dados do banco de dados utilizam uma *Steiner Tree* [41].

O grafo gerado por uma *Candidate Network* é um grafo de esquema direcionado G = (N, A), onde os N nós do grafo representam as informações de esquema - nome de tabelas e atributos, e o conjunto de arestas (A) que representa as relações de chave estrangeira e chave primária no banco de dados [25, 3, 4]. A geração das *Candidate Network* possuem duas etapas: *Candidate Network Generation* e *Plan Generator*. Na

etapa de *Candidate Network Generation* as consultas são mapeadas no esquema da banco de dados e são criadas as configurações que associam o nome de tabela, a coluna e o registro da tupla, obedecendo as restrições lógicas do banco de dados. Posteriormente, os grafos criadas são submetidas à etapa de *Plan Generator*, onde são verificados em relação à eficiência das associações. As melhores configurações são convertidas em linguagem *SQL* e o resultado posteriormente apresentado ao usuário [25]. São exemplos de técnicas que utilizam a estratégia de *Candidate Network*: *DISCOVER* [25], *DBXplorer* [3], *Efficient* [24], *Keymantic* [5] e Ramada et al.[35].

Por outro lado, as técnicas baseadas em *Steiner Tree* possuem como resultado um grafo de dados G = (N, A), onde os N nós do grafo representam as tuplas do banco de dados, e o conjunto de arestas (A) que conecta os N nós. Com isso, sendo direcionados ou não, representa a restrição lógica do banco de dados, definida pela relação chave primária e chave estrangeira. Um grafo é gerado a partir de um determinado nó e as arestas que o conectam formam um subgrafo. Para cada consulta, um subgrafo é gerado na tentativa de encontrar o subgrafo de menor custo. O grafo de dados gerado é construído na memória de acesso aleatório (*RAM*) da máquina [29]. São exemplos de técnicas que utilizam a estratégia de *Steiner Tree*: *BANKS* [1], *Bidirectional* [27], *DPBF* [18], *BLINKS* [23], *DBSemSXplorer* [20] e *Ghanbarpour* [21].

A técnica *BANKS* [1] é uma das técnicas que implementa um grafo de dados direcionado e possui acesso prévio aos dados. Cada tupla do banco de dados é modelada como um nó e a chave estrangeira é modelada como aresta direcionada do grafo, esse direcionamento será em relação à chave primária. Ao receber uma consulta, a técnica inicialmente identifica todas as tuplas que serão mapeadas como nós. Então, um subgrafo é construído baseado nas relações existentes. As buscas são realizadas utilizando o algoritmo *Backward Expanding* proposto em [10]. A implementação do algoritmo constrói um grafo baseado no caminho mínimo que, dado uma consulta para cada palavra-chave é verificado o conjunto de tuplas no qual esteja presente. O subgrafo gerado é percorrido em busca do nó com o maior número de palavras da consulta submetida, então este nó é denominado nó central. O nó central é uma folha e a partir dele são gerados subgrafos enraizados percorridos de forma ascendente. Uma consulta com muitas palavras pode gerar um número elevado de subgrafos o que acaba dificultando os resultados visto que o tempo de processamento e a memória são diretamente afetados.

BANKS-II é uma versão do *BANKS* com a implementação do algoritmo *Bidirectional Expanding* [27]. O algoritmo possui como proposta a construção do subgrafo realizando a busca considerando que o nó eleito será a folha e a raiz. Desta forma, as buscas acontecem em dois sentidos. O algoritmo também propõe de maneira diferente a importância do nó, passando a considerar o peso das arestas individualmente - denominado *edge-score*. Após, há uma nova combinação dos pesos dos nós individualmente

- denominado *node-prestige-score* e, por fim, ocorre a combinação das duas estratégias. As expansões dos subgrafos acontecem para os caminhos com menor ramificação preferencialmente. Portanto, ao iniciar uma pesquisa e encontrar o nó com o maior número de palavras em relação a consulta submetida, os subgrafos são gerados e percorridos em ambos os sentidos. Na expansão, prioriza-se o grafo de menor caminho gerando um grafo mínimo. A expansão das arestas do nó com maior número de palavras-chave é feita para nós considerados raízes potenciais. O grafo resposta será dado após o ranqueamento, considerando o peso das arestas individuais e a combinação com o nó que possui o maior número de palavras-chave. Os subgrafos são gerados na memória RAM.

Keymantic [5] é uma técnica de pesquisa que implementa um grafo de esquema e não possui acesso prévio aos dados. A técnica inicia o processamento com a construção de uma matriz de pesos referente a associação entre os termos da consulta e termos do esquema da banco de dados e outra matriz para os termos que representam valor na banco de dados. Quando uma consulta com mais de uma palavra não pode ser totalmente mapeada com as informações do esquema do banco de dados, é direcionada para verificação quanto a fração da consulta ser parte dos registros do banco de dados e esta configuração é dita como mapeamento parcial.

A técnica utiliza o algoritmo húngaro[5] para gerar os melhores mapeamentos para os termos de esquema, não somente um, prezando assim pela completude, para os mapeamentos classificados como parciais mapeados diretamente para termos de valor são selecionados. Novamente, vários mapeamentos são selecionados e mesclados em razão da sua pontuação e como saída tem-se um conjunto de melhores mapeamentos, sejam eles para termos de metadados, metadados e tuplas ou somente tuplas. Os mapeamentos são transformados em configurações que geram interpretações e em conseguinte as transformadas em linguagem *SQL*.

Mapeamentos e interpretações ambíguas podem ser geradas em função das informações inicialmente submetidas pelo usuário e da estrutura lógica do banco de dados que é alvo da busca. Quando disponibilizadas as interpretações é necessária uma intervenção do usuário selecionando uma das várias interpretações possíveis apresentadas. Após a seleção da interpretação pelo usuário, uma consulta e linguagem *SQL* é gerada e os dados apresentados. A interação com o usuário é parte do processo de busca para identificar, no seu ponto de vista, qual das interpretações geradas é significativa e relevante. São apresentadas as 100 melhores interpretações e disponibilizadas ao usuário que pode escolher uma interpretação e gerar os *scripts SQL* e posteriormente gerar, outros resultados escolhendo outra interpretação [5].

Ramada et al.[35] é outra técnica que implementa o grafo de esquema. Após submetida uma consulta a técnica realiza um pré-processamento que verifica a existência de funções de agregação, ordenação. Então são incluídos sinônimos, hiperônimos, hipôni-

mos e outros termos que possam agregar semântica à consulta. Diferentemente de outras técnicas que utilizam um banco de dados por vez, a técnica identifica os bancos de dados com informações úteis para a consulta baseado nos metadados extraídos e inseridos em uma tabela de metadados (TME). Esta tabela contém informações de descrição e acesso ao banco de dados.

Após a identificação dos bancos de dados úteis um ranqueamento é construído baseado no número de vezes que as palavras da consulta aparecem na TME. Sua estrutura de mapeamento da consulta é similar a *keymantic* com aprimoramentos referente às técnicas de similaridade de *string* e considera a interdependência e a proximidade entre as palavras da consulta submetida. Desta forma, a consulta é mapeada seja para informações de metadados, metadados e tuplas ou somente tuplas, então as interpretações são geradas e construídas as consultas em linguagem *SQL* a serem executadas posteriormente. Os resultados são ranqueados baseados no quão significativo os dados retornados são para a consulta [35] .

2.1.1 Grafo de Dados e Grafo de Esquema

Para auxiliar no entendimento de como cada uma das técnicas descritas acima se comportam em relação a uma fração de dados e como os dados são modelados será descrito um exemplo didático. Considere uma fração da banco de dados original do *IMDB* demonstrada pelas tabelas a segui descritas. A Tabela *title* 2.1 apresenta as informações sobre filmes ou programas de televisão com *id*, *title* e *producer_year* .

Tabela 2.1: *Tabela title - IMDB*

id	***title***	***producer_year***
15680	*A Stranger in My Arms*	1959
79601	*But What About the Revolutionary George Pocker?*	2005

A Tabela *name* 2.2 apresenta informações sobre as pessoas que podem estar associados aos filmes apresentados, com *id* e *name*.

Tabela 2.2: *Tabela name - IMDB*

id	name
931665	*Strong, Robert*
1998817	*Nemcich, Julie*
221432	*Davidson, John*
2059868	*Green, Jarod*
1190063	*Cook, Sophie*
356378	*Goldsworthy, Robin*
418278	*Hickman, George*
1864787	*Radevski, Tony*
764445	*Planet, Troy*
1785734	*Gross, Frank*
1902447	*Truman, Anna*
2037811	*Wilder, Robert*
169370	*Chester, George*

E por fim a Tabela *cast_info* 2.3 apresenta as relações entre as duas tabelas. O Campo *id* da Tabela *name* é uma chave primária e na Tabela *cast_info* será apresentado como *person_id* se comportando como chave estrangeira e igualmente a coluna *id* da Tabela *title* e coluna *movie_id* na Tabela *cast_info*.

Tabela 2.3: *Tabela cast_info - IMDB*

id	person_id	movie_id	person_role	note	role_id
15095588	1998817	79601		*writer*	4
15772101	2059868	79601			5
17348300	1998817	79601			8
1670485	221432	15680	263	*uncredited*	1
1532545	221432	79601			3
6793946	931665	15680	263	*uncredited*	1
2663913	356378	79601	519		1
3120463	418278	15680	263	*uncredited*	1
8623686	1190063	79601	1589655		2
13616131	1864787	79601		*producer*	3
17872200	1785734	15680	263		9
5598776	76445	79601	5712		1
13955931	1902447	79601		*producer*	3
15561543	2037811	15680		*novel And Rio*	4
1262438	169370	15680	1825	*uncredited*	1

Se a consulta submetida for *"actor Julie"* haverá grafos diferentes para as técnicas baseada em grafo de dados e para aquelas que geram grafo de esquema. Para técnicas baseadas em grafo de dados será construído um grafo em razão da similaridade entre a as palavras-chave e os valores encontrados nas tuplas do banco de dados. Desta forma, as tuplas são mapeadas como nós do grafo e os relacionamentos, ou seja, associação entre chave primária e chave estrangeira, são mapeadas como arestas. Como resultado, espera-se um grafo de dados semelhante ao da Figura 2.1.

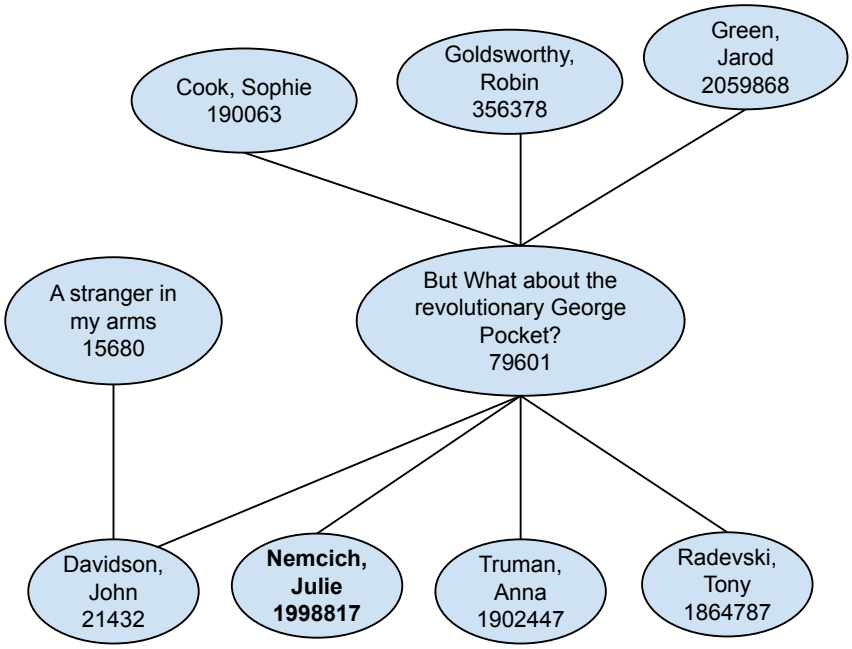

Figura 2.1: *Exemplo de Grafo de Dados.*

A palavra *"actor"* está relacionado a várias instâncias da tabela *name*, enquanto o nome *"Julie"* será encontrado somente um registro da tabela, conforme pode ser observado nas Tabelas *title* 2.1, *name* 2.2 e *cast_info* 2.3. O nó principal do grafo de dados seria o registro da Tabela *name* de *id* 1998817 - *name Nemcichh, Julie* e a partir deste registro seria criado um grafo com os dados associados a ele por chave primária e estrangeira, sendo disponibilizado o grafo abaixo. Então de acordo com a técnica de pesquisa escolhida um ou vários subgrafos seriam apresentados.

Já para as técnicas baseadas em grafo de esquema, com a mesma consulta como exemplo, *"actor Julie"* os valores da consulta são mapeados primeiro em relação aos metadados da tabela, e posteriormente verificado a similaridade das palavras-chave em relação as tuplas o banco de dados. Desta forma, o grafo gerado possui como nó, a coluna da tabela e como aresta, as associações entre chave primária e estrangeira conforme pode ser observado na Figura 2.2. A palavra "*actor*" poderia ser relacionado a Tabela *name* e coluna *name*, assim como poderia ser considerada a Tabela *title* e o campo

title, e aconteceria de forma similar com a palavra *Julie*, a similaridade é calculado individualmente por cada técnica.

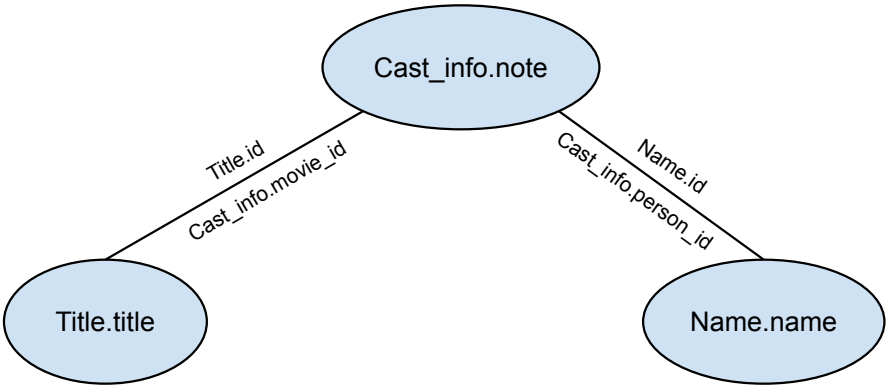

Figura 2.2: *Exemplo de Grafo de Esquema.*

Os grafos apresentados são similares e ambos seriam capazes de apresentar os registros vinculados a Tabela *name* de *id* 1998817 - *name Nemcichh, Julie*, porém é perceptível o número de nós gerados. Para o grafo de dados seriam gerados mais registros na memória *RAM* em contraste ao gerado em um grafo de esquema.

2.2 *Benchmarks*

As pesquisas científicas além de apresentar suas teorias e resultados, precisam, de algum modo, realizar comparações e avaliações com outros métodos e técnicas iguais ou similares existentes para justificar e persuadir com argumentos a utilização da teoria apresentada [22, 42]. Para realizar estes experimentos é necessário a definição de um conjunto de ações previamente estabelecidas e com objetivos claros. Este conjunto de ações precisam ser finito, imparcial e sistematicamente organizadas. A este conjunto de ações damos o nome de avaliação de referência ou *benchmark*. Para ser aceito e utilizado pela comunidade o *benchmark* precisa ser modelado com impessoalidade e rigor científico ao conjunto de ações [9, 39].

A Computação é uma das áreas que tem se beneficiado dessa ferramenta de observação. Quando uma pesquisa científica é proposta para comprovar sua eficiência ou eficácia são realizados testes com cargas de trabalho que se assemelham à realidade. Muitas vezes são escolhidos *benchmarks* consolidados na área concernente ao trabalho que se desenvolve. Existem *benchmarks* na computação para, por exemplo, comparar o desempenho de um *software* em diferentes *hardwares*, comparar o desempenho de diferentes *softwares* em um *hardware*, comparar o desempenho de diferentes *hardwares* em uma família compatível ou diferentes versões de um *software* em um *hardware* [9, 39, 11].

Estas avaliações podem também identificar tópicos do objeto de avaliação que precisa de atenção. Porém, uma avaliação de desempenho precisa ser específica para o domínio desejado, se métricas e cargas de trabalho inadequadas forem adotadas serão apresentados resultados distorcidos. Embora existam vários estudos não há uma métrica única que seja capaz de ser utilizada em qualquer contexto. Da mesma forma, não conseguimos determinar absolutamente que somente o desempenho do algoritmo é suficiente para comprovar que uma certa aplicação se destaca em relação às demais, ainda seria necessário verificar outros cenários [9, 22, 39].

Uma avaliação pode ser tendenciosa se analisada por apenas uma perspectiva. Se o autor propõe um algoritmo, por exemplo, se o mesmo construir sua própria avaliação, o conhecimento dos detalhes de sua implementação pode levá-lo a construir uma sequência de passos que beneficie seu estudo. Desta forma, as avaliações precisam ser padronizadas e devem comparar ações iguais em configurações diferentes em um domínio específico e de fácil compreensão [9, 39].

Na comunidade de Banco de Dados, há *benchmarks* consolidados e que são constantemente utilizados. Os *benchmarks* também são constantemente aprimorados com inserções de novos parâmetros para aumentar seu poder de avaliação [39]. Especificamente para bancos de dados relacionais, podemos identificar alguns *benchmarks* apresentados e estudados ao longo dos anos.

O *Wisconsin* foi a primeira avaliação padronizada de que se tem registro. Originalmente construído para realizar medições e comparações de desempenho de operações como seleção, projeção, junções simples e múltiplas, agregações simples e funções agregadas, exclusão e alteração. A carga de trabalho controlada não abrange os recursos críticos intencionalmente e os testes foram realizados considerando somente um usuário em ambientes reais. A banco de dados definida era sintética e seus dados estavam distribuídos em formato de número, texto e com configurações relativas a índice primário agrupado e não agrupado. Quando proposto, *Wisconsin* foi fundamental para comparação dos sistemas gerenciadores de bancos de dados que estavam sendo disponibilizados comercialmente [17, 37].

Na tentativa de solucionar problemas relativos à falta de projeção das consultas do *Wisconsin* em relação à mudança do tamanho de um banco de dados foi apresentado o *AS3AP - ANSI SQL Standard Scalable and Portable*. O *AS3AP* se propõe a comparar arquitetura e recursos diferentes de sistemas de bancos de dados relacionais em uma variada carga de trabalho. Projetado para ser escalável e portátil na comparação de vários sistemas com o mínimo de esforço humano. Os testes podem ser realizados com um único usuário que incluem consultas e atualizações, ou múltiplos usuários que inclui consultas e atualizações simultâneas com carga de trabalho mista [37, 40].

Transaction Processing Perfomance Council (TPC) é uma organização sem fins

lucrativos que concentra e disponibiliza alguns *benchmarks* para bancos de dados. Foi criada em 1988 com a intenção de equalizar as avaliações e diminuir as divergências decorrentes de avaliações feitas por clientes dos sistemas de bancos de dados de dados e pelas próprias empresas. Desde então, é possível se tornar membro da organização ou submeter o *benchmark* com apoio e colaboração mútua. Existem *benchmarks*, ativos e obsoletos que estão disponíveis para acesso, download e utilização. Os resultados das avaliações são publicados no site para apreciação da performance do sistema de banco de dados [37, 39].

Os *benchmarks* TPC avaliam o desempenho do sistema de banco de dados como um todo, são complexos em relação à documentação e a execução. Há *benchmark* para avaliação de transações em sistemas transacionais, de desempenho em sistemas de apoio a decisão, de carga de trabalho em sistemas de transações *on-line* [37, 39].

De acordo com os princípios apresentados por [39], podemos definir *benchmark* como sendo uma quadrupla, **B = (D, C, T, M)**, composta por um conjunto de dados ou *Datasets* **D**, um conjunto de configurações **C**, um conjunto de testes **T** e um conjunto de métricas **M**.

O conjunto de dados - **D** é composto por *scripts* que darão origem a estrutura e aos dados do banco de dados. Em alguns casos há também um conjunto de *scripts* de otimização. O conjunto de configurações - **C** são as definições de *hardware* e *software* que serão utilizados. O conjunto de testes - **T** é o conjunto de ações padronizadas e previamente estabelecidas que serão realizadas. No contexto de avaliação de pesquisa por palavras-chave a bancos de dados relacionais, um conjunto de palavras-chave que serão imputadas nos *softwares* de busca e são consideradas consultas. Ainda no conjunto de testes há de ser definido os resultados esperados e a semântica do resultado esperado. Por fim, o conjunto de métricas - **M** que serão impostas às respostas e como serão aplicadas [39].

Não obstante a todas as características dos *benchmarks* acima relatados, ainda não foram encontradas definições técnicas ou convenções sobre o tamanho em *bytes* que um *dataset* necessita ou a quantidade de tabelas, relações, relacionamentos etc. Não há descrição formal para a quantidade de operações necessárias a serem feitas ou se é necessário realizar ciclos de repetições ininterruptos de tarefas ou realizar pausas. Há apenas o pensamento generalizado da necessidade de simular um ambiente real [9].

2.3 *Benckmarks* para Técnicas de Consultas por palavras-chave a Bancos de Dados Relacionais

Apesar dos *benchmarks* citados na Seção 2.2 se enquadrarem na definição e todos eles apresentarem um banco de dados, um conjunto de configurações, o conjunto de testes e as métricas não é possível adotar qualquer um deles em uma avaliação qualquer, ou que não estejam nas mesmas características do que quando elaborado. Desta forma, é necessário perceber o quão restrito e específico um *benchmark* pode ser. Há na literatura pouca definição de arquitetura seja das técnicas de consultas por palavra-chaves a banco de dados relacionais que utilizam grafo de dados ou para as técnicas que utilizam grafo de esquema.

Bergamaschi [6] propõe uma arquitetura de *benchmark* que considera as especificidades de cada técnica, inclusive sobre qual grafo é implementado seja ele de dados ou de esquema. A avaliação precisa ser orientada ao sistema e ao usuário, construindo uma tarefa mais ampla e que verifique mais do que o resultado inicial e o resultado final. Fatores como os algoritmos utilizados, o tipo de grafo gerado, a classificação dos resultados impactam diretamente tanto na busca quanto no resultado da consulta.

Desta forma, é necessário construir uma arquitetura ampla. Essa arquitetura precisa avaliar além da pesquisa por palavra-chave propriamente dita outros aspectos da variabilidade e da incerteza do processo de busca. A arquitetura de avaliação apresentado na Figura 2.3 é uma proposta para sistemas de buscas com palavras-chave a bancos de dados relacionais e possui 4 módulos *Keyword Identification*, *Keyword Search*, *Result Filtering and Ranking* e *Presentation*.

O módulo *keyword identification* pretende identificar a semântica definida pelo usuário, isto é, o que exatamente ele gostaria de receber como resposta. A inserção de ontologias, propriedades de radicalização da palavra ou a expansão da consulta são ações que ao serem realizadas auxiliam a identificação da intenção do usuário. Quando um usuário submete um conjunto de palavras-chave, em sua percepção a informação desejada lhe parece algo muito simplificado. Porém, ao se analisar a banco de dados, uma palavra pode ser desde uma coluna da tabela, até várias informações que são encontradas nas tuplas de dados. A fragmentação da informação compromete o resultado final por buscar a correspondência exata da palavra-chave buscada com as informações contidas nas tuplas ou em valores de esquema da banco de dados.

O módulo *keyword search* pretende fazer a relação entre o conjunto de palavras-chave submetidas às suas possíveis e esperadas respostas. Neste módulo encontram-se as técnicas existentes que modelam a base com um grafo de dados, grafo de esquema ou de maneira híbrida. Ao buscar um conjunto de palavras-chave utilizando as técnicas que geram um grafo de dados, quando a busca é realizada diretamente nas tuplas da base,

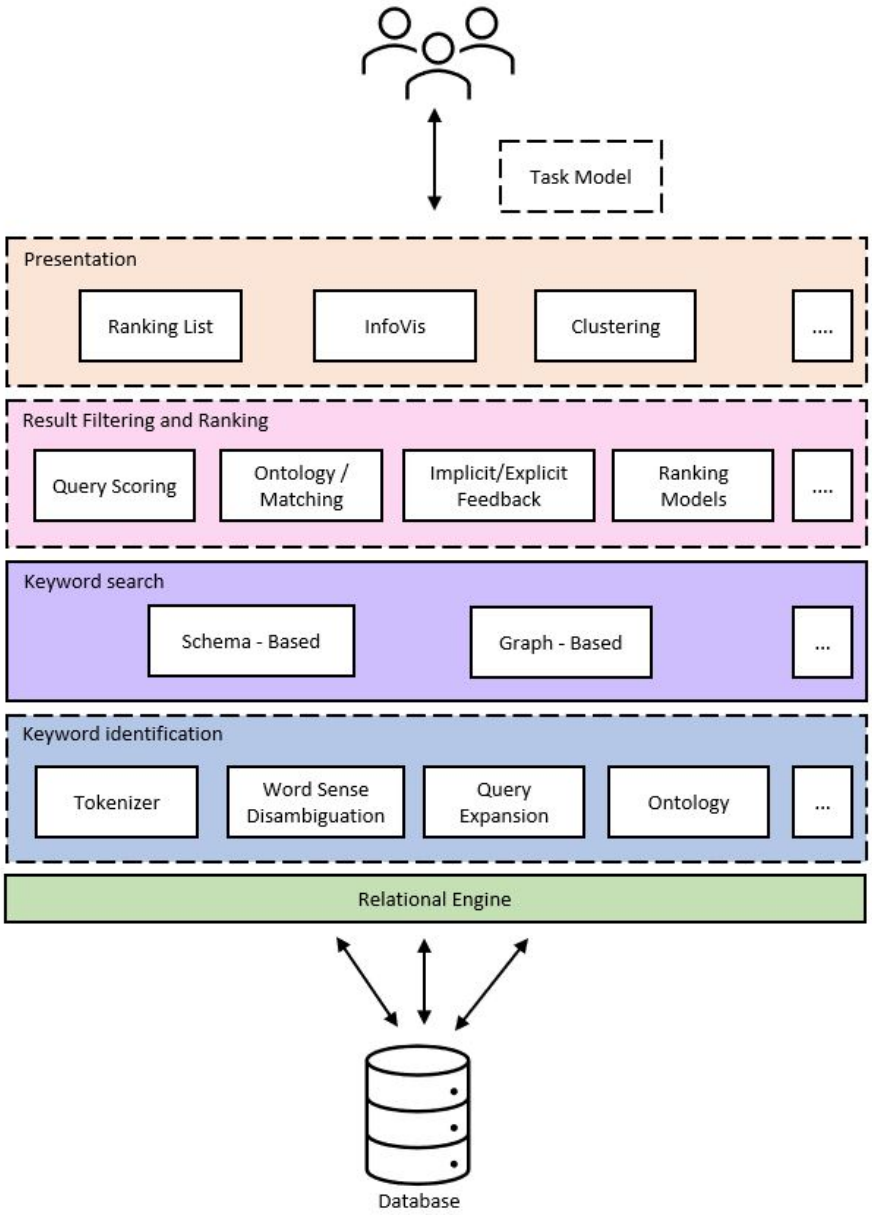

Figura 2.3: *Possível arquitetura de um Benchmark para consultas por palavras-chave. Inspirado em [6].*

as possíveis respostas são classificadas em uma lista *top-k* de relevância. Ao utilizar uma técnica que implementa o grafo de esquema, que explora os metadados, precisa-se encontrar uma correspondência entre o conjunto de palavras-chave e as informações de metadados ou tentar identificar em qual tabela e coluna a informação estará. Em uma proposta híbrida [7] constrói-se uma estrutura probabilística que pretende identificar a melhor correspondência de metadados ou valores em tuplas da banco de dados.

O módulo de *result filtering and ranking* pretende construir uma correspondência e classificar os resultados retornados. Para as técnicas que implementam grafos de dados deseja-se encontrar a melhor proximidade da palavra-chave ou conjunto de palavras-chave

a relação aos dados da banco de dados. Para as técnicas que implementam grafos de esquema deseja-se encontrar a maior pontuação para o conjunto de palavras-chave. A classificação dos resultados, ou *ranking* está habitualmente acoplada a técnica utilizada ou ao módulo de *keyword search* que determina quais os resultados são considerados relevante e estes obterão maior pontuação ou ditos com maior proximidade.

O módulo *presentation* pretende discutir sobre como a técnica apresenta os resultados ao usuário. Os resultados podem ser apresentados por uma *interface*, um arquivo de texto, um identificador ou a própria tupla. O nível de detalhes apresentado ao usuário permite a verificação pelo usuário se os dados retornados são relevantes, segundo seu julgamento.

Uma avaliação de referência se propõe mais do que explicitar quais atividades ou ações possuem comportamento ou *performance* cobiçáveis. Espera-se que apresente quais atividades ou ações que apresentaram falhas e essas atividades serão as que em uma próxima implementação deveriam ter maior atenção. O aperfeiçoamento constante é obtido através de avaliações que utilizam bases de dados reais e casos de testes convincentes.

É preciso entender o que exatamente está sendo avaliado, se a avaliação será sobre o algoritmo implementado ou sobre os resultados retornados. Ao submeter uma consulta a uma técnica espera-se o maior proximidade possível do conjunto de resultados esperados em relação ao conjunto de resultados retornados. O conjunto de resultados esperados deve ser o mesmo inclusive para técnicas com abordagens diferentes. Desta forma, a diferença será em função do tempo que cada técnica utilizará para processar os resultados.

Além de estabelecer as técnicas e os resultados esperados, o conjunto de métricas é um item importante a ser definido. Os conjuntos de métricas utilizado podem não ser apropriados em função da diversidade de algoritmos e diferentes disposição de resultados. O conjunto de resultados esperados e o conjunto de métricas demonstra o que se deseja avaliar a eficiência ou a eficácia.

2.4 Métricas

As métricas são maneiras de quantificar um elemento ou conjunto de elementos e demonstrar estatísticas, quanto a alguma habilidade ou atributo. A definição de métricas adequadas é capaz de demonstrar resultados relevantes da avaliação [36]. Na comunidade de Recuperação de Informação, as métricas utilizadas com maior frequência são: *recall* ou revocação, *precision* ou precisão e *f-meansure*.

Exemplificando o conjunto de resultados retornados na Figura 2.4 representam os subconjuntos A e B. O subconjunto A são os resultados esperados e relevantes que

estão no conjunto de resultados retornados. Já o subconjunto B são os resultados que não são esperados e não relevante neste contexto, mas fazem parte do subconjunto de resultados retornados. O subconjunto C são os resultados considerados relevantes e que a consulta não foi capaz de selecioná-los, assim como o subconjunto D são os resultados que não são considerados como relevantes e realmente não deveriam ser retornados. Assim, para cada consulta serão construídos os subconjuntos A, B, C e D [36].

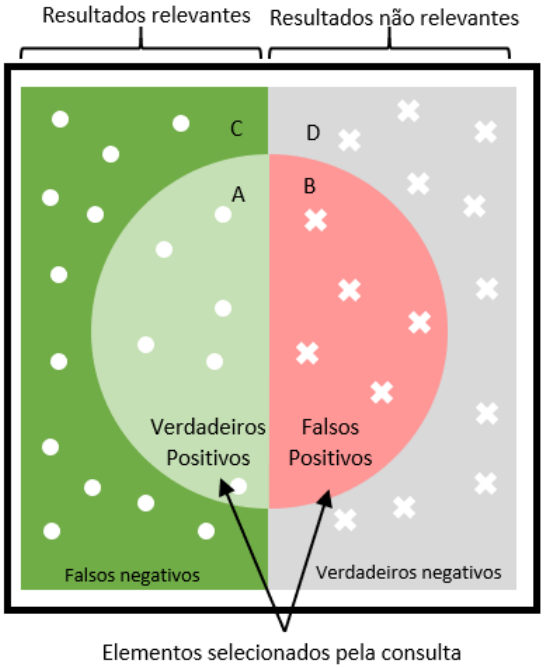

Figura 2.4: *Exemplo elucidativo sobre um universo de elementos a ser julgado.*

Tabela 2.4: *Análise dos resultados retornados*

	relevante	não relevante	
recuperados	A	B	A + B
não recuperados	C	D	C + D
-	A + C	B + D	-

- ***Precisão*** [13] é uma métrica estatística entre 0 e 1, e trata-se do percentual entre o conjunto de resultados recuperados relevantes e o conjunto total de resultados recuperados. Na Figura 2.4, o subconjunto A é o conjunto de resultados relevantes retornados e o subconjunto B sendo os resultados não relevantes, como descrito na fórmula 2-1. Quanto mais próximo a 1 maior será a eficácia da técnica em relação à consulta submetida.

$$P = \frac{A}{A+B} \tag{2-1}$$

- ***Média Simples de Precisão*** [13] é denotado pelo valor médio da precisão de cada resultado relevante conforme descrito na fórmula 2-2.

$$MSP = \frac{\sum_{i=0}^{N}}{N} \tag{2-2}$$

- ***Revocação*** [13] é a verificação do conjunto de resultados esperados em função dos resultados relevantes e não recuperados, conforme descrito na fórmula 2-3. Nesse conjunto há registros esperados que foram ou não recuperados. A medida retornada entre 0 e 1, e trata-se do percentual em função dos resultados esperados serem recuperados ou não, em sua totalidade.

$$R = \frac{A}{A+C} \tag{2-3}$$

- ***Média Simples da Revocação*** [13] é denotado pelo valor médio da revocação de cada resultado relevante conforme descrito na fórmula 2-4.

$$MSR = \frac{\sum_{i=0}^{N}}{N} \tag{2-4}$$

- ***F-measure*** [13] é a medida harmônica que combina precisão e revocação, conforme a fórmula 2-5 :

$$F\text{-}measure = 2 \times \frac{precisao \times revocacao}{precisao + revocacao} \tag{2-5}$$

2.5 Considerações Finais

Neste Capítulo foram apresentados os principais conceitos sobre as técnicas de consultas por palavras-chave a bancos de dados relacionais, a separação das técnicas em função do grafo gerados, ora de dados ora de esquema. Assim como os componente de um *benchmark* genérico, e um breve resgate sobre a origem dos *benchmarks* na área de banco de dados. Apresentado ainda, uma discussão sobre os *benchmarks* atuais, as possíveis partições de um *benchmark* para consultas por palavras-chaves com a utilização de conceitos e *benchmark* já existentes da área de banco de dados e as métricas que poderão ser utilizadas.

CAPÍTULO 3

Trabalhos Relacionados

Neste Capítulo são discutidos os trabalhos relacionados existentes que avaliam as técnicas de consulta com palavras-chave a bancos de dados relacionais, na Seção 3.1 os primeiros *benchmarks* propostos são descritos, e na Seção 3.2 o estudo mais recente é detalhado.

No contexto de pesquisa por palavras-chave a bancos de dados relacionais há poucos estudos relacionados a *benchmark*. Implementações de novas técnicas posteriores a 2010 utilizam, em sua maioria, o *benchmark* existente proposto por [14, 16] enquanto outras técnicas utilizam sua própria metodologia de avaliação. Estudos recentes reforçam a necessidade de revisão e reorganização dos cenários atuais de avaliação [12, 6]. Para comodidade na identificação dos trabalhos e seus respectivos autores, a Tabela 3.1 lista os *benchmarks* a serem explorados neste Capítulo.

Tabela 3.1: *Benchmarks Analisados.*

Ref.	Título	Ano
Benchmark Coffman [14, 16]	*A Framework for Evaluating Database Keyword Search Strategies*	2010
	An Empirical Performance Evaluation of Relational Keyword Search Techniques	2014
Benchmark Oliveira Filho [32]	*Benchmark* para Métodos de pesquisa por Palavras-Chave a Bancos de Dados Relacionais	2018

3.1 Benchmark Coffman

O *benchmark* proposto em [14] aborda a eficácia das técnicas de pesquisa por palavras-chave a bancos de dados relacionais. O mesmo define 03 banco de dados, sendo o *Mondial* um banco de dados de informações geográficas, *IMDB* um banco de dados de filmes e *Wikipédia* um banco de dados de artigos de propósito geral, conforme detalhado na Tabela 3.2 que apresenta o tamanho em *MBy*, a quantidade de relações e a quantidade de registros de cada banco. O banco de dados *Mondial* apesar de possuir menor tamanho apresenta maior complexidade em relação às suas restrições e relacionamentos entre as tabelas. É utilizado um número de 50 consultas para cada banco de dados e

tanto o conjunto de consultas como os resultados esperados para cada consulta foram arbitrariamente definidos, ou seja, não há uma justificativa para as escolhas.

Tabela 3.2: *Bancos de dados definidos por Coffman*

Bancos de Dados	Tamanho em *MB*	Quantidade de Relações	Quantidade de Tuplas
Mondial	9	28	17.115
IMDB	516	6	1673.074
Wikipédia	550	6	206.318

A relevância de cada resultado esperado em relação ao conjunto de consultas é uma medida binária, assim como decidir o que seria considerado relevante partiu de uma definição pessoal. As técnicas avaliadas foram *Efficient* , *Effective*, *SPARK*, *BANKS*, *Bidirectional*, *DPBF*, *BLINKS*, *DISCOVER* e *CD*, e estavam distribuídas entre técnicas que implementam grafo de esquema e grafo de dados. As métricas utilizadas para avaliação foram o *Top-1* que identifica o número de consultas para as quais o primeiro resultado é relevante. O *Reciprocal Rank* verificado a partir da classificação dos resultados relevantes de uma classificação para uma determinada consulta. A *precision* que trata da assertividade do resultado retornado em relação ao esperado, considerado relevante e a partir dela calcula-se a *Average Precision* e a *Mean Average Precision*. Para aplicar as métricas foram utilizados os top-1000 resultados de cada técnica para cada consulta.

Previa-se que as técnicas baseadas em grafos de dados teriam um desempenho inferior às técnicas baseadas em grafos de esquema. Porém, verificou-se que o tamanho do banco de dados foi um fator determinante em relação aos comportamentos de cada técnica e os resultados das consultas. As técnicas baseadas em grafo de dados sustentam problemas de escalabilidade em relação as técnicas baseadas em grafo de esquema em função da necessidade de mais recursos de hardware.

Em [16] a avaliação além da eficácia acrescenta a eficiência das técnicas de consulta por palavras-chave a banco de dados relacionais. Utilizou-se os mesmos bancos de dados, conjunto de consultas, e conjunto de resultados esperados construídos anteriormente [14]. As técnicas avaliadas foram o *DISCOVER*, *DISCOVER-II*, *DPBF*, *BLINKS*, *BANKS*, *BANKS-II* e *STAR*. Houve a reimplementação das técnicas *DISCOVER*, *DISCOVER-II*, *DPBF* e *BANKS* enquanto *BANKS-II*, *BLINKS* e *STAR* foi utilizado a implementação original.

É evidenciado [16] que nas reimplementações houve um esforço para otimizar os algoritmos e os resultados obtidos, se comparados aos relatos originais, se tornaram melhores. Para os experimentos definiu-se um tempo de espera/execução de cada consulta de 1 hora e uso de 5 *Gb* de memória *RAM*, um conjunto de resultados com 5 tuplas e no máximo 3 tentativas de execuções diferentes para cada técnica em cada consulta.

Ao executar a avaliação uma consulta poderia ser identificada como não executada se excedesse os recursos computacionais previamente definidos.

Em contraste ao estudo anterior [14], duas métricas são utilizadas para avaliar a eficiência, são elas tempo de execução e tempo de resposta. Tempo de execução é o tempo entre a submissão da consulta até o término da execução do algoritmo, enquanto o tempo de resposta é o tempo total gasto desde a submissão da consulta até a disponibilização dos resultados. *Recall*, *precision*, *mean average precision* continuaram sendo utilizadas para verificar a eficácia das técnicas.

Ao executar a avaliação nas técnicas previamente definidas, verificou-se que o número de termos em cada consulta não impacta na quantidade de resultados, mas impacta no tempo que a técnica necessitava para finalizar a execução da consulta. Para exemplificar as consultas definidas a Tabela 3.3 apresenta um fragmento das palavras-chave utilizados nas avaliações, é apresentado o número da consulta, a própria consulta, a semântica esperada e o *__search_id* que nos indicará exatamente tupla desejada. Desta forma, o tamanho do resultado é inversamente proporcional ao número de consultas concluídas com êxito. Se o tamanho do resultado esperado for significativamente grande será necessário proporcionalmente mais tempo de execução e mais memória.

Tabela 3.3: *Exemplo de Consultas IMDB*

Nº	**Consulta**	**Semântica**	***__search_id***
1	*denzel washington*	*Relevant results contain a single tuple from the person relation that is the tuple of the specified individual*	([39927668], [ ])
2	*clint eastwood*		([39172749], [ ])
3	*john wayne*		([39931125], [ ])
4	*will smith*		([39807078], [ ])
5	*harrison ford*		([39214967], [ ])
6	*julia roberts*		([40463372], [ ])
7	*tom hanks*		([39295438], [ ])
8	*johnny depp*		([39141807], [ ])
9	*angelina jolie*		([40255278], [ ])
10	*morgan freeman*		([39223764], [ ])

Para algumas técnicas, o tempo de resposta é muito próximo ao tempo de execução. O que pode justificar essa proximidade é a apresentação dos resultados aos usuários, ou seja, as técnicas não disponibilizam todos os resultados simultaneamente apresentando-os de forma incremental. Verificou-se o *recall*, a métrica que define a proporção de resultados esperados que foram recuperados e outra vez evidencia que o tamanho da banco de dados é fator determinante para que as técnicas consigam produzir resultados. *Precision* apontou que quanto maior a necessidade *p@1*, *p@10* menor a eficácia, e, se aumentarmos o *recall* a *precision* diminui.

Ao fim dos experimentos verificou-se que as técnicas baseadas em grafo de esquema apresentam um número maior de resultados que as baseadas em grafo de dados, o que não garante que o resultado esperado pertença ao grupo de resultados recuperados. A reimplementação das técnicas citadas anteriormente reflete resultados diferentes de tempo comparados às implementações originais. Por outro lado, percebeu-se que os resultados variam de acordo com o tamanho e complexidade de relacionamentos das bases de dados. Os melhores resultados foram obtidos na menor banco de dados com consultas inclusive, classificadas como não executadas para algumas bases e técnicas.

3.2 Benchmark Oliveira Filho

No *benchmark* proposto por Oliveira Filho [32] as técnicas *Keymantic* [5] e Ramanda [35] foram avaliadas e ambas implementam grafo de esquema. A técnica *keymantic* possui acesso aos dados e necessita de intervenção do usuário na escolha de uma configuração para então demonstrar o resultado, conforme esclarecido no Capítulo 2. A técnica de Ramada et al. não possui acesso prévio aos dados e acessa inicialmente uma tabela de metadados própria da técnica e somente após identificar se há banco de dados relevante para a consulta acessa a banco de dados realizando as buscas.

Os bancos de dados utilizados foram o *Mondial*, um banco de dados de informações geográficas, *IMDB* um banco de dados de filmes, o *DBLP* um banco de dados de informações bibliográficas na área de ciência da computação e *Northwind* um banco de dados sintética de vendas conforme informações de tamanho em *MB*, quantidade de relações e de tuplas dispostos na Tabela 3.4.

Tabela 3.4: *Bancos de dados definidos por Oliveira Filho*

Banco de Dados	**Tamanho em *MB***	**Quantidade de Relações**	**Quantidade de Tuplas**
Mondial	11	33	27.210
IMDB	429	7	5518.540
DBLP	58	6	881.876
Northwind	1.1	13	3.308

Um conjunto de 50 consultas e sua respectiva semântica foi definido por banco de dados sendo definidos arbitrariamente. Em razão das técnicas avaliadas implementarem grafo de esquema e em razão de tais técnicas verificarem primeiro os metadados das tabelas definiu-se a classificação dos termos da consulta em se tratando de termos de valor para a banco de dados ou termos de esquema da banco de dados. Para cada consulta definiu-se uma semântica esperada mas não estabeleceu um conjunto de resultados esperados [32].

Em contraste às propostas de [14, 16], os termos da consulta são classificados, e essa classificação é relativo a sua localização como termo de esquema - termos que poderão ser encontrados como metadados; ou termos de valores para termos que serão encontrados em meio as tuplas da banco de dados. Foi calculado uma média do número de palavras-chave nas consultas que é de 3,36, e percebe-se uma predileção por palavras que seriam classificadas como termos de esquema.

A avaliação de relevância é definida como binária, isto é, para cada retorno é definido se os resultados são relevantes e não, apesar da definição formal não há relatos de quais são os resultados relevantes. A Tabela 3.5 exemplifica uma fração das palavras-chave escolhidas para a banco de dados *IMDB* e a descrição da semântica da consulta, mas o resultado esperado por cada consulta não é apresentado.

Tabela 3.5: *Exemplo de Consultas IMDB*

Nº	**Consulta**	**Sentido**
1	*“will smith”*	Informações sobre *Will Smith*
2	*"julia roberts"*	Informações sobre *Julia Roberts*
3	*"gone with the wind"*	Informações sobre *Gone with the wind*
4	*"star wars"*	Informações sobre *Star Wars*
5	*casablanca*	Informações sobre *Casablanca*
6	*actors*	Lista de atores
7	*movies*	Lista de filmes
8	*directors*	Lista de diretores
9	*Genre*	Lista de gêneros de filmes
10	*movies directors*	Lista de diretores de filmes

As técnicas utilizadas nas avaliações são as disponibilizadas pelos respectivos autores. As métricas utilizadas são similares às utilizadas em [14] e as avaliações se limitaram à eficácia das técnicas. *Precision* e *recall* são métricas utilizadas em sistemas de recuperação de informação e são base para o cálculo de outras métricas. Métricas como *Mean Reciprocak Rank*, *Average Precision* e *Mean Average Precision* que foram utilizadas são dependentes de *precision* e *recall*. O número de *Top-1* resultados relevantes e a *precision* em *n (P@n)* também são definidas como métricas.

Para executar as avaliações definiu-se primeiramente as técnicas a serem utilizadas, em seguida configurou-se o ambiente de testes e executou-se as consultas nas bases de dados. Em posse dos resultados da submissão das consultas às técnicas, julgou-se os itens em relação à relevância. O tempo de execução de cada consulta em cada técnica foi limitado em 1 hora. Com o início das quantificações verificou-se que o número de relações é um fator determinante no número de consultas que ultrapassaram o tempo limite e, portanto, consideradas não finalizadas.

A técnica Ramada et al.[35] apresenta, em um contexto geral, resultados diferentes a *Keymantic* [5] dada a construção das configurações conforme descrito no Capítulo 2. *Keymantic* verifica a correspondência dos termos da consulta aos metadados e se não encontrar inicia a construção das configurações pontuadas pela associação de cada termo a cada coluna das tabelas da banco de dados. A pontuação é dada pela similaridade da palavra-chave com a coluna da banco de dados.

Em contraditório Ramada et al. [35] verifica inicialmente uma tabela própria de metadados, sem acessar a base inicialmente, que contém informações das bases de dados possíveis de serem verificadas e localização, acesso e uma descrição em formato de conjunto de palavras que dá contexto a cada base. Se este contexto não for devidamente discriminado, apesar de haver bancos de dados relacionadas a técnica sequer identificará a mesma como relevante e, consequentemente não a utilizará nos experimentos. Após essa identificação, a técnica verifica a correspondência da palavra-chave aos metadados ou aos dados propriamente dito.

Resultados das métricas são apresentados, porém não descreve quais seriam exatamente os resultados relevantes, ou um conjunto de resultados relevantes. Subentende-se apenas que o primeiro resultado retornado é considerado resultado relevante. Após todas as avaliações verificou que, para as bases de dados com menor quantidade de relações e tuplas as técnicas de consulta obtiveram melhores resultados. Por fim, todas as avaliações, teste e quantificações foram manualmente executadas.

3.3 Considerações Finais

Neste Capítulo foram apresentados os dois *benchmarks* para técnicas de consultas por palavras-chave a bancos de dados relacionais. Individualmente, os *benchmarks* foram evidenciados em relação a seus componentes - construção do conjunto de consultas, semântica e resultados esperados, o que houvesse, a metodologia de ação em relação aos experimentos, métricas utilizadas e seus resultados apresentados.

CAPÍTULO 4

Benchmark Proposto

Neste Capítulo será apresentado o *benchmark* proposto, na seção 4.1 são elencados os itens necessários para as avaliações, na seção 4.2 os resultados esperados e a semântica pretendida para cada consulta. Por fim, na 4.3 um paralelo para apresentar a dissimilitude entre os *benchmarks* existentes e o *benchmark* proposto neste estudo.

A proposta deste trabalho é definir um *benchmark* capaz de avaliar não só a eficácia, mas também como as técnicas de consulta por palavras-chave a banco de dados relacional comportam-se em relação ao tratamento semântico das palavras-chave da consultas. Conforme esclarecido na seção 2.3, é preciso fragmentar a avaliação da técnica desde o tratamento da consulta até o resultado esperado. Uma técnica de consulta com palavras-chave a banco de dados relacionais possui vários pontos de atenção.

Nesta pesquisa, será objeto de estudo o tratamento semântico das consultas em relação a expansão da consulta, a proximidade sintática, a limpeza da consulta, utilização de dicionários e funções de agregação. Verificaremos qual o comportamento da técnica em relação à cada característica que deseja-se que a mesma contemple. A eficiência não foi objeto deste estudo pois as técnicas não apresentam possibilidade de comparação em todas as suas etapas.

A limpeza, expansão da consulta, a proximidade sintática, a utilização de dicionário e função de agregação são características que são capazes de aumentar, corrigir ou excluir termos do conjunto de palavras-chave que podem ser significativos. Um erro de digitação, por exemplo, pode ser resolvido com a proximidade sintática ou com a derivação da palavra-chave. A expansão da consulta e os dicionários permitem a inserção de palavras sinônimas e aumentam a possibilidade de correspondência entre o conjunto de palavras-chave e os metadados ou os valores da banco de dados [27, 5, 35].

A limpeza da consulta e remoção de *stopwords* acontece com a exclusão de termos como artigos, preposições, pronomes e outras palavras que não auxiliam na correspondência do conjunto de palavras-chave aos valores do banco de dados, e que não representam ganho semântico. O tratamento de termos que representam funções de agregação em *SQL* possibilita o entendimento semântico de palavras no conjunto de palavras-chave e a utilização de funções de ordenação auxiliam no detalhamento da ordem

pretendida dos resultados [27, 5, 35].

Técnicas de consulta por palavras-chave a bancos de dados relacionais devem, ao final do processo de submissão da consulta, conseguir o maior número de resultados relevantes, segundo o julgamento do usuário. Trabalhos anteriores [16, 32] propuseram avaliar a eficácia de tais técnicas. Todavia, estes trabalhos não discutem como o conjunto de palavras-chave é analisado, mesmo diante do fato de que as técnicas possuem diferentes formas de tratar o conjunto de palavras-chave. Portanto, a presente proposta objetiva, além de avaliar as técnicas em relação a eficácia, ou seja, aos resultados recuperados para cada palavra-chave definida, estabelecer bancos para avaliação do tratamento dado a cada conjunto de palavras-chave.

A Figura 4.1 ilustra um processo de submissão da consulta a uma técnica de pesquisa por palavras-chave a bancos de dados relacionais. Diferentemente das propostas anteriores, pretende-se verificar se é realizado algum pré-processamento em relação a consulta em função das características explicitadas a seguir. Ao submeter uma consulta, espera-se que as técnicas realizem pré-processamentos no intuito de "enriquecê-las". Após o pré-processamento a técnica acessará informações sobre metadados ou a banco de dados e então são construídos os mapeamentos da consulta no grafo de esquema ou no grafo de dados. Finalmente, os resultados são apresentados ao usuário que submeteu a consulta.

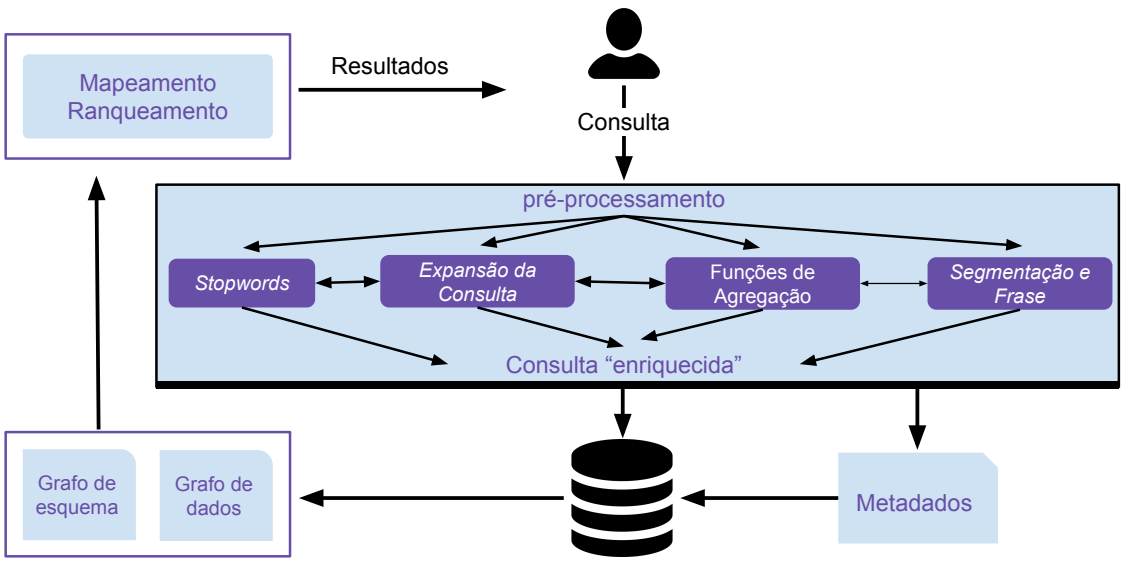

Figura 4.1: *Ilustração de técnicas de pesquisa por palavras-chave a bancos de dados relacionais.*

O pré-processamento, quando existente, pode auxiliar na possibilidade de atender de forma mais precisa a expectativa do usuário que submeteu a consulta. A baixa taxa de precisão dos resultados, o tempo gasto para executar uma consulta, isso quando a consulta é finalizada, a heterogeneidade das implementações como demonstrado em

[16, 32], deixa evidente porque até o presente momento não é possível utilizar as técnicas de consulta por palavras-chave a bancos de dados relacionais em larga escala ou escala industrial.

4.1 Componentes

Um *benchmark* é formado por um conjunto de ações previamente estabelecidas com objetivos claros. Conforme descrito na seção 2.2, um *benchmark* é composto por um conjunto de dados, um conjunto de configurações, o conjunto de consultas ou testes e o conjunto de métricas. Cada um destes componentes será apresentado a seguir de acordo com a proposta deste trabalho.

4.1.1 Bancos de Dados

Serão utilizados três bancos de dados de domínio público e disponíveis *online* para *downloads* (*IMDB*, *DBLP* e *Mondial*). Esses mesmos bancos de dados foram utilizadas em [32], porém, no presente trabalho serão consideradas as atualizações mais recentes dos dados. O banco de dados *Northwind*, que também foi utilizada em [32], não será considerada neste estudo por se tratar de um banco sintético, isto é, não apresenta dados reais. Além disso, em uma análise prévia sobre as palavras-chave do referido trabalho, não foram encontradas palavras que contivessem as características necessárias, discutidas na subseção 4.1.2.

Para cada banco de dados pode ser necessário alguma alteração para adequação as regras das técnicas que necessitam das definições de relacionamentos entre chave primária e estrangeira. A seguir, serão descritos os bancos de dados a serem utilizadas na avaliação de técnicas de consulta. Em todas as tabelas dos três bancos de dados foi adicionado uma coluna de nome *__search_id* de valor numérico e único em todo o banco de dados para identificação dos registros no conjunto de resultados esperados para cada consulta.

Mantido pela *Amazon en Seattle - WA*, o *IMDb*[1] é um banco de dados *on-line* com informações sobre filmes, programas de televisão, vídeos caseiros e videogames, e *streams* da *internet*. Nele encontram-se dados sobre elenco, equipe de produção e biografias de pessoal, resumo de enredo, curiosidades, avaliações e classificações do público geral ou de fãs. O banco de dados propriamente dito [2], atualmente[3] encontra-se com os seguintes arquivos:

[1] https://www.imdb.com/
[2] https://datasets.imdbws.com/
[3] *Download* realizado em 24 de Fevereiro de 2020

- *title.akas.tsv.gz* - contém informações sobre os títulos como nome, região, língua, dentre outros e possui 176 *MB*;
- *title.basics.tsv.gz* - também contém informações sobre os títulos como tempo de duração, ano de início e fim da execução e o gênero, dentre outros e possui 463 *MB*;
- *title.crew.tsv.gz* - contém informações sobre diretores e autores e possui 174 *MB*;
- *title.episode.tsv.gz* - contém informações sobre episódios como número e temporada, dentre outros e possui 95 *MB*;
- *title.principal.tsv.gz* - contém informações sobre por título, elenco por título e possui 1,35 *GB*;
- *title.ratings.tsv.gz* - contém informações sobre a classificação e o número de votos, possui 34,70 *MB*;
- *name.basics.tsv.gz* - contém informações sobre as pessoas que atuaram no filme como profissão e aniversário, e possui 528 *MB*.

O banco de dados original não estava normalizada e para que pudesse ser utilizada nas técnicas a normalização precisou ser realizada. O arquivo *name.basics.tsv.gz* que contém informações sobre as pessoas que trabalharam em determinados filmes possui as informações de *nconts* - identificador único da pessoa; *primaryName* - nome; *birthYear* - ano nascimento; *deathYear* - ano morte; primaryProfession - as profissões da pessoa; e *knownForTitles* - filmes pelos quais a pessoa é conhecida. As colunas *primaryProfession* e *knownForTitles* possuem mais de uma informação, e foram utilizadas para construir uma entidade associativa entre a pessoa e o filme *movies_actors*, e foram utilizados neste momento o valor encontrado em *primaryProfession* de *actor* e *actress*.

O arquivo *title.crew.tsv.gz* possui os campos *tconst* identificador do filme, *directors* e *writers* - identificador da pessoa e pode conter mais de um valor. Este arquivo foi utilizado para criar a tabela associativa *movies_directors* e neste momento a coluna *writers* não foi utilizada.

O arquivo *title.akas.tsv.gz* possui as colunas *titleId* - identificador do filme e que é associado ao campo *tconst*, *ordering* - um número sequencial em relação ao filme, *title* - o título do filme na *region* e *language*, *types* - é a classificação da produção como festival como *alternative video*, *alternative*, *imdbDisplay*, *alternative tv*, *working*, *video*, *alternative dvd*, *tv*, *dvd video*, *original*, *tv video*, *imdbDisplay video*, dentre outras, *attributes* - que trata se o título é original ou não e, por fim, *isOriginalTitle* - um campo com valor 0 ou 1 para indicar se aquele é o título original do filme. Nesta tabela é possível notar que o campo *title* apesar de ser um identificador, este campo isoladamente se repete, mas ao ser combinado com a campo *ordering* não há repetição. Está foi a tabela principal que deu origem a tabela *movies*, foram considerados todos os filmes cujo campo

isOriginalTitle fosse igual a 1, o campo *IMDBDisplay* e *ordering* não foram utilizados por não serem motivo de busca observável até o momento.

Para adicionar o campo *rating* na tabela *movies*, foi utilizado o campo *averageRating* do arquivo *title.ratings.tsv.gz* que possui os campos *tconst* - identificador do filme, *averageRating* - a classificação da produção e *numVotes* - número de classificações que a produção já recebeu até o *download* do arquivo.

O *title.basics.tsv.gz* possui os campos *tconst* - identificador do filme, *titleType* - os valores podem variar entre *tvSeries*, *videoGame*, *tvSpecial*, *tvShort*, *tvMovie*, *tvEpisode*, *video*, *movie* e *tvMiniSeries*; *primaryTitle* - título da produção, *originalTitle* - nome original da produção, *isAdult* - valor de 0 ou 1, *startYear* - ano de lançamento da produção, *endYear* - ano final da produção - mais adequado para séries, se for um filme ou episódio único este campo não é preenchido, *runtimeMinutes* duração da produção em minutos, e *genres* - os gêneros para esta produção podendo ser mais de um. O campo *titleType* foi adicionado a Tabela *movies* e com o campo *genres* foi criada uma tabela com todos os possíveis gêneros e depois criada uma tabela associativa *movies_genres* que relaciona o filme ao gênero.

Por fim, o arquivo *title.episode.tsv.gz* que possui os campos *tconst* - identificador do episódio do filme, *parentTconst* - identificador do filme, *seasonNumber* - o número da temporada da séries e *episodeNumber* - o número do episódio de uma temporada de uma série, este arquivo foi utilizado por completo.

Após estas alterações banco é composto das Tabelas *movies*, *person*, *genres*, *episodes*, *movies_actors*, *movies_directors* e finalmente a *movies_genres*, conforme detalhado na Tabela 4.1 há um total de 40.240.524 registros, 5,51 *GB* de dados e um total de 7 relacionamentos entre as tabelas, que pode ser verificado no Diagrama Entidade Relacionamento apresentado na Figura 4.2.

Tabela 4.1: *Tabelas da banco de dados IMDB*

Banco	***Tabelas***	***Quantidade de registros***	***Tamanho em KB***	***Chaves estrangeiras***
IMDB	*episode*	2.866.700	381952	1
	genres	28	2,08	0
	movies	6.579.412	1101004,8	0
	movies_actors	15.399.455	1562378,24	2
	movies_directors	4.379.357	458752	2
	movies_genres	10.014.086	933888	2
	person	1.001.486	1342177,28	0
	Total	***40.240.524***	***5780154,4***	***7***

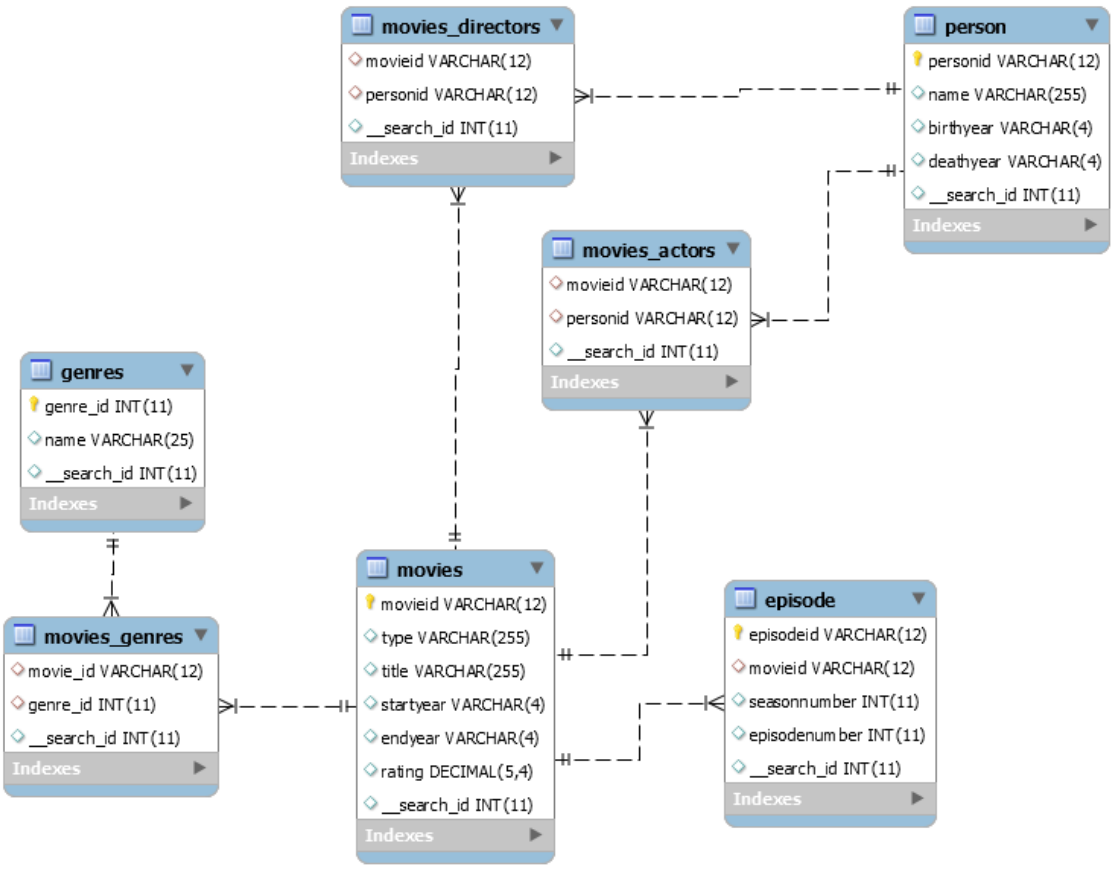

Figura 4.2: *Diagrama Entidade-Relacionamento da banco de dados IMDB.*

Outro banco de dados utilizado neste trabalho é a *Digital Bibliography Library Project* (*DBLP*)[4]. A *DBLP* é um banco de dados de publicações de documentos na área de Ciência da Computação. Este banco de dados indexa mais de 4,4 milhões de publicações, publicadas por mais de 2,2 milhões de autores, contemplando mais de 40.000 volumes de periódicos, mais de 39.000 conferências e *workshops*, e mais de 80.000 monografias. Tudo isso, disponibilizado atualmente[5] com o arquivo *dblp.xml* com 2,36 *GB*.

Para processamento e visualização do arquivo disponibilizado foi utilizado o *Pentaho Data Integration*[6]. Após a inserção dos dados em um Sistema Gerenciador de Banco de Dados, foi realizado uma analise e somente as tabelas descritas na Tabela 4.2 foram utilizadas. Das tabelas utilizadas foram mantidos todos os registros totalizando 14.169.596 tuplas e 2,8 *GB* de dados.

[4] https://dblp.uni-trier.de/
[5] *Download* realizado em 07 de Janeiro de 2020
[6] https://bitlybr.com/JOSfgQ

Tabela 4.2: *Tabelas da banco de dados DBLP*

Banco	*Tabelas*	*Quantidade de registros*	*Tamanho em KB*	*Chaves estrangeiras*
DBLP	*article*	2.166.665	956416	3
	author	651.005	58777,6	0
	author_article	8.708.567	820224	2
	book	17.922	4689,92	4
	booktitle	13.952	1392,64	0
	inproceedings	2.526.762	1048576	2
	proceedings	81.047	40345,6	4
	publisher	2.162	162	0
	series	1.514	5,47	0
	Total	***14.169.596***	***2930589,23***	***15***

Na Figura 4.3 é possível verificar os relacionamentos e a complexidade do banco de dados.

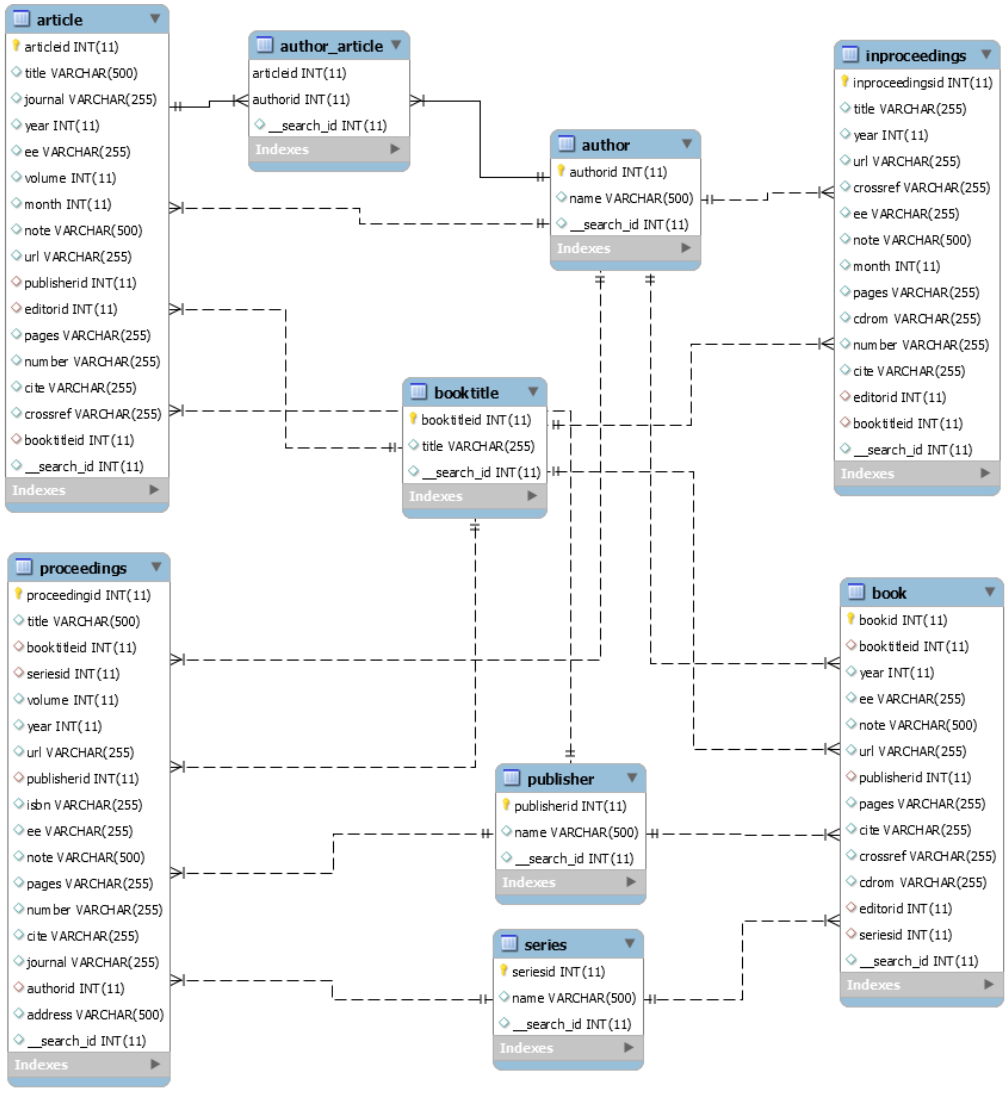

Figura 4.3: *Diagrama Entidade-Relacionamento da banco de dados DBLP.*

Por fim, este trabalho considera o banco de dados *Mondial*[7], banco este que provê informações geográficas globais. Disponibiliza arquivos em vários formatos como *XML*, *RDF* e SQL. Inclusive, são fornecidos arquivos SQL para os sistemas gerenciadores de Bancos de dados *Oracle*, *PostgreSql* e *MySql*. O banco de dados disponibilizado atualmente[8] possui um *script* de criação do *schema* com 7,41 *KB* e outro arquivo para os *inputs* com 1,33 *MB*.

Este banco de dados não precisou passar por nenhum processamento ou alteração, as tabelas são descritas na Tabela 4.3. O banco de dados por sua vez, apesar de um número menor de registros, 21.344 tuplas e 2,54 MB de dados possui um total de 63 relacionamentos conforme pode ser visto na Figura 4.4.

[7] http://www.dbis.informatik.uni-goettingen.de/Mondial/
[8] *Download* realizado em 20 de Janeiro de 2020

Tabela 4.3: *Tabelas da banco de dados Mondial*

Banco	***Tabelas***	***Quantidade de registros***	***Tamanho em KB***	***Chaves estrangeiras***
Mondial	*Borders*	320	29,2	2
	City	3.111	473	2
	Continent	5	0,412	0
	Country	288	36,9	1
	desert	63	6,88	0
	economy	238	32,3	1
	encompasses	242	25,4	2
	ethnicgroup	*540*	*54,6*	*1*
	geo_desert	154	16,9	3
	geo_estuary	265	28,7	2
	geo_island	418	46	3
	geo_lake	253	27	3
	geo_mountain	295	33,7	2
	geo_river	851	89,3	3
	geo_sea	735	78,7	3
	geo_source	219	23,4	3
	island	279	40,3	0
	islandin	349	40,3	4
	ismember	8.008	823	2
	lake	130	21,7	1
	language	144	14	1
	located	857	126	6
	locatedon	434	55,1	4
	mergewith	54	5,13	2
	mountain	240	36,5	0
	mountainisland	67	6,8	1
	organization	153	29,9	3
	politics	238	34,1	1
	population	238	27,3	1
	province	1.450	229	1
	religion	454	45,6	1
	river	218	54,9	4
	sea	34	2,66	0
	Total	***21.344***	***2594,682***	***63***

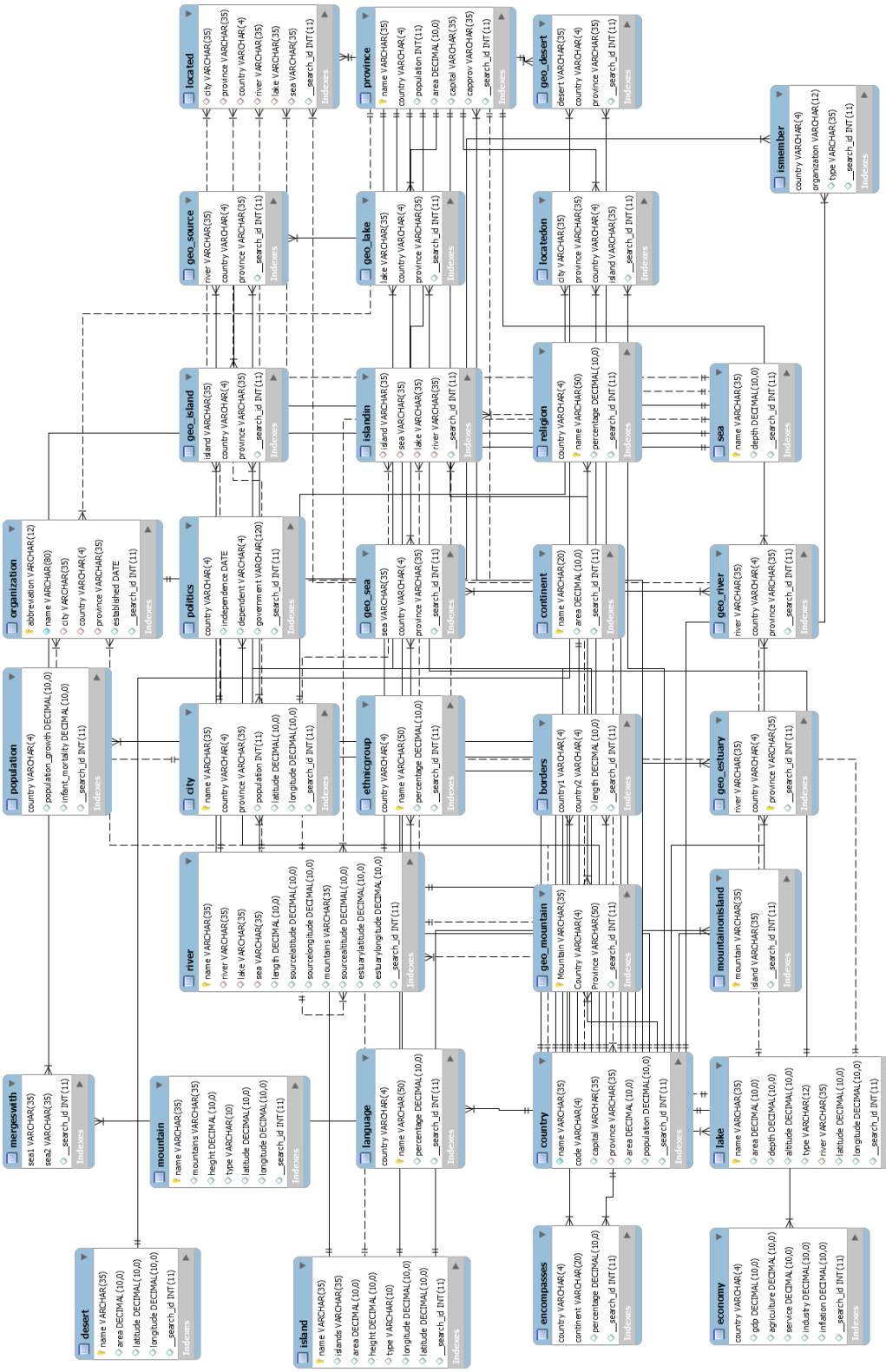

Figura 4.4: *Diagrama Entidade-Relacionamento da banco de dados Mondial.*

4.1.2 Consultas

As consultas, ou conjunto de palavras-chave propostas por Oliveira Filho [32] não serão utilizadas neste trabalho, pois deseja-se analisar características que não poderiam ser verificadas com o referido conjunto. Apesar de verificar a eficácia das técnicas em relação às palavras escolhidas, deseja-se verificar a inferência das características listadas a seguir. As palavras-chave serão definidas previamente, assim como os resultados esperados e a semântica pretendida. As consultas serão submetidas às técnicas e posteriormente analisadas e quantificados.

Uma preocupação ao construir as consultas estava em encontrar palavras que refletissem as ações de um usuário que se enquadrasse em uma ou várias das características descritas a seguir. Estas consultas precisam assemelhar-se com palavras e frases do dia-a-dia com a inclusão ambiguidades ou a inserção de termos que não auxiliam na busca ou na capacidade de encontrar sentindo no conjunto de palavras-chave agrupadas pelo uso de aspas duplas. Desta forma, optamos por buscar palavras ou conjunto de palavras que realmente foram utilizadas em algum momento, em um contexto real. Para que a escolha não se tratasse de um conjunto de palavras criadas sem nenhum critério científico e que fosse bem definido, objetivo e replicável, buscamos por listas de palavras, frases ou expressões relacionadas à cada banco de dados.

Desta forma optamos inicialmente por buscar palavras e ou frases na redes social *Twitter*, utilizando *hastags* específicas. Ao iniciar a captura dos tuítes e apesar de fazê-lo em várias datas diferentes a quantidade de tuítes iguais e similares foi maior do que o esperado. Com a banco de dados *Mondial* não foi possível utilizar o *twitter*, pois a *hastag #mondial* é utilizado em uma marca de eletrodomésticos e não para a banco em questão. Assim como utilizar as *hastag #geography*, *#mapamundi*, ou *#mundi* não revelou palavras que pudessem ser utilizadas. Para a banco *DBLP* apesar de haver tuítes a quantidade a diversidade era pequena. Para a banco *IMDB* haviam significativamente mais tuítes, mas 80% deles relativos a classificação de algum título.

Para utilizar palavras que não fossem criadas aleatoriamente e sem nenhum critério o do *site Keyword Tools*[9] foi utilizado. No *site* uma palavra ou um conjunto de palavras é submetido a uma busca das palavras, perguntas ou proposições mais buscadas em *sites* como o *Google*, *Bing*, *Youtube*, *Amazon*, *eBay*, *Play Store*, *Instagram* e *Twitter*. Foram retiradas palavras principais de cada contexto e submetidas a uma pesquisa, desta pesquisa as listas disponibilizadas foram analisadas e utilizadas as que se enquadravam nas características desejadas e que também representassem diversidade em relação às bancos de dados. As consultas selecionadas podem ser observadas no Apêndice A e são classificadas em relação às características demonstradas a seguir, assim como a

[9] https://keywordtool.io/pt

semântica desejada para cada consulta. Para cada consulta selecionada é definida em quais características se enquadra, a semântica desejada.No Apêndice B serão apresentados o universo de registros que a técnica deveria realizar a busca dos resultado assim como quais exatamente são os dados esperados. Em razão de em algumas consultas a quantidade de resultados ser significativo e poderia comprometer a legibilidade da tabela, é apresentado apenas o número de registros esperados, os valores exatos podem ser observados em um arquivo externo que pode ser acessado pelo link[10].

Para entender melhor as consultas selecionadas discutiremos as características que as técnicas deveriam considerar e obviamente, a maneira que cada técnica executa a consulta interfere diretamente no resultado final e as consultas de cada banco de dados foram submetidas em todas as técnicas. As características a serem analisadas são (1) *Stopwords*, (2) Expansão da consulta, (3) Funções de Agregação e (4) Segmentação e frases são detalhadas a seguir.

4.1.2.1 *Stopwords*

As *stopwords* incluem os artigos definidos ou indefinidos, proposições essenciais ou acidentais; conjunções alternativas, conclusivas, explicativas, aditivas ou adversativas; pronomes do caso reto, oblíquos ou átomos. Elas não adicionam significado às palavras escolhidas e não deveriam interferir no resultado final. A intenção é verificar se a técnica consegue identificar tais *stopwords* e retirá-las da consulta, considerando somente as demais palavras que as circundam [12]. Se considerarmos a consulta *"movies of commedy"*, a palavra *"of"* é uma *stopword* e deve ser retirada. Assim, uma técnica deveria considerar quaisquer resultados contendo as palavras-chave *"movies"* ou *"comedy"*.

Para avaliação do aspecto relativo ao tratamento *stopwords*, será definida a semântica de cada consulta, um conjunto de resultados será definido não importando a classificação de cada resposta especificamente. Isto é, independente do conjunto de resultados retornados da pesquisa realizada pela técnica o que importa é a *stopword* não estar contida nos resultados retornados.

4.1.2.2 Expansão da Consulta

A expansão da consulta é a capacidade da técnica utilizar dicionários de sinônimos e polissemia. A utilização da expansão da consulta é na tentativa de realizar a proximidade sintática ou que tenham o mesmo significado ampliando a possibilidade de correlações significativas. Além disso, deseja-se analisar se a técnica é capaz de realizar o *steming*, ou seja verificar a formação da palavra em relação a raiz, radical, derivação

[10] https://bitlybr.com/IWCxgkl

prefixal e sufixal, desde que mantenha o sentido e significado. Realizar a radicalização da palavra em todas as palavras imputadas aumenta a capacidade de encontrar similaridade [35]. A verificação de uma palavra exata restringe o tratamento de ambiguidades e o não conhecimento exato dos campos da banco de dados. Se considerarmos a consulta *"movies of comedy"*, a palavra *"movies"* poderia ser associada a *"movie"*, *film* e *"picture"*, enquanto *"comedy"* poderia estar associada a *"drollery"*, *"clowning"* e *"funniness"*.

Neste grupo as consultas selecionadas necessitam ser verificadas em sua construção léxica e sinônimos, a quantidade de resultados retornados pode variar de acordo com a semântica atribuída a cada consulta.

4.1.2.3 Funções de Agregação

Identificar se entre as palavras-chave submetidas há uma palavra que pode ser traduzida, como um comando da linguagem *SQL* que pode melhorar a precisão dos resultados retornados. Esses comandos da linguagem *SQL* são funções de agregação que é a computação de várias linhas em um único resultado [19]. Comandos como soma, média, maior e menor valor e até mesmo uma contagem podem ser traduzidas em linguagem *SQL*. Suponha que o usuário submeta a seguinte consulta *recent movie "star wars"*. Este poderia ser traduzido como o *max(year) movie "star wars"*.

Neste grupo a consulta selecionada necessita ser verificada em relação a possibilidade de substituição de palavras por funções disponíveis na linguagem *SQL*.

4.1.2.4 Segmentação e Frase

A segmentação é a capacidade de identificar que algumas palavras precisam ser tratadas como um conjunto para que o significado desejado seja encontrado. Algumas palavras só fazem sentido ou possuem significado relevante se associadas a outras palavras ou a um conjunto de palavras[35]. Supondo que o usuário submeta a seguinte consulta *movie "star wars"* é esperado que a técnica entenda que *"star wars"* precisa ser tratada de forma conjunta e não como palavras isoladas.

Neste grupo a técnica precisa ter a capacidade de entendimento de conjuntos de palavras e não somente a verificação de cada palavra de forma isolada. As aspas duplas auxiliarão na identificação de que várias palavras precisam ser tratadas como uma palavra-chave, em outras a junção das palavras precisa ser entendida semanticamente.

4.1.3 Métricas Utilizadas

Analisaremos as consultas baseado nas informações do universo total de dados. Para cada consulta imputada foi definida a semântica e os resultados esperados, e após a submissão da consulta serão listados os resultados retornados. Isto é, a definição de

quais métricas serão efetivamente utilizadas para quais conjuntos de palavras-chave só é possível após a construção de conjunto de resultados esperados.

Conforme descrito na Seção 2.4 as métricas utilizadas para verificar a eficácia serão Precisão, Revocação, Média Simples da Precisão, Média Simples da Revocação e *F-measure*. Este trabalho não organizou os resultados em uma lista ordenada, mas em um conjunto de resultados esperados e encontrados, para tanto não utilizou métricas para tais fins.

As métricas apresentadas até o momento são relativos ao resultado final, e apesar do avaliação ser feita até a apresentação dos resultados, o foco principal é a verificação do pré-processamento de cada uma das características semânticas descritas anteriormente. Será então verificado se foi realizado ao menos um pré-processamento conforme desejado para as consultas detalhadas e descritas no Apêndice A. Será apresentado em forma de uma pergunta **"Realizou o processamento desejado?"** e a resposta apresentada será **Sim** ou **Não**. Se houver mais de uma característica sendo verificada, será inserida a abreviação da característica realizada. A métrica é binária pois em razão da não apresentação ao usuário todas as etapas internas de processamento da consulta nas técnicas *BANKS-II* e *Keymantic*, é possível apenas deduzir a existência ou não do pré-processamento.

4.2 Resultados Esperados e Semântica

Considerando que este é ponto de principal questionamento dos *benchmarks* para consultas com palavras-chave a banco de dados relacionais, será também o ponto de maior atenção. Para cada consulta foi definido uma semântica que será utilizada para definir o conjunto de resultados esperados. Estes resultados podem ser um conjunto de um ou vários resultados dependerá do que se deseja verificar em cada consulta. A semântica definida está disposta no Apêndice A, e os resultados esperados de forma resumida no Apêndice B e detalhado no arquivo[11]. Nos três bancos de dados foram inseridos uma coluna __search_id para que a tupla seja identificada unicamente e será este o valor disponibilizado como resultado relevante, quando possível. Como em alguns casos os conjunto de resultados esperados é significativamente grande serão disponibilizados arquivos em formato csv[12] para verificação.

Nas consultas que se deseja verificar a limpeza e a retirada de *stopwords* será verificado se a palavra fez parte da consulta, ou se foi desconsiderada antes da submissão da consulta. Já nas consultas cuja verificação será da característica de utilização de dicionários - expansão da consulta, será verificado se novas palavras foram adicionadas a

[11] https://bitlybr.com/IWCxgkl
[12] https://bitlybr.com/xUWnt

consulta e assim aumentando os resultados disponibilizados ao usuário. Nas consultas que se deseja verificar que as palavras-chave da consulta, ou uma fração delas pode ser entendida como uma função de agregação da linguagem *SQL* será verificado se está alteração foi realizada e o resultado é uma computação de algumas linhas em um único resultado e esta linha será disponibilizada, quando possível, em razão do campo __search_id. Para as consultas que se deseja verificar o entendimento da técnica em razão da junção de várias palavras, como uma expressão será verificado se a técnica foi capaz de entender e processar a consulta em função da expressão. Será verificado também a combinação de algumas destas características.

Os resultados esperados são baseados no conhecimento e análise dos bancos de dados a partir da semântica esperada. Todos os resultados esperados que podem ser verificado nos arquivos, especificados em cada consulta por palavras-chave, serão considerados relevantes. Esse conjunto será a pilar para execução de todas as métricas.

4.3 Paralelo com outras Abordagens

Após expostas as principais características do *benchmark* proposto a Tabela 4.4 apresenta uma visão geral em razão aos *benchmark* existentes.

Tabela 4.4: *Paralelo entre benckmarks*

	Coffman	**Oliveira Filho**	**Benchmark Proposto**
Bancos de Dados	*Mondial* *IMDB* *Wikipédia*	*Mondial* *IMDB* *DBLP* *Northwind*	*Mondial* *IMDB* *DBLP*
Qtd. de consultas	50	50	50
Características de Consultas	Não especificado	Não especificado	4 grupos *Stopword* Expansão da Consulta Funções de Agregação Segmentação e Frase
Resultados esperados	Definidos	Não definidos	Definidos
Semântica da Consulta	Não definidos	Definidos	Definidos
Avaliação de relevância	Relevância Binária	Relevância Binária Verificação de Configurações	Dependente da semântica e dos resultados esperados
Métricas	Top-1 resultados relevantes *Reciprocal Rank* *Average Precision* *Mean Average Precision* Tempo de Execução Tempo de Resposta	Top-1 resultados relevantes Precisão em n(P@n) *Reciprocal Rank* *Mean Reciprocal Rank* *Average Precision* *Mean Average Precision*	Precisão Média Simples da Precisão Revocação Média Simples da Revocação *F-measure*

Analisando a tabela percebe-se que os bancos de dados *IMDB* e *Mondial* estão presente em todas propostas. Percebe-se a necessidade de bancos de dados com complexidades diferentes em número de relações, número de colunas ou quantidade de tuplas. A quantidade de consultas como orientado na comunidade de Banco de Dados permanece a mesma, uma quantidade mínima de 50 para descrever cenários diferentes.

Diferentemente dos *benchmarks* existentes as consultas são separadas em 04 grupos em função da característica a ser avaliada, o processamento semântico que deseja verificar. Uma consulta pode estar em um ou vários grupos de características ao mesmo tempo e os grupos não possuem um número fixo de consultas para cada característica. Este é o principal diferencial em razão dos *benchmarks* apresentados no Capítulo 3, porque enquanto um *benchmark* define o resultado esperado para cada consulta sem definir a semântica e outro define a semântica para cada consulta sem definir o conjunto de resultado esperado, nesta proposta será apresentado tanto a semântica quanto o resultado esperado.

A relevância é a característica que possui maior variação em relação às propostas existentes e é dependente do julgamento do usuário. Para incluir uma tupla no conjunto de resultados esperados serão analisados quais tratamentos das características desejadas foram realizados. Desta forma, a relevância será verificada em relação aos resultados esperados e dependente da semântica.

As métricas serão aplicadas em contexto comparativo entre os resultados esperados e os resultados retornados. O tempo de resposta e execução ou tempo de resposta não são alvos deste trabalho, pois o formato da apresentação dos resultados das técnicas avaliadas são divergentes, impossibilitando tal comparação e a necessidade de intervenção do usuário, em uma das técnicas avaliadas interfere nesta medição de tempo.

4.4 Considerações Finais

Neste Capítulo foi apresentado especificamente o *benchmark* proposto e seus componentes. Inicialmente quais bancos de dados foram selecionadas, seu tamanho em *byte* e em número de registros. Posteriormente, apresentadas as metodologias de definição das consultas a ser utilizada, e com quais as características semânticas se buscavam palavras ou conjunto de palavras-chave. As métricas a serem utilizadas foram definidas em seguida, mas destaca-se que a não necessidade de um ranqueamento e sim da construção de um conjunto de resultados. Após todas essas definições, contruí-se os resultados esperados e a semântica para cada consulta e para finalizar novamente uma verificação de pontos de equidade ou não, mas agora para os *benchmarks* atuais e o proposto neste trabalho.

CAPÍTULO 5

Contexto dos Experimentos e Análise dos Resultados

Este Capítulo apresenta como foram elaboradas as consultas e a semântica na seção 5.1, demonstra ainda o ambiente em que foram realizados os experimentos na seção 5.2 detalhando as técnicas utilizadas e as configurações necessárias. Assim como, os bancos de dados e alterações que neles foram realizadas e posteriormente cada técnica se comportou em cada banco de dados em relação a cada característica semântica. Além disso, promove uma discussão sobre os resultados na seção 5.3 em três perspectivas diferentes que seria dos bancos de dados, das técnicas de consulta por palavras-chave a bancos de dados relacionais e na perspectiva do processamento das características semânticas.

5.1 Construção das Consultas

A fim demonstrar a metodologia utilizada na construção das consultas as etapas desta construção, a definição da palavra e a lógica da semântica definida serão apresentadas a seguir. A captura dos tuítes ocorreu de novembro de 2019 até Maio de 2020 e as definições das consultas e da semântica aconteceu de Janeiro a Maio de 2020, então as palavras mais buscas são referentes a esse período. Na construção das consultas pretendia-se simular a ambiguidade comum em consultas construídas pelo usuário final.

5.1.1 IMDB

Conforme descrito na seção 4.1.2, houve o acompanhamento em várias datas do *twitter* e relação a *#imdb*, mas a maior parte dos tuítes se tratavam de uma classificação "*I rate ...* ". Para exemplificar esse grande número de tuítes, a consulta número 30 - i rate "*The Lion King*" foi acrescentada na lista de consultas - e a mesma possui a semântica de "Classificação do filme *The Lion King*". Outra consulta retirada de um tuíte foi a de número 33 - *cast of "Jack and Jill"* - o tuíte original - "*I applied to be in the cast of Jack*

and Jill on IMDB and they accepted me. This is the best day of my life!" - e a semântica aplicada foi Todos os participantes do filme *Jack and Jill*.

Além destes tuítes com o nome de filmes como *Star is Born*, *Star Wars*, *Fast and Furious*, ou séries como *Game of Thones*, *Yes prime Minister*, *Supernatural*, *Smallville*, *Vikings* apareciam frequentes então estas palavras foram submetidas no *site Keyword Tools*, em todas as plataformas, em língua Inglesa, os assuntos mais buscados sejam como *Keyword Suggestios*, *Questions* ou *Prepositions* eram novamente verificados. Desta forma a consulta número 12 "*actor dead Fast and Furious*", ou a número 19 - "*episode of Yes, Prime Ministrer*", ou a consulta 29 "*directed a star is born*", ou a consulta 08 "*game of thrones actors*". Todas elas foram construídas utilizando os dois contextos, o *twitter* e o *site Keyword Tools* e a semântica foi construída com o resultado da submissão das palavras-chave no *site* de buscas *google*.

Na tentativa de encontrar outras possibilidades, o *site* do *IMDB* foi acessado, e encontrado uma lista de palavras mais buscadas[1]. Ao acessar esta lista palavras algumas chamaram a atenção. A consulta 21 foi construída com palavras exatamente como encontradas no site - "*good versus evil*", mas a semântica não foi construída usando o resultado da pesquisa no próprio site do *IMDB*, pois tratava-se da descrição dos filmes ou séries, campo esse não constante no banco de dados utilizado, e percebendo a existência de um filme com exatamente este nome, colocou-se como resultado desejado o próprio filme. Outras palavras como *action hero*, *anime*, *musical number*, *based on...*, *dark hero*, *directed by star*, *husband wife relationship*, *superhero*, *police detective* foram combinado novamente com outras palavras do *site Keyword Tools*.

5.1.2 DBLP

Em relação aos demais bancos de dados, apesar de ter encontrado poucos *tuítes #dblp*, alguns dos encontrados foram utilizados. A consulta 26 *research on "open data"* é originário do tuíte "*Interesting research on open data"* e como semântica "Os artigos ou *proceedings* que possuem no título ao expressão dados abertos". A consulta 18 - *the 2019 proceedings indexed ieee* - originária do tuíte "*@IEEESSP any reason why the 2019 proceedings are not indexed by @dblp?*" como semântica foi definido "todos os proceedings de 2019 e indexados pelo IEEE".

Outras consultas como a Consulta 02 *author profile* - originária do tuíte "*@amyjko @amyjko Congrats and thanks for sharing! Hey @dblp_org, can you maybe reflect this in Amy's author profile*" e possui como semântica "a lista de todos os autores". Ou a Consulta 35 "*how many paper have been published by ACM in* 2019" originária do

[1] https://www.imdb.com/search/keyword/

tuíte - "*@APierantonio: This is a scandal! Look at how many papers have been published by IEEE Access in 2019*" e como semântica "a quantidade de publicações aconteceram pelo IEEE em 2019".

Após a utilização de vários tuítes as palavras, ou o conjunto de palavras mais repetidas foram conduzidas ao *site Keyword Tools* para verificar quais associações com outras palavras poderiam ser realizadas, para finalizar a lista que segue abaixo das consultas e semânticas definidas para esta base. Como a Consulta 06 *conference proceedings publishers* e definido como semântica "As conferencias de *proceedings* publicadas", ou a Conculta 11 - *journal article without volume number* e como semântica "Os *articles* que possuem jornais relacionados sem número de volume".

Neste banco de dados foi utilizado somente os tuítes e as combinações de palavras frequentemente utilizadas nos tuítes e o *site Keyword Tools*, e para a semântica o sentido dos tuítes, e nos casos da utilização do *site Keyword Tools* o resultado da submissão das palavras-chave no *site* de buscas *google*.

5.1.3 Mondial

Conforme descrito na seção 4.1.2 este banco de dados não foi possível encontrar nenhuma consulta através do twitter. O nome *Mondial* está relacionado a uma marca, então foi utilizado o *site World Atlas*[2] para encontrar palavras que pudessem compor as consultas. No site há diversas páginas apresentado as cidades, os estados, os países, a população, os oceanos, a economia, a religião. As palavras como "*populations cities and countries*", "*oceans details*" ou "*bodies of water rivers, seas and more*" foram encaminhados ao *site Keyword Tools* e então encontradas *Keyword Suggestios*, *Questions* ou *Prepositions* que refletiam as mais buscadas no *google*.

No *site World Atlas* foi possível construir consultas e semântica como a Consulta 13 - *oceans name* - Lista com nome de todos os oceanos, ou a Consulta 35 - *country with most borders* - País que possui a maior extensão de fronteira, assim como a Consulta 40 - "*largest continent population in the world*" - "O continente mais populoso do mundo".

A consulta 21 "*country largest population in the world*" foi construída após a inserção das subpalavras *populations cities* no *site Keyword Tools*. E após a construção da consulta foi verificado o resultado desta pergunta no próprio *site World Atlas* onde foi possível definir a semântica. A consulta 26 *USA States name them* foi outra consulta construída com a submissão das palavras *states name* no *site Keyword Tools* e um dos muitos resultados foi a junção das palavras-chave apresentadas.

[2]https://www.worldatlas.com/aatlas/world.htm

A fim de demonstrar a ambiguidade da construção das consultas foram criadas consultas com apenas uma palavra como a consulta 9 - Tocantins que o resultado poderia ser ou um estado do Brasil ou um Rio na tentativa de demonstrar a ambiguidade.

5.2 Contexto dos Experimentos

Para execução dos experimentos foram criadas máquinas virtuais no *software VMware Worstation 15Pro*, em uma máquina física que possui 4 núcleos de processador *Intel Core(TM) i7-6500U CPU 2.50 GHz*, 16 *GB* de memória *RAM*, disco rígido *SSD* de 480 *GB* e outro disco rígido de 1 *TB*, Sistema Operacional *Windows* 10 *Professional* de 64 *bits*. Os recursos disponíveis para as máquinas virtuais são os mesmos, isto é, 2 núcleos de processadores, 8 *GB* de memória *RAM* e 100 *GB* de disco rígido.

Os experimentos foram realizados com prazo limite de 60 minutos para execução da consulta e com os recursos computacionais disponíveis. As consultas que não apresentaram resultados dentro deste prazo são interrompidas. Neste caso, a técnica ainda estava realizando processamentos, mas resultados não foram apresentados. Há também consultas que por motivos não explícitos a própria técnica finalizou a execução e também não apresentou resultados. Há consultas que as técnicas finalizaram, mas por motivos individuais de seu processamento não apresentou resultados, ou para outros casos foi apresentado mapeamentos em *SQL*, esses mapeamentos não eram capazes de retornar algum registro do banco de dados.

5.2.1 Técnicas Selecionadas

As técnicas foram selecionadas por disponibilizarem versão para execução e testes, seja por um arquivo executável ou através dos códigos fonte. Buscamos o maior número possível de técnicas, considerando tanto as que implementam grafos de dados quanto aquelas que usam grafo de esquema. Mesmo após contatos a autores de outras técnicas, além das técnicas que serão utilizados, não foram encontradas ou disponibilizadas versões para execução com possibilidade de inserção e adaptação dos bancos de dados selecionados.

A técnica *BANKS*[3] disponibiliza uma versão na *web* que poderia ser utilizada, mas os testes só poderiam ser executados nos bancos de dados disponibilizadas pela técnica e que estão com versões defasadas (*DBLP* - 2005 e 2009; *IMDB* e *IIT Bombay ETD Database* sem registro da data de inclusão). Não é possível baixar o banco disponibilizado para utilização e não é possível inserir novos ou atualizar os bancos de dados disponíveis.

[3] http://www.cse.iitb.ac.in/banks/banks-demo/SearchForm

Diante disto, foram analisadas três técnicas (*Banks-II* ou *Bidirectional*, *Keymantic* e Ramada) com características diferentes. O *Banks-II*, conforme visto no Capítulo 2, implementa um grafo de dados e possui acesso prévio aos dados. Já *Keymantic* e Ramada et al. implementam grafo de esquema e não possuem acesso prévio aos dados. Apesar das duas últimas técnicas implementarem o mesmo tipo de grafo, há particularidades significativas para a avaliação de cada uma delas.

5.2.1.1 *Banks-II*

Uma versão do código fonte está disponível no *GitHub*[4] e implementa o algoritmo *Dijkstra* em Linguagem *Java*. Para funcionamento, foi necessária a instalação do *JDK* versão 1.8.0_201 e do Sistema Gerenciador de Banco de Dados *PostgreSql*[5] versão 9.6. É possível, caso seja necessário, editar o código-fonte ou somente executar o arquivo *.jar* disponibilizado.

Na construção desse código, foi considerado que a avaliação seria feita sobre os bancos de dados propostos por [16]. Desta forma, é necessário que os bancos de dados deste estudo possuam a coluna especificada. Ao submeter uma consulta para um banco que não possua a coluna *__search_id*, é apresentado um erro informando que a coluna não existe e a submissão não prossegue.

Os fragmentos de código *Java* que explicitam a necessidade desta coluna no banco de dados pode ser verificado no trecho de código apresentado no código I.1 no Anexo I linha 7. Da mesma forma, os arquivos de criação de esquema do banco de dados disponibilizado por [16] podem ser visto no fragmento de código na linguagem de consulta no código I.2 Anexo I.

A execução da técnica propriamente dita é simples e intuitiva, ao iniciar a execução, seja pelo código em uma IDE, seja por meio do *jar* disponível, são solicitadas as informações de conexão com o banco de dados como nome de usuário, senha, local, porta e nome do banco de dados. Ao iniciar a execução, a consulta deve ser inserida e submetida ao banco de dados. A técnica imprime na tela as ações que foram realizadas gradativamente.

A técnica acessa o banco de dados e todas as tuplas são carregadas na memória *RAM* e, então, a verificação de igualdade da consulta em relação aos registros do banco de dados é iniciada. A técnica busca por registros iguais à consulta submetida ou que tenham mais palavras do que as dispostas na consulta. A consulta não é fracionada e a ordem das palavras-chave, se houver mais de uma, também não é alterada e o conteúdo da consulta deve estar em apenas uma coluna. Se na consulta houver aspas duplas, elas são tratadas

[4] https://github.com/prof18/banks
[5] https://www.postgresql.org/

como texto a ser encontrado no banco e não como um gatilho que poderia identificar o uso da característica Segmentação e Frase, descrito na seção 4.1.2.4. Não é verificado em nenhum momento se a consulta submetida é parte ou é um metadado do banco de dados.

Para exemplificar esse processamento observem-se as consultas para o banco de dados *Mondial* de número 09 - "Tocantins" e 13 - "*Oceans name*" conforme pode ser observado no Anexo I e código I.3 e código I.4. Na Consulta 009 possui como semântica "Informações sobre o Rio Tocantins, ou sobre o estado de Tocantins" e esperava-se encontrar dois registros, identificados pelo *__search_id* = 17319 e *__search_id* = 19089. Ao se submeter tal consulta os resultados esperados foram retornados, mas a técnica não realizou todo o processamento desejado.

Da mesma forma como na consulta 013 que possui como semântica "lista com nome de todos os oceanos" e esperava-se que a técnica realizasse uma expansão da consulta adicionando a palavra *sea* que é um sinônimo de *ocean*. Neste caso, retornaria a tabela *sea* completa, mas a técnica não retornou nenhum registro pois buscava exatamente "*Oceans name*".

5.2.1.2 *Keymantic*

A técnica *Keymantic* disponibiliza uma versão executável disponível no *Source Forge*[6]. Para utilização desta versão, foi necessário o *Java* versão 1.8.0_201 x86 e o Sistema Gerenciador de Banco de Dados *MySQL* versão 5.0.22. Conforme especificado em sua documentação, é necessário criar um arquivo em *XML* com o esquema de relacionamento das tabelas no banco de dados, conforme o Código II.1 disposto no Anexo II, e, outro arquivo contendo o detalhamento de cada tabela e suas relações, conforme o Código II.2 disponibilizado no Anexo II.

Essas informações são essenciais para a devida execução da técnica e a formação do grafo de esquema e o direcionamento de cada consulta. Outra configuração necessária é a inserção de informações sobre usuário e senha do banco de dados, assim como as informações do *driver* de conexão com o banco de dados no arquivo *costConfiguration.properties*.

Os arquivos II.1 e II.2 são utilizados na verificação de tipo de dados das colunas para viabilizar a associação com os valores submetidos na consulta e os encontrados no banco de dados. A técnica fraciona as palavras-chave submetidas na consulta e, se aspas duplas forem incluídas, todas as palavras em seu interior são consideradas uma única palavra. A técnica associa inicialmente as palavras-chave aos metadados, mais especificamente aos nomes das colunas. Como pode haver mais de uma palavra-chave na consulta as palavras que não são mapeadas como metadados são associadas às tuplas

[6]https://sourceforge.net/projects/keymantic/

do banco. A associação da palavra-chave é feita baseada no tipo de dado da coluna da tabela descrito no arquivo *xml* de configuração da tabela. Para uma mesma consulta são gerados vários mapeamentos considerando cada palavra em colunas diferentes.

Para exemplificar o processamento, tome-se a consulta para o banco de dados *Mondial* de número 09 "Tocantins" conforme pode ser observado no Anexo II e código II.3, uma fração do resultado apresentado. Na Consulta 009, que possui como semântica "Informações sobre o Rio Tocantins, ou sobre o estado de Tocantins" a técnica não identificou a consulta como um valor de metadados. Então, a consulta foi associada aos valores de registro gerando, em alguns casos, vários mapeamentos explicitados na linguagem *SQL*. Tais mapeamentos diferem-se somente em qual coluna está sendo associada a palavra-chave naquele momento.

5.2.1.3 Ramada

O código-fonte foi disponibilizado pelos próprios autores. Para a execução desta técnica são requeridos a *IDE Eclipse*, *Java* versão 8_201, *JDK* versão 1.8.0_20, Sistema Gerenciador de Banco de Dados *MySQL* versão 8.0, *conector Mysql* versão 5.1.47 e dicionário de sinônimos *WordNet* 2.1. Conforme a especificação, é necessário criar uma tabela, definida como TME, disposta no Anexo III e Código III.1, que armazena as informações sobre os possíveis bancos de dados a serem verificadas. Como resultado final, é construído um arquivo de nome *output* com as informações geradas no processamento e mapeamento da consulta. Este arquivo contém, inclusive, as expressões *SQLs* geradas e os resultados retornados por estas expressões.

A técnica foi projetada com a possibilidade de verificar se a consulta submetida pode encontrar informações úteis em mais de um banco de dados. Esta verificação é realizada com as informações dispostas na tabela TME, e após o acréscimo de sinônimos adquiridos através do dicionário *WordNet*. As palavras encontradas como sinônimo, hiperônimo e radicais são adicionadas ao conjunto de palavras que serão submetidos inicialmente a tabela TME, se uma destas palavras estiver na coluna *dc_description*, *dc_subject*, *dc_title* e *dc_identifier* da tabela TME, o banco de dados é considerada útil. Somente após ser considerada útil o banco de dados é acessado. Um mapeamento das tabelas é realizado a fim de identificar as relações, e tipos de dados para construção dos mapeamentos.

5.2.2 Bancos de Dados

Ao iniciar a execução dos experimentos nas técnicas, observou-se que para os bancos de dados *IMDB* e *DBLP* não havia resultados retornados. Em função dessa situação, optou-se por reduzir o banco de dados para que os experimentos fossem

possíveis. Em relação ao banco de dados *IMDB*, a redução baseado na coluna *startyear* da produção que será a partir de 1980 e constará somente produções do tipo *movies* e *tvSeries*. A partir da redução da tabela *movies* de 6.579.412 para 473.410 registros as demais tabelas foram atualizadas contendo apenas informações que estavam associadas aos filmes. Então há somente pessoas que hora podem associar-se como atores ou diretores, assim como os episódios e gêneros que podem ser associados.

A Tabela 5.1 abaixo, apresenta as quantidades para todas as tabelas e o percentual de redução para cada tabela. O conjunto de resultados esperados foi mapeado usando o banco reduzido e como as consultas foram construídas com o auxílio de *sites* que refletem as informações mais buscadas em 2019 e 2020, a retirada dos filmes com ano de início anterior a 1980 não casou grande impacto.

Tabela 5.1: *Tabela da fração banco de dados IMDB.*

Banco	***Tabelas***	***Quantidade tuplas originais***	***Quantidade tuplas atual***	***% de redução***
IMDB	*episode*	2.866.700	2.416.136	15,7%
	genres	28	28	0%
	movies	6.579.412	473.410	92,8%
	movies_actors	15.399.455	518.460	96,6%
	movies_directors	4.379.357	316.156	92,7%
	movies_genres	10.014.086	653.459	93,4%
	person	1.001.486	392.532	99%
	Total	***40.240.524***	***4.344.112***	***89,2%***

Em relação ao banco de dados *DBLP*, foi utilizado como parâmetro para redução dos dados a coluna *year* das tabelas *article*, *inproceedings*, *proceedings* e *book*. O ano selecionado foi 2010, então as tabelas só possuem registro cujo ano de publicação seja maior ou igual a 2010. Na Tabela 5.2 consta a quantidade de registros originais e a quantidade de registros com o banco de dados reduzida; houve uma redução de 55,9% do número de registros mas em algumas tabelas não houve redução.

Tabela 5.2: *Tabelas da fração do banco de dados DBLP*

Banco	***Tabelas***	***Quantidade tuplas originais***	***Quantidade tuplas atual***	***% de redução***
DBLP	*article*	2.166.665	1.320.630	39%
	author	651.005	479.733	26%
	author_article	8.708.567	2.935.503	66,2%
	book	17.922	6072	66,1%
	booktitle	13.952	8403	39,7%
	inproceedings	2.526.762	1.439.722	43%
	proceedings	81.047	46.622	57,5%
	publisher	2.162	2.162	0%
	series	1.514	1.514	0%
	Total	***14.169.596***	***6.240.361***	***55,9%***

O banco de dados *Mondial* não sofreu nenhuma alteração.

5.2.3 Características Semânticas

A observação sobre o processamento ou não das características semânticas foi realizada individualmente, pois foram inseridas nas consultas, situações que exigiria da técnica um pré processamento e não somente realizar a busca da palavra em relação aos metadados ou aos próprios valores do banco de dados.

Exemplificando na consulta 13 do banco de dados *Mondial*, que pode ser observado no Apêndice A, Tabela A.2 e que possui como semântica "lista com nome de todos os oceanos" for observado somente as palavras bastava encontrar a Tabela "*ocean*" e uma coluna "*name*". Mas não existe uma tabela de nome *ocean* e sim *sea*. Logo, foi posto como pretensão que a técnica utilizasse uma expansão da consulta pois segundo o dicionário Colins[7] *sea* é sinônimo de *ocean*. Então seria possível encontrar a tabela e retornar todos os registros da mesma.

Outro exemplo agora para o banco de dados *IMDB* consulta 10 "*Steven Spielberg directoring*" com semântica pretendida "Todos os filmes dirigidos por *Steven Spielberg*", que pode ser observado no Apêndice A, Tabela A.1, onde deveria ser realizado a radicalização da palavra *directoring* para *director* e assim consegui associar os filmes ao diretor.

[7]https://www.collinsdictionary.com/pt/dictionary/english/ocean

5.2.3.1 *Stopwords*

Conforme explicitado na seção 4.1.2 espera-se que as técnicas sejam capazes de ignorar determinadas palavras inseridas intencionalmente e que não agregam valor semântico a consulta.

- *Banks-II*

A técnica não realiza interferência em relação à consulta submetida nos bancos de dados. A técnica verifica a correspondência exata da consulta e valores existentes no banco de dados ou se a consulta completa é uma subpalavra de algum registro do banco de dados, conforme descrito na seção 5.2.1.1. Se a consulta for composta por várias palavras-chave, a técnica procura identificar a existência do conjunto de palavras-chave na ordem apresentada. Em razão da estratégia adotada pela técnica as *stopwords*, não foram retiradas em qualquer das consultas submetidas para qualquer dos bancos de dados utilizadas no experimento.

No banco de dados *DBLP*, a técnica não foi capaz de finalizar as consultas no tempo e com os recursos computacionais supracitados. Das consultas apresentadas na Tabela A.3 de número 14 a 35, somente para a consulta 15 a técnica foi capaz de encontrar um registro no banco que atendesse ao solicitado, mas não finalizou a execução no tempo estabelecido.

Para o banco de dados *IMDB* e as consultas apresentadas na Tabela A.1 de número 15 a 34, a técnica não conseguiu processar as consultas no tempo previamente estabelecido e com os recursos computacionais disponíveis.

E, por fim, em relação ao banco de dados *Mondial* e as consultas apresentadas na Tabela A.2 de número 17 a 41, a técnica não foi capaz de processar as consultas no tempo previamente estabelecido e com os recursos computacionais disponíveis.

- *Keymantic*

Ao iniciar a execução, a técnica fraciona as palavras da consulta, tratando-as individualmente. Se uma das subpalavras for associada a uma coluna ou tabela, as demais palavras a serem verificadas são comparadas preferencialmente a esta coluna ou tabela. Todas as palavras são utilizadas nos mapeamentos gerados e, apesar do grande número, a quantidade de resultados retornados destes mapeamentos é pequena. Para os bancos de dados *DBLP*, *IMDB* e *Mondial* e apesar de todas serem disponibilizados mapeamentos após a execução dos *SQLs* nenhum resultado foi retornado.

- Ramada

A técnica não exclui qualquer palavra inserida na consulta, e as trata sempre como subpalavras fracionadas. Para o banco de dados *DBLP*, das consultas apresentadas na Tabela A.3 de número 14 a 35, somente as consultas 22 e 23 tiveram o processamento realizado dentro do tempo previsto, as demais consultas foram finalizadas por exceder a razão do tempo e os recursos computacionais disponíveis.

Para o banco de dados *IMDB*, as consultas apresentadas na Tabela A.1 de número 15 a 34, somente a consulta de número 23 foi finalizada e com resultados retornados, para as demais consultas a técnica não conseguiu realizar o processamento no tempo previamente estabelecido e com os recursos disponíveis.

Para o banco de dados *Mondial* e as consultas apresentadas na Tabela A.2 de número 17 a 41, nenhuma delas foi finalizada até o prazo máximo previamente estabelecido e com os recursos disponíveis. Para as consultas que foram finalizadas em qualquer uma das bancos o processamento desejado de retirada de *stopwords* não foi realizado.

5.2.3.2 Expansão da Consulta

Conforme explicitado na seção 4.1.2 espera-se que a técnica seja capaz de observar os sinônimos e verificar a construção léxica das palavras e adicionar estes novos termos a consulta.

- *Banks-II*

Conforme descrito na seção 5.2.1.1, a técnica não realiza alteração da consulta, tampouco expande a consulta ou realiza verificações em relação a sinônimos ou radical das palavras. Para o banco de dados *DBLP* e *IMDB*, a técnica não conseguiu processar as consultas no tempo e com os recursos computacionais disponíveis. Para o banco de dados *Mondial*, a técnica foi capaz de apresentar resultados para algumas consultas e alguns desses resultados ditos como relevantes. Para as consultas 4, 5, 7, 8, 15 e 16, apresentou algum ou todos os resultados esperados, mas também apresentou resultados que não eram esperados. Para a consulta 6, todos os resultados retornados não estavam presentes no conjunto de resultados esperados.

- *Keymantic*

A técnica não realiza acréscimos para as consultas. Para o banco de dados *DBLP* em duas das 16 consultas, a técnica apresentou erro e não conseguiu gerar nenhum mapeamento para disponibilização dos *SQLs*, e nenhuma das 14 consultas resultantes apresentou resultado. Para o banco de dados *IMDB*, mesmo não apresentando o processamento desejado, das 19 consultas apenas uma apresentou resultados que estavam contidos no

conjunto de resultados esperados. Para o banco de dados *Mondial*, a técnica, mesmo não realizando o processamento desejado, apresentou resultados que estavam no conjunto de resultados esperados em 8 diferentes consultas - 4, 5, 7, 8, 9, 10, 15 e 16. Para a consulta 13, os resultados apresentados não estavam no conjunto de resultados esperados.

- Ramada

A técnica realiza expansão da consulta para qualquer consulta que seja submetida, com a utilização do dicionário de sinônimo *WordNet*. Apesar de realizar a expansão da consulta em relação aos sinônimos, não realizou a radicalização das palavras. Em todas os bancos *DBLP*, *IMDB* e *Mondial* e todas as consultas a expansão foi construída, mas em alguns casos não foi suficiente para atender a semântica da consulta.

5.2.3.3 Função de Agregação

Conforme explicitado na seção 4.1.2, espera-se que a técnica identifique e substitua uma ou um conjunto de palavras por uma função de agregação em linguagem *SQL*

- *Banks-II*

Conforme descrito na seção 5.2.1.1 a técnica não realiza alteração na consulta submetida, e para os bancos de dados *DBLP*, *IMDB* e *Mondial* a técnica não conseguiu processar as consultas no tempo e com os recursos computacionais disponíveis.

- *Keymantic*

A técnica não foi capaz de realizar o entendimento e a substituição de algumas palavras por uma função de agregação em SQL. Conforme exposto anteriormente, não há acréscimos ou alteração nas consultas submetidas. Para os bancos de dados *DBLP*, *IMDB* e *Mondial*, vários mapeamentos foram criados e expressões *SQLs* disponibilizadas, mas ao serem executadas nenhum resultado foi retornado.

- Ramada

Antes de iniciar a execução dos comandos *SQL*, a técnica realiza um pré-processamento e um deles é a substituição de algumas palavras por funções de agregação em *SQL*. Inspecionando o código da técnica, é possível identificar o mapeamento das seguintes funções de agregação: máximo, mínimo, média, soma, contagem, agrupamento e ordenação. Esse mapeamento é feito através da identificação de algumas palavras, por exemplo, para mapear um valor máximo obtido pela função de agregação *SQL max*,

é necessário encontrar uma das palavras *higher*, *highest*, *maximum*, *maximal*, *larger*, *largest*, *greater*, *greatest* ou *max* na consulta.

Para o banco de dados *DBLP* e as consultas referentes as funções de agregação, somente uma consulta conseguiu ser finalizada, mas as palavras usadas na consultas não estavam no conjunto de palavras relativas as funções de agregação, então, o *SQL* e o resultado retornado não estavam no conjunto de resultados esperados.

Para o banco de dados *IMDB*, apesar de finalizar a execução das consultas 13, 14, 35, 37, 39 e 41 em nenhum dos processamentos houve a utilização de função de agregação na geração dos resultados retornados e nenhum resultado esperado foi gerado.

Para o banco de dados *Mondial* e as consultas referentes a função de agregação, somente a consulta 46 foi finalizada e o *SQL* gerado continha uma função de agregação. O conjunto de resultados retornado não fazia parte do conjunto de resultados esperados, mas o processamento que mapeia palavras para uma função de agregação em *SQL* foi realizado. Outra consulta, a 33, apesar de não finalizada constrói alguns mapeamentos utilizando a função de agregação *max*, que era exatamente o processamento desejado - "*select distinct max(borders.country1),count(ies) from borders where borders.country1 like '%in%' AND borders.country2 like '%territory%'* ".

5.2.3.4 Segmentação e Frase

Conforme explicitado na seção 4.1.2 espera-se que a técnica processe não individualmente as palavras, mas como um conjunto ou um subconjunto em relação a consulta original.

- *Banks-II*

Conforme descrito na seção 5.2.1.1 a técnica não realiza alteração na consulta submetida, e para aos bancos de dados *DBLP* e *IMDB*, a técnica não conseguiu processar as consultas no tempo e com os recursos disponíveis. Para o banco de dados *Mondial*, apesar de 03 das 19 consultas forem finalizadas e apresentarem resultado o processamento necessário não foi realizado. Somente nas consultas nas quais todas as palavras-chave precisavam ser entendidas como uma expressão, resultados que estavam no conjunto de resultados esperados foram retornados. A quantidade de registros do conjunto de resultados retornados foi maior, em número absoluto, que o conjunto de resultados esperados.

- *Keymantic*

Se na inserção da consultas, a palavra ou palavras que necessitam ser entendidas como uma frase estiverem entre aspas duplas, a técnica entende que essa expressão que é

composta por várias palavras e precisa ser entendida como uma palavra única. Para o banco de dados *DBLP*, *IMDB* e *Mondial*, quando o uso das aspas foi utilizado e o processamento foi realizado, mas nem sempre os resultados retornados estavam contidos nos resultados esperados.

- Ramada

A técnica não realiza pré-processamentos na intenção de identificar a consulta completa como uma expressão e que precisa ser considerada como uma palavra única ou que duas ou mais subpalavras da consulta possam ser processadas como expressão. Para o banco de dados *DBLP*, e apesar de conseguir finalizar as consultas de número 10, 11, 44, 45 e 50, e, as consultas 44 e 50 apresentarem resultados que estavam contidos no conjunto de resultados esperados, a técnica não foi capaz de realizar o processamento desejado. Para o banco de dados *IMDB*, a técnica foi capaz de conseguir apresentar resultados para as consultas 10, 11, 12, 32, 45 e 49 em nenhuma das consultas houve resultados e o processamento desejado não foi realizado. Para o banco de dados *Mondial*, a técnica não foi capaz de realizar o processamento no tempo previamente estabelecido e com os recursos disponíveis.

5.3 Análise dos Resultados

Nesta seção serão apresentados os resultados dos experimentos em três perspectivas diferentes: (1) dos bancos de dados; (2) das técnicas e (3) nas características semânticas. Inicialmente será explicitado como a submissão das consultas se comportaram em relação aos bancos de dados, seguido de como se comportaram os experimentos em relação às técnicas de busca e por fim como se comportaram em relação as características semânticas.

5.3.1 Perspectiva dos Bancos de Dados

Ao iniciar a execução dos experimentos, foi verificado que para os bancos de dados *IMDB* e *DBLP* as técnicas não eram capazes de produzir resultado algum. Para que as técnicas fossem executadas as bancos foram reduzidas conforme descrito na seção 5.2.2 e mesmo sendo reduzidas algumas técnicas não apresentaram resultados. Foram submetidas à cada banco, 50 consultas para cada técnica, observando de maneira geral em relação a banco foram 150 consultas. As consultas poderiam figurar em três ambientes diferentes: (1) Executadas - o que demonstra que a técnica concluiu a execução e apresentou resultados não importando se seriam relevantes ou não; (2) Falha - são as consultas para as quais as técnicas interromperam a execução sem apresentar resultado

algum; e, (3) Tempo - são as consultas que atingiram o limite de tempo de execução usando os recursos computacionais disponíveis.

A Figura 5.1 apresenta o número de consultas executadas, com falha ou interrompidas em razão do tempo para cada banco *IMDB* em azul, *DBLP* em laranja e *Mondial* em cinza. A inexistência da barra e, determinado ambiente explicita a não ocorrência da situação no banco de dados.

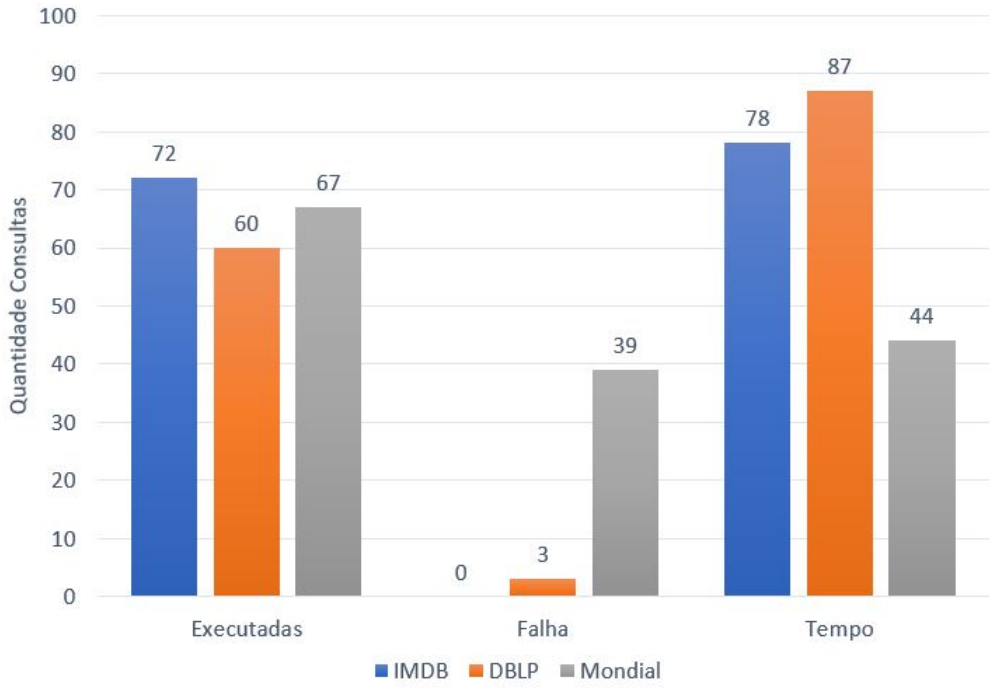

Figura 5.1: *Comportamento das Consultas por banco de dados.*

Pode ser observado que conforme a Figura 5.1, para o banco de dados *IMDB*, 72 consultas, ou seja, 48% do total foram executadas e as demais consultas 72, 52% atingiram o tempo máximo disponível e foram interrompidas e nenhuma das consultas houve falha . Para o banco *DBLP* temos que 60 consultas, 40% foram executadas enquanto 03 consultas, 2%, apresentaram falha e a própria técnica finalizou a execução da consulta sem apresentar resultados e outras 87 consultas atingiram o tempo máximo e foram finalizadas, ou seja 58% do total. Ao reduzir a quantidade de registos destes bancos de dados esperava-se que fossem obtidos melhores resultados.

Ainda verificando sobre a execução da consultas no banco de dados *Mondial* os resultados apresentados foram 67 consultas executadas - 45% permanecendo na média das consultas executadas, para as consultas que foram interrompidas em relação ao tempo a técnica obteve um resultado de 44 consultas, 29% e houve um aumento em relação as consultas que apresentaram falha em 39 delas, 26% do total.

A Tabela 5.3 apresenta de forma resumida as informações apresentadas na seção 4.1.1 para auxiliar nas conjecturas feitas a seguir.

Tabela 5.3: *Resumo dos bancos de dados.*

banco de dados	Qte Relações	Número Registros
IMDB	7	4.344.112
DBLP	15	6.240.361
Mondial	63	21.344

Esperava-se que após a redução dos bancos as execuções seriam mais bem sucedidas em relação ao conjunto de resultados esperados, mas isso não aconteceu. Mesmo fazendo as reduções em relação as linhas, simplificando-a em relação a possíveis relacionamentos, os resultados não foram melhores. Se considerarmos os bancos em relação a complexidade de relacionamentos entre as tabelas podemos observar que *Mondial*, o menor banco de dados em quantidade de registro, é o banco com maior complexidade e em número de relacionamento entre as tabelas - 63 o que não garantiu melhores resultados.

O banco de dados *IMDB* que, em relação aos demais é classificadas por quantidade de registros é o segundo maior banco de dados, suas consultas variaram somente entre executadas ou interrompidas em função do tempo e que também não ajuda. Visto que, é o banco com menor complexidade em relação às associações entre tabelas. Já o banco de dados *DBLP* que é o maior em quantidade de registros possui consultas nos três cenários.

Desta forma, se observamos apenas o número de consultas executadas o banco de dados *IMDB* possui o maior número de consultas executadas e o menor número de consultas finalizadas por falha, mas apresenta o segundo maior número de consultas finalizadas por tempo e coincidentemente este é o banco com menor complexidade. O banco de dados *DBLP*, segundo em relação a complexidade do banco, apresentou um menor numero de consultas executas, houve para este banco consultas que apresentaram falha e um número grande de consultas finalizadas por tempo.

Neste sentido, o banco de maior complexidade, *Mondial*, apresentou o maior número de consulta com falha, e quase o mesmo número de consultas não executadas em função do tempo. Desta forma é possível perceber que a complexidade afeta no desempenho. Porém esta análise esta sendo realizada apenas na perspectiva do exito da execução da consulta mas não quanto ao precisão dos resultados.

5.3.2 Perspectiva das Técnicas

Banks-II é uma técnica que considera que a consulta submetida será encontrada sempre em uma única coluna. Baseado nisso, quando se tem certeza de como os dados estão armazenados é uma técnica que recupera com assertividade os resultados esperados. A consulta também pode ser parte da informação contida em uma célula que a técnica

conseguirá encontrar, mas se houver ao menos uma palavra que difere da consulta, mesmo contendo a maior parte da consulta não apresentará como resultado válido. Quando a consulta submetida foi, por exemplo, a de número 15 disposta no Apêndice A, Tabela A.2 "*Arabian Desert*", a técnica encontrou exatamente o resultado esperado, mas se a consulta submetida for a de número 25, "*river in cuzco*" e houver a necessidade de associar duas tabelas, *River* e *City*, a técnica não é capaz de realizar essa associação e a fragmentação da consulta.

Um problema relevante é a geração de um grafo de dados com todas as tuplas do banco de dados na memória *RAM*. Ao iniciar a execução, a técnica carrega o banco de dados na memória *RAM* e incrementalmente constrói o grafo ou grafos dos possíveis resultados. A técnica foi capaz de apresentar resultados, mesmo que não relevantes, somente para o banco de dados *Mondial*. Para os demais bancos as consultas foram interrompidas em razão do tempo, mesmo disponibilizando 8 *Gb* de memória *RAM* e 1 hora de limite de tempo.

A Figura 5.2 tem o objetivo de apresentar como foi a execução das consultas nos bancos de dados, especificamente para a técnica *Banks-II*, conforme explicitado anteriormente as consultas foram classificadas em executadas, falha e tempo. São expostos os três bancos de dados e o número referentes a classificação de cada consulta e cada banco. No Apêndice B é possível observar individualmente o comportamento de cada consulta.

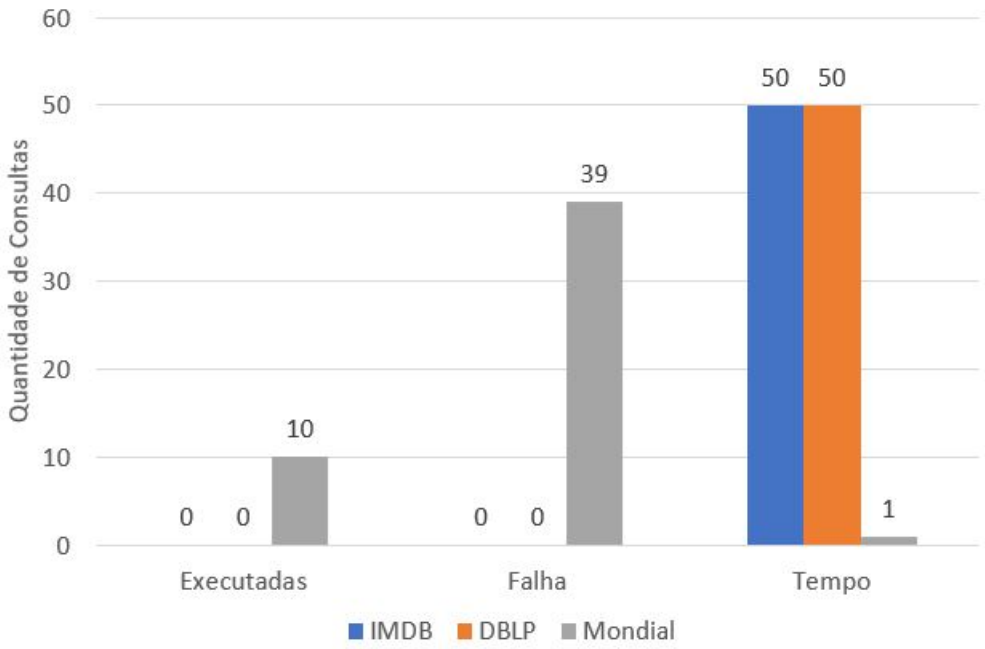

Figura 5.2: *Comportamento das Consultas Banks-II / banco de dados.*

E conforme pode ser observado para o banco de dados *IMDB* e *DBLP*, as consultas foram interrompidas em razão do tempo de execução. Para o banco de dados

Mondial há consultas classificadas como executadas, com falha e interrompidas em razão do tempo. Desta forma podemos concluir que, em razão da geração do grafo de dados na memória *RAM* a técnica conseguiu finalizar consultas apenas para o banco de dados, que neste contexto, possui a menor quantidade de registros, não importando que este seria o banco de dados de maior complexidade em relação às relações entre as tabelas.

Após a submissão das consultas, os resultados foram organizados e o conjunto de resultados retornados. Foi comparado consulta a consulta ao conjunto de resultados retornados e após a tabulação dos dados as métricas foram aplicadas. A Figura 5.3 é possível visualizar o comportamento da técnica em relação a precisão dos resultados apresentados do quais estes poderiam estar no conjunto de resultados esperados.

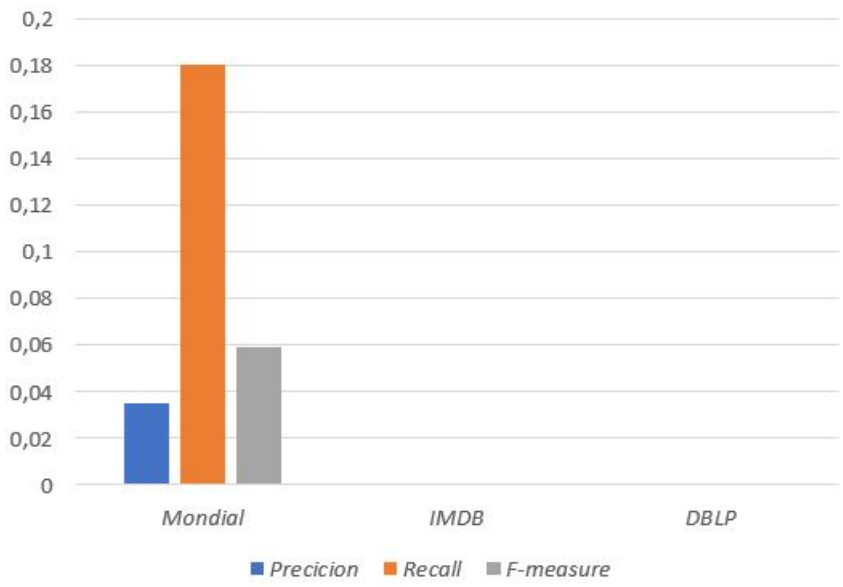

Figura 5.3: *Resumo Métricas Banks-II / banco de dados.*

Os resultados gerados para cada consulta do banco de dados e, que podem ser observados individualmente no Apêndice B e Tabela B.3, B.2 e B.1, a média simples da precisão ficou em 0,035134 enquanto a média simples de revocação em 0,179753 e a *f-measure* em 0,058779. Para os bancos de dados *IMDB* e *DBLP* não houve resultados recuperados que estavam no conjunto de resultados esperados. Desta maneira, para as consultas submetidas que o conjunto de resultados retornados não fosse vazio haviam resultados esperados e não esperados. Como a métrica revocação é calculada baseada no número de resultados recuperados relevantes e não relevantes a média simple da revocação é maior que a média simples da precisão pois sempre houve mais resultados retornados que os esperados.

Keymantic é uma das técnicas avaliadas que utiliza grafo de esquema e ao submeter uma consulta, cada subpalavra-chave é mapeada como um metadado, se for

encontrada a subpalavra não é buscada como valor no banco de dados. Utilizando o exemplo da Consulta 14 para o banco de dados *IMDB* encontrada no Apêndice A e Tabela A.1 "*best movies ratings* 2020", a palavra *movies* é mapeada como metadado, assim como a palavra *ratings*. Já as palavras *best* e 2020 são mapeadas como valor e são criadas associações com a tabela *person* e *genres*, na tentativa de encontrar os valores nestas tabelas. Observando que a semântica desejada é um conjunto com "melhores filmes classificados e lançados em 2020" e conhecendo a estrutura da tabela *movies*, percebe-se que não seria necessário a associação com as tabelas *person* ou *genres*. Se a mesma consulta for submetida utilizando o singular "*best movie rating* 2020", o processamento realizado é o mesmo.

A Figura 5.4 apresenta o comportamento das consultas especificamente para a técnica *Keymantic*, e conforme explicitado anteriormente as consultas foram classificadas em executadas, falha e tempo, conforme esclarecido na seção 5.3.1. São expostos os três bancos de dados e o número referentes a classificação de cada consulta e cada banco. No Apêndice B é possível observar individualmente o comportamento de cada consulta. Conforme pode ser observado para o banco de dados *IMDB* e *Mondial*, todas as consultas foram executadas. Para o banco de dados *DBLP* há consultas classificadas como executadas e com falha. Mas para nenhum dos três bancos de dados há consultas que foram interrompidas em função do tempo de execução.

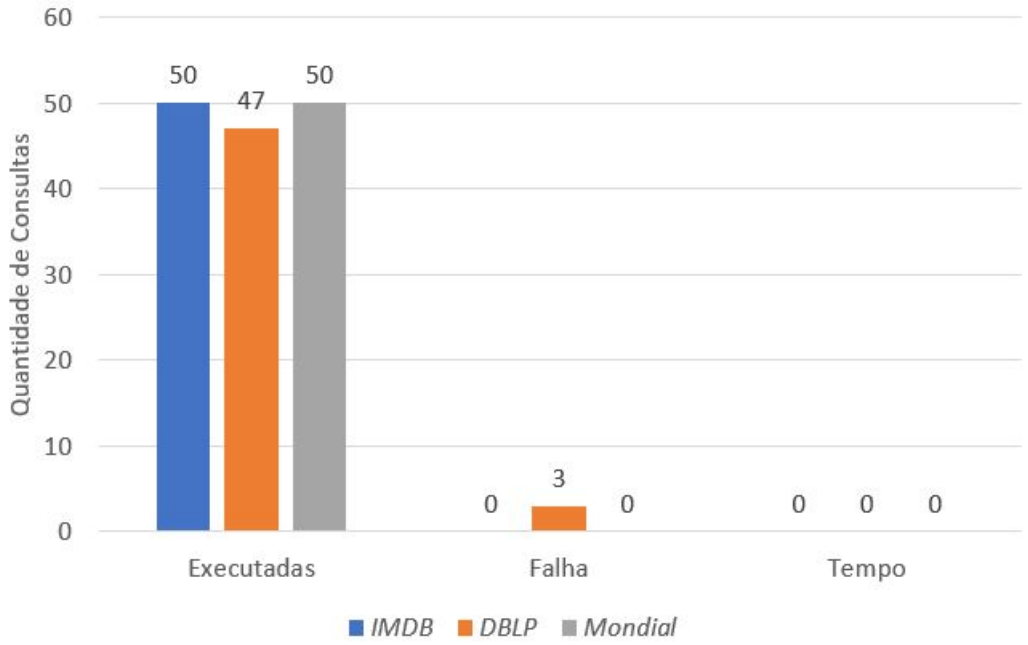

Figura 5.4: *Comportamento das Consultas Keymantic / banco de dados.*

Observando o comportamento da técnica *Keymantic*, são gerados mapeamento que associam as subpalavras da consulta aos metadados das tabelas ou aos valores de tu-

plas do banco dados. De acordo com o observado anteriormente, a execução da consulta não assegura que haverá conjunto de resultados retornados não vazios, e mesmo os conjuntos não vazios não assegura a precisão dos resultados. Diante do exposto, e observando que 147 das 150 consultas, ou seja, 98% do total foram executadas conjecturando-se uma possível precisão.

Na Figura 5.5 é apresentado a média de mapeamentos gerados por banco de dados, cada mapeamento pode ser resumido em uma consulta *SQL*. A quantidade de mapeamentos gerados por consulta pode ser observado no Apêndice B e Tabelas B.4, B.5, e B.6. Deste modo, percebe-se que são gerados muitos mapeamentos por consulta na tentativa de um dos mapeamentos apresentar a precisão necessária para a consulta, o que não se pôde observar como verdadeiro. Ao realizar um mapeamento a técnica explora a possibilidade de cada subpalavra assemelhar-se a uma coluna da tabela elegida como potencial, o que dificulta a precisão da técnica.

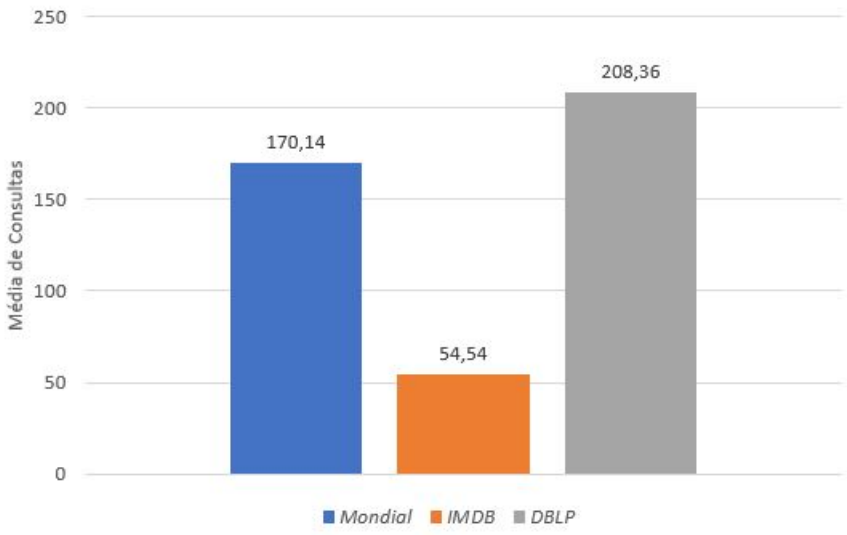

Figura 5.5: *Média de Mapeamentos Keymantic / banco de dados.*

Na fração do resultado da consulta 14 do banco de dados *Mondial*, disposta o Código 5.1 é possível verificar tal situação. São somente 03 consultas, mas o que difere de uma para outra é onde cada subpalavra *countries* é associada ora na coluna latitude, ora na coluna longitude, ora na coluna nome.

Código 5.1 Fração Resultado Processamento Consulta 14 banco de dados *Mondial* e Técnica *Keymanic*

```
1  SELECT borders.length, country.Name, country.Code, country.Capital,
2  country.Area, country.Population, city.name, city.Population,
3  city.Latitude, city.Longitude, sea.Name, sea.Depth FROM borders,
4  country, city, located, sea WHERE country.Code = borders.Country1
5  AND country.Code = city.Country AND city.Name = located.City  AND
6  sea.Name = located.Sea AND  lower(city.name) like lower('\%without\%')
7  AND  lower(city.Latitude) like lower('\%countries\%')
8
9  SELECT borders.length, country.Name, country.Code, country.Capital,
10 country.Area, country.Population, city.name, city.Population,
11 city.Latitude, city.Longitude, sea.Name, sea.Depth FROM borders,
12 country, city, located, sea WHERE country.Code = borders.Country1
13 AND country.Code = city.Country  AND city.Name = located.City
14 AND sea.Name = located.Sea  AND  lower(city.Longitude) like
15 lower('\%countries\%') AND  lower(city.name) like lower('\%without\%')
16
17 SELECT borders.length, country.Name, country.Code, country.Capital,
18 country.Area, country.Population, city.name, city.Population,
19 city.Latitude, city.Longitude, sea.Name, sea.Depth FROM borders,
20 country, city, located, sea WHERE country.Code = borders.Country1
21 AND country.Code = city.Country  AND city.Name = located.City
22 AND sea.Name = located.Sea  AND  lower(city.name) like lower('\%countries\%')
23 AND  lower(city.Latitude) like lower('\%without\%')
```

A Figura 5.6 apresenta a média de precisão em relação a cada banco de dados. Os resultados gerados para cada consulta do banco de dados pode ser observado individualmente no Apêndice B e Tabelas B.4, B.5, e B.6. São apresentados resultados para a média simples da precisão, média simples da revocação e a *f-measure* para cada banco de dados em relação aos resultados recuperados e os resultados esperados.

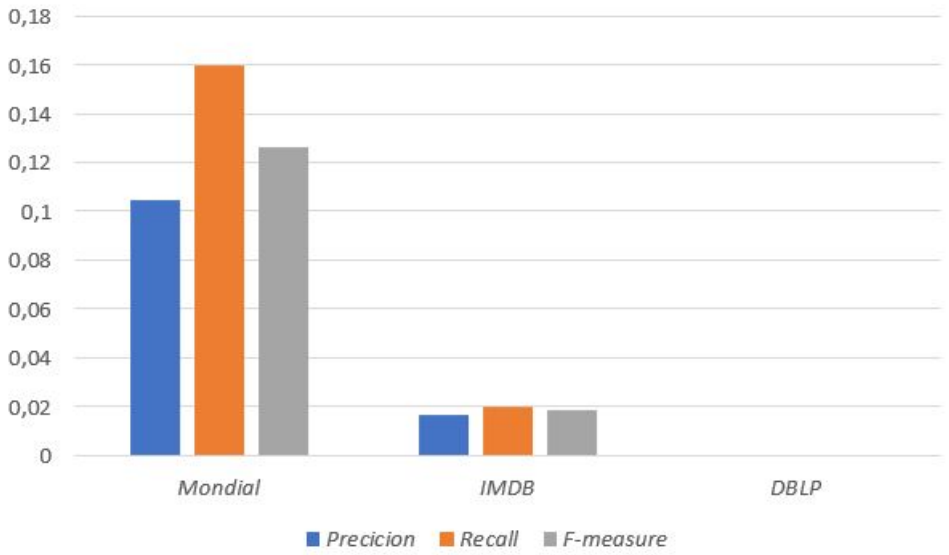

Figura 5.6: *Resumo Métricas Keymantic / banco de dados.*

Em relação ao banco de dados *DBLP*, a técnica *Keymantic* gerou a maior quantidade de mapeamentos e consequentemente era esperado uma precisão em relação aos dados, mas em função do fracionamento da consulta e a comparação de cada fração com um campo das tabelas diferentes e a utilização do operador "*and*" os resultados não eram encontrados. Para os bancos de dados *IMDB* e *Mondial* houve resultados recuperados que pertenciam ao conjunto de resultados esperados, mas a maior parte dos mapeamentos gerados não era capaz de gerar resultados válidos. Em relação a métrica revocação que é calculada baseada no número de resultados relevantes e não relevantes a média simples da revocação é maior que a média simples da revocação pois sempre houve mais resultados retornados que os esperados.

A técnica *Keymantic* gera vários mapeamentos na tentativa de apresentar o maior número possível de resultados e que algum deles esteja no conjunto de resultados esperados. Dessa forma, não podemos associar o número de mapeamento a uma possível precisão em relação aos resultados esperados.

Ramada et al. é uma das técnicas que constrói um grafo de esquema, mas diferentemente de todas as técnicas investigadas neste trabalho, realiza pré-processamento com as consultas submetidas. A expansão da consulta é realizada sempre, e somente após a expansão tentar identificar se há banco de dados que possui o contexto da consulta. Nos experimentos realizados, havia informação somente do banco que se desejava acessar, assim, os campos da tabela TME foram preenchidos com os valores de metadados do banco de dados que se desejava utilizar.

A Figura 5.7 apresenta o comportamento das consultas dispostas no Apêndice B em relação a executadas, falha e tempo para os três bancos de dados. Em algumas consultas, a técnica precisou ser interrompida em razão do tempo de execução, entretanto,

observou-se um processamento parcial e mapeamentos criados. Na maioria dos casos havia o direcionamento correto da consulta, ou seja, se a semântica da consulta fosse uma função de agregação *max* por exemplo, a consulta construída pela técnica apresentava a função de agregação desejada.

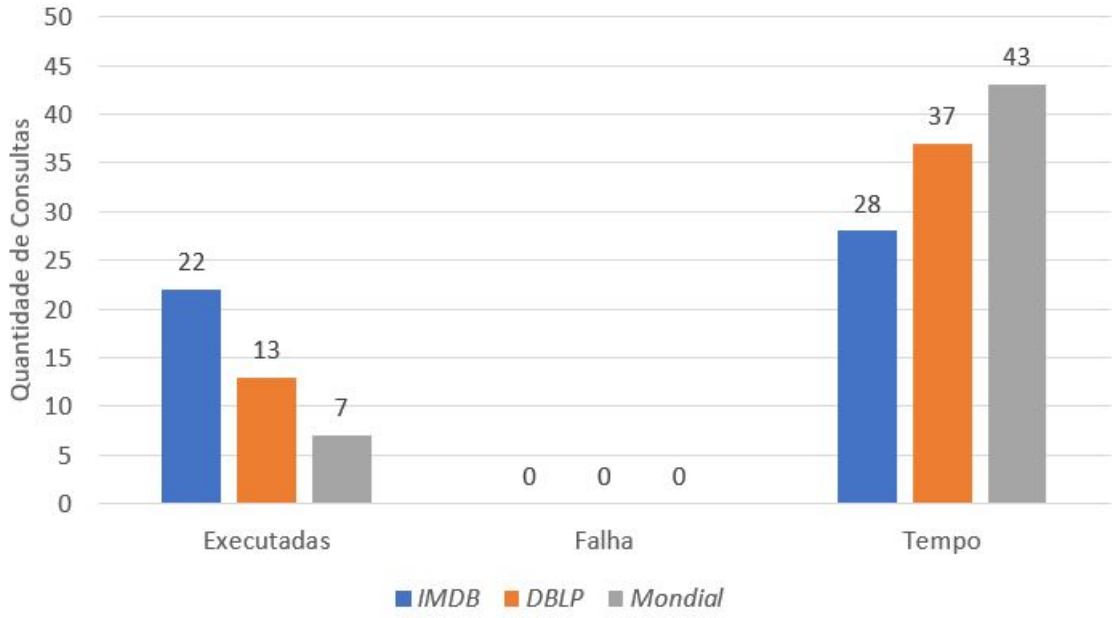

Figura 5.7: *Comportamento das Consultas Ramada et al./ banco de dados.*

A Figura 5.8 apresenta as métricas em relação a precisão da técnica para os bancos de dados e as informações detalhadas podem ser observadas no Apêndice B e Tabelas B.7, B.8 e B.9. É possível observar que a técnica não apresentou nenhum resultado para o banco de dados *Mondial*, diferentemente das anteriores. Enquanto isso, apresentou resultados que eram esperados para o banco *DBLP*, o que outras técnicas não fizeram. O que novamente nos demonstra a diferença das técnicas e as particularidades individuais. Enquanto Ramada et. al consegue apresentar resultados para o banco de dados com maior número de registros mas não consegue apresentar para um banco de dados com um número de registros bem menor, porém, de maior complexidade de relacionamentos.

Analisando, por exemplo, a consulta 46 disposta no Apêndice A e Tabela A.2 "*Largest Continents*", que possui como semântica desejada o "Continente com a maior área territorial" e que precisaria ser verificado a utilização de uma função de agregação, a técnica apresentou dois resultados, que não eram o desejado. Contudo, ao se verificar o comando *SQL* que foi utilizado para gerar os resultados em ambos havia a utilização da função de agregação desejada, mas o mesmo foi feito em campo incorreto. As consultas geradas foram "***select distinct max(continent.name) from continent***" e "***select distinct max(encompasses.continent) from encompasses***".

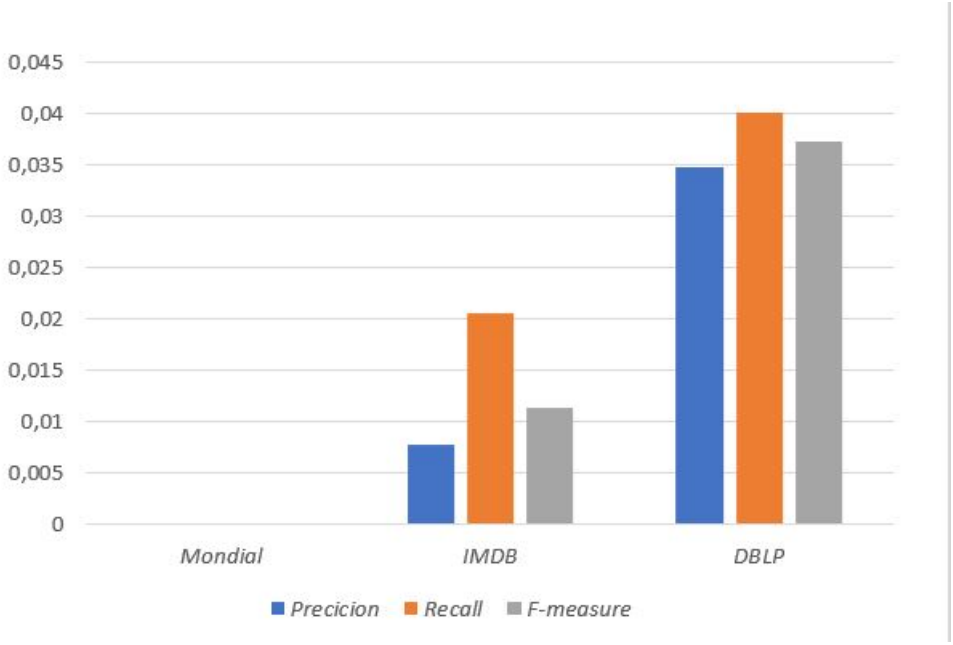

Figura 5.8: *Resumo Métricas Ramada et al./ banco de dados.*

Após apresentados todos os resultados, percebe-se a necessidade do direcionamento de cada uma das técnicas. Cada um delas é eficiente em um determinado ponto. A Figura 5.9 apresenta a média simples da precisão, média simples da revocação e *f-measure* apresentadas anteriormente por cada técnicas. Observando a Figura é possível perceber que em relação à precisão, o banco de dados que apresenta melhor desempenho é o *Mondial*, independente da técnica definida, seguida da banco *DBLP* e *IMDB*. O que se ajusta em relação à ordem crescente do tamanho, em quantidade de registros, dos bancos de dados.

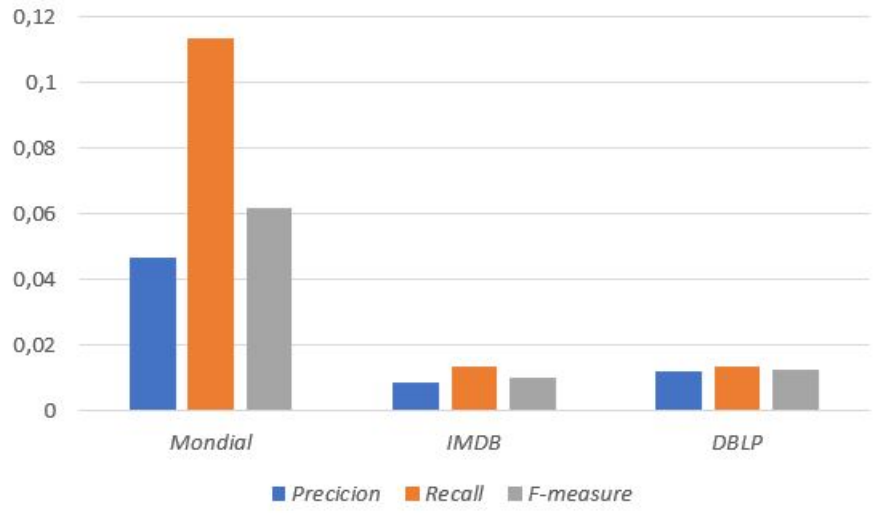

Figura 5.9: *Média das Métricas por Banco de Dados.*

O banco de dados *Mondial* e *IMDB*, possui resultados em duas técnicas enquanto o banco *DBLP* possui resultados apenas em uma técnica. O banco de dados *IMDB* apesar de ser grande em número de registros o mesmo é simples em relação a complexidade de seus relacionamentos. *Banks-II* obteve melhor resultado, em relação à submissão das consultas, no banco de dados com um menor número de registros, já *Keymantic* obteve bons resultados, em relação à submissão de consulta, nos dois menores bancos de dados, enquanto, Ramada et al.apresentou melhores resultados, em relação à submissão de consulta, para os bancos de dados com menor complexidade de relacionamentos.

5.3.3 Perspectiva das Características Semânticas

O objetivo principal deste trabalho é a verificação do pré-processamento em relação às características semântica descritas na seção 4.1.2. Considerando que quando escrevemos uma consulta para realizar uma busca de informações não escrevemos somente palavras-chave e que algumas vezes não desejamos saber o que está exatamente escrito, mas uma interpretação que o conjunto de palavras utilizadas é possível realizar. Pensamos que ao lidar com um software ele precisa se comportar como nós seres humanos o que, vem a ser uma falácia.

Os desenvolvedores de sistemas ao construírem os algoritmos agregam além de conhecimentos técnicos o conhecimento cotidiano de sua vivência. Mas, mesmo construindo o *software* mais robusto, que parece ser possível até o presente momento não há *software* que realize todas as ações que são desejadas. Temos *softwares* especialistas que são bons em determinadas tarefas, mas não em todas.

Ao propor a verificação da análise semântica nas técnicas de pesquisa por palavras-chave o bancos de dados relacionais, pretendia-se verificar quão próximo ao dialeto e incertezas do usuário as técnicas estariam. Os usuários de sistemas, por vezes, não sabem nem como pesquisar um filme por exemplo em buscadores da *internet* ou quais seriam as melhores palavras que poderiam ser utilizadas.

Dessa incerteza surgiu o conjunto de consultas para cada banco de dados, feitas por pessoas reais com palavras que não auxiliam ou com significado diferente do que realmente estava escrito, conforme descrito na seção 5.1. Cada uma das características utilizadas neste trabalho, em algum momento, foi propósito de discussão em trabalhos anteriores como em [8, 35, 7, 2], assim são motivos de observação a algum tempo.

Conforme detalhado anteriormente na seção 4.1.2, as consultas selecionadas foram distribuídas em relação às características conforme Apêndice A e Tabelas A.1, A.2 e A.3. Após a execução dos experimentos, foi verificado se cada técnica realizou o processamento desejado e informado individualmente na Tabela B. A Tabela 5.4 apresenta de forma sintetizada quais característica cada técnica realizou o processamento

e as características semânticas serão abreviadas para **S** - *Stopwords*, **EX** - Expansão da Consulta, **AG** - Funções de Agregação e por fim **FR** - Segmentação e Frase.

Tabela 5.4: *Análise dos resultados retornados em relação as características semânticas.*

	IMDB				*DBLP*				*Mondial*			
	S	EX	AG	FR	S	EX	AG	FR	S	EX	AG	FR
Banks-II												
Keymantic				✓				✓				✓
Ramada et al.		✓	✓			✓	✓			✓	✓	

É possível perceber que se uma característica é realizada em um banco de dados por uma técnica, é realizada em todas. Isso implica que se a técnica não conseguiu apresentar resultados que estavam no conjunto de resultados esperados não quer dizer que não houve o processamento das características. Este pré-processamento é observado antes da submissão da consulta e quando realizados aumentam a quantidade de palavras submetidas.

A técnica *Banks-II* não realiza nenhum processamento na consulta, entende a consulta como uma expressão e submete-a como uma expressão e também não fraciona a consulta. Desta forma, a única característica que poderia ser observada seria a de Segmentação e Frase, mas somente se tratasse da consulta completa e não da fragmentação e assim não considerado que técnica fizesse tal processamento. Assim, nenhuma das características semânticas desejadas foi atribuída a técnica *Banks-II*. Neste caso o não pré-processamento não interferiu na apresentação de resultados, sejam esses resultados relevantes ou não.

A técnica *Keymantic* fraciona a consulta em subpalavras, exceto se elas estiverem entre aspas duplas, desta forma, o processamento de segmentação e frase foi considerado para a técnicas. As demais características não foram observadas; a técnica não considera a retirada, substituição ou acréscimo de nenhuma palavra. Assim, somente a característica Segmentação e Frase foi atribuída a técnica.

Antes de iniciar o processamento da qualquer consulta, a técnica Ramada et al. expande as palavras-chave em relação a sinônimos. Após a expansão, são mapeadas algumas palavras que podem indicar que a intenção do usuário seja a utilização da função de agregação, e se isso for encontrado, a palavra-chave é substituída para uma função de agregação e as demais palavras são utilizadas na construção da consulta. Todavia, a técnica não é capaz de identificar uma expressão na consulta submetida ou que seja uma expressão. Assim, a característica expansão da consulta e a agregação foi atribuída a técnica.

Nenhuma das técnicas investigadas realizou a retirada de *stopwords* ou eliminação de qualquer palavra mesmo que elas não acrescentem ganho semântico na consulta.

5.4 Considerações Finais

Neste Capítulo foram apresentados inicialmente o contexto dos experimentos, e como primeira situação descrevemos o tamanho, em número de registros, dos bancos de dados *IMDB* e *DBLP*. Mesmo reduzindo os bancos de dados, ele ainda estava significativamente grande em relação *Mondial* e houve técnicas que não conseguiram apresentar os dados. Mas, ao observar as técnicas em relação aos bancos, percebe-se que complexidade de relacionamentos interfere na execução das consultas, e neste caso sem considerar a precisão. O tamanho em número de registros é obviamente um problema para a técnica que gera um grafo de dados, que foi discutido e questionado no Capítulo 2.

Quanto a precisão da técnica ao submeter uma consulta e a mesma apresentar resultados válidos, se observarmos as figuras dispostas ao longo da seção 5.3 podemos perceber que ainda há um longo caminho a ser precorrido para aprimorar a precisão, e que tanto o tamanho da base quanto a complexidade dos relacionamentos afeta a execução. Com a fragmentação das ações da técnica de consulta, iniciando na identificação da palavra e a adesão de critérios semânticos pela técnica pode auxiliar na precisão dos resultados. Porque, conforme foi demonstrado, um número grande de candidatos a resposta não tem relação com a precisão, não é o número de resultados que precisam ser acrescidos, mas o processamento para encontrar estes resultados.

Quanto a análise do processamento ou não das características semânticas, observa-se que nenhuma das características é processada por mais de uma técnica, houve inclusive característica que não foi processada por nenhuma técnica. Se levarmos em consideração o ano de publicação de cada técnica (*Banks-II* [3] - 2002; *Keymantic* [5] - 2010; e Ramada et al. [35] - 2020) podemos observar a evolução.

Os resultados dos experimentos de cada uma das técnicas podem ser observados nos arquivos externos BANKS-II[8], Keymantic[9] e Ramada et al[10].

[8] https://bitlybr.com/r2CM
[9] https://bitlybr.com/gUeR
[10] https://bitlybr.com/l3dJPU

CAPÍTULO 6

Conclusão

Em razão das particularidades de cada uma das técnicas - *Banks-II*, *Keymantic* e Ramada, (1) o grafo gerado (dados ou esquema), (2) como os resultados são apresentados não é possível, nem é intenção deste trabalho, realizar uma comparação entre essas técnicas. As técnicas foram o meio para compreender a existência ou não de um processamento semântico das consultas submetidas.

Ao iniciar este trabalho haviam dois cenários: o primeiro descrito em [14, 16], que apresentava uma avaliação de eficiência e depois eficácia em técnicas de pesquisa por palavras-chave conforme abordado na seção 3.1 e que ao final apresenta os resultados. Entretanto, o estudo discutiu sobre o tamanho dos bancos de dados, que interferiram no resultado. O segundo cenário descrito em [32] e abordado na seção 3.2 que introduz o conceito de separação das palavras por termos de metadados ou valor, em parte por utilizar somente as técnicas de pesquisa que construíam o grafo de esquema. Mais uma vez, houve a conjectura sobre o tamanho do banco de dados, mas desta vez em relação a sua complexidade de relacionamentos entre as tabelas.

Este estudo pretendia verificar qual, e se havia, pré-processamentos para as consultas submetidas, pois seja em língua portuguesa ou inglesa, que é o idioma dos bancos de dados, sempre há neologismos e acepções diferentes. Estes possuem relativa significância na interpretação de uma frase. A escrita, por exemplo, de um *tuíte* por parte de um usuário é repleto de ambiguidades e intenções diferentes. Trabalhar com essa incerteza trouxe para este estudo a evidência de que estamos nos caminho correto.

É preciso verificar quanto ao tamanho, em número de registros do banco de dados e quanto a complexidade das tabelas destes banco, mas principalmente que estamos vivendo uma época em que há dados abundantes por toda parte e precisamos processá-los. Se observarmos os dados apresentados, podemos evidenciar que mesmo não realizando o pré-processamento de uma determinada característica semântica as técnicas não conseguiram ser precisas.

Foi possível perceber que não são todas as técnicas utilizadas neste estudo que realizam algum tipo de processamento semântico. Em especial o tratamento de *stopwords*, que nenhuma das técnicas realiza e em estudo [2] sobre técnicas de consulta

que utilizam linguagem natural, e analisando algumas características semânticas, dentre essas características o o processamento de *stopwords*. O estudo verificou 24 técnicas e somente 03 delas realizam o processamento de *stopwords*, e as técnicas que realizam esse processamento são datadas de 2015 e 2017.

Analisando os resultados obtidos dos experimentos realizados e a proposta principal deste trabalho, é possível perceber que a maioria das técnicas não realiza todos os processamentos semânticos requeridos. Todos os documentos gerados por este trabalho são lista de consultas, semântica, resultados esperados, os bancos de dados com estrutura e inserção de dados, assim como os resultados gerados pela avaliação, se encontram disponíveis para consulta e apreciação.

6.1 Contribuições

As contribuições deste trabalho são apresentados a seguir:

- **Definição de Bancos de Dados** - foram definidos 03 bases de dados atuais (*IMDB*, *DBLP* e *Mondial*) com cargas reais para serem utilizadas em avaliações futuras. Para as bases de dados *IMDB* e *DBLP* foi definido uma base e uma fração da mesma.
- **Definição e disponibilização de *benchmark*** - Foi definido um *benchmark* com a formalização da metodologia de avaliação semântica de técnicas de consultas por palavras-chave com definições de métricas, consultas, semântica, resultados esperados.

6.2 Limitações

Considerando os principais objetivos, o presente trabalho possui algumas limitações que devem ser consideradas, sendo elas:

- Não foram utilizadas técnicas avançadas de estatística na avaliação dos resultados gerados. Optou-se por utilizar somente as métricas de precisão, revocação e *f-measure* em razão de não realizar ranqueamentos.
- A não utilização de um ranqueamento de resultados, mas um conjunto em razão da diversidade de apresentação dos resultados.
- O tamanho dos bancos de dados, pois como foi observado ao longo do trabalho os melhores resultados foram apresentados nos menores bancos.

6.3 Trabalhos Futuros

Durante o desenvolvimento deste projeto foram identificadas os seguinte trabalhos futuros:

- O *benchmark* proposto por este estudo se ateve a avaliar o processamento semântico e a eficácia dos resultados retornados pelas técnicas de consultas por palavras-chave a bancos de dados relacionais. No entanto, não foi adicionado nenhuma forma de avaliar a eficiência. Como trabalho futuro, devem ser consideradas métricas que afiram o tempo gasto por cada técnica para concluir cada consulta para os bancos de dados propostas por esta avaliação.
- A avaliação deste trabalho foi realizada de forma manual, ou seja, foram executadas as consultas em cada técnica para cada banco de dados, após a obtenção dos resultados da consulta foi verificado se os resultados retornados eram relevantes. Como proposta de trabalho futuro, pretende-se estudar uma forma de automatizar o processo buscando agilizar o processo de avaliação para outras técnicas.
- Este estudo utilizou de técnicas de pesquisa por palavras-chave a bancos de dados relacionais implementados pelos próprios autores o que gerou resultados diferentes para cada técnica. Como trabalho futuro sugere-se a padronização e a reimplementação das técnicas para que os resultados sejam apresentados de forma idêntica.

Referências Bibliográficas

[1] ADITYA, B. **BANKS : Browsing and Keyword Searching in Relational Databases *.** *Vldb*, p. 1083 – 1086, 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, 2019.

[3] AGRAWAL, S.; CHAUDHURI, S.; DAS, G. **DBXplorer: A system for keyword-based search over relational databases**. *Proceedings - International Conference on Data Engineering*, p. 5–16, 2002.

[4] BERGAMASCHI, S.; DOMNORI, E.; GUERRA, F.; TRILLO LADO, R.; VELEGRAKIS, Y. **Keyword search over relational databases: a metadata approach**. *Sigmod*, p. 565, 2011.

[5] BERGAMASCHI, S.; EMILIA, R.; LADO, R. T. **Keymantic : Semantic Keyword-based Searching in Data Integration Systems**. *Vldb*, p. 1637–1640, 2010.

[6] BERGAMASCHI, S.; FERRO, N.; GUERRA, F.; SILVELLO, G. **Keyword-based search over databases: A roadmap for a reference architecture paired with an evaluation framework**. *Transactions on Computational Collective Intelligence XXI*, p. 1–20, 2016.

[7] BERGAMASCHI, S.; GUERRA, F. **Quest: A keyword search system for relational data based on semantic and machine learning techniques**. *Proceedings of the ...*, 6(12):0–3, 2013.

[8] BERGAMASCHI, S.; GUERRA, F.; INTERLANDI, M.; TRILLO-LADO, R.; VELEGRAKIS, Y. **Combining user and database perspective for solving keyword queries over relational databases**. *Information Systems*, 55:1–19, 2016.

[9] BERRY, M.; CYBENKO, G.; LARSON, J. **Scientific benchmark characterizations**. *Parallel Computing*, 17(10-11):1173–1194, 1991.

[10] BHALOTIA, G.; HULGERI, A.; NAKHE, C.; CHAKRABARTI, S.; SUDARSHAN, S. **Keyword searching and browsing in databases using banks**. In: *Proceedings 18th International Conference on Data Engineering*, p. 431–440, Feb 2002.

[11] BIENIA, C.; LI, K. **Benchmarking modern multiprocessors**. Princeton University Princeton, NJ, 2011.

[12] CARRARO, A.; AGOSTI, M.; SILVELLO, G. **Keyword Search in Relational Databases: Architecture, Approaches and Considerations**. PhD thesis, Ateneo di Padova, February 2017.

[13] CLEVERDON, C. **The Cranfield Tests On Index Language Devices**. *Aslib Proceedings*, 19(6):173–194, 1967.

[14] COFFMAN, J.; WEAVER, A. C. **A framework for evaluating database keyword search strategies**. *Proceedings of the 19th ACM international conference on Information and knowledge management - CIKM '10*, p. 729–738, 2010.

[15] COFFMAN, J.; WEAVER, A. C. **Structured data retrieval using cover density ranking**. In: *Proceedings of the 2Nd International Workshop on Keyword Search on Structured Data*, KEYS '10, p. 1:1–1:6, New York, NY, USA, 2010. ACM.

[16] COFFMAN, J.; WEAVER, A. C. **An empirical performance evaluation of relational keyword search techniques**. *IEEE Transactions on Knowledge and Data Engineering*, 26(1):30–42, 2014.

[17] DEWITT, D. J. **The wisconsin benchmark: Past, present, and future.**, 1993.

[18] DING, B.; YU, J. X.; WANG, S.; QIN, L.; ZHANG, X.; LIN, X. **Finding top-k min-cost connected trees in databases**. *Proceedings - International Conference on Data Engineering*, p. 836–845, 2007.

[19] ELMASRI, R.; NAVATHE, S. B. **Fundamentals of database systems:[6-th edition]**, 2011.

[20] FAKHRAEE, S.; FOTOUHI, F. **DBSemSXplorer: Semantic-based Keyword Search System over Relational Databases for Knowledge Discovery**. *Proceedings of the Third International Workshop on Keyword Search on Structured Data - KEYS '12*, p. 54–62, 2012.

[21] GHANBARPOUR, A.; NADERI, H.; PERMISSIONS, F.; GHANBARPOUR, A.; NADERI, H. **A Model-based Keyword Search Approach for Detecting Top-k Effective Answers**. *The Computer Journal*, 2018.

[22] GRAY, J. **Database and transaction processing performance handbook.**, 1993.

[23] HE, H.; WANG, H.; YANG, J.; YU, P. S. **BLINKS : Ranked Keyword Searches on Graphs**. *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, p. 305–316, 2007.

[24] HRISTIDIS, V.; L.GRAVANO.; PAPAKONSTANTINOU, Y. **Efficient IR- Style Keyword Search over Relational Databases**. *Data Base*, p. 850–861, 2003.

[25] HRISTIDIS, V.; PAPAKONSTANTINOU, Y. **DISCOVER: keyword search in relational databases**. *Vldb*, p. 670–681, 2002.

[26] ISHAK, M. B.; LERAY, P.; AMOR, N. B. **Probabilistic relational model benchmark generation: Principle and application**. *Intelligent Data Analysis*, 20(3):615–635, 2016.

[27] KACHOLIA, V.; PANDIT, S.; CHAKRABARTI, S.; SUDARSHAN, S.; DESAI, R.; KARAMBELKAR, H. **Bidirectional expansion for keyword search on graph databases**. *VLDB '05 Proceedings of the 31st international conference on Very large data bases*, p. 505–516, 2005.

[28] KARGAR, M.; AN, A.; CERCONE, N.; GODFREY, P.; SZLICHTA, J.; YU, X. **Meanks: Meaningful keyword search in relational databases with large and complex schema**. *Proceedings - International Conference on Data Engineering*, 2015-May:411–422, 2015.

[29] LI, G.; JI, S.; LI, C.; FENG, J. **Efficient Type-Ahead Search on Rel ational ata: a TASTIER Approach**. *Sigmod*, p. 695–706, 2009.

[30] LUO, Y.; WANG, W.; LIN, X. **SPARK: A keyword search engine on relational databases**. *Proceedings - International Conference on Data Engineering*, 00:1552–1555, 2008.

[31] MARINS, W. F. **Ontologias de domínio na interpretação de consultas a bancos de dados relacionais**. *Proceedings of IADIS International Conference WWW/Internet*, p. 79–86, 2015.

[32] OLIVEIRA FILHO, A. D. C. **Benchmark para Métodos de Consultas por Palavras-Chave a Bancos de Dados Relacionais**. Master's thesis, Universidade Federal de Goiás, 2018.

[33] PU, K. Q.; YU, X. **FRISK: Keyword query cleaning and processing in action**. *Proceedings - International Conference on Data Engineering*, p. 1531–1534, 2009.

[34] QUEIROZ, L. T. **Um Benchmark para Avaliação de Técnicas de Busca no Contexto de Análise de Mutantes SQL**. Master's thesis, Universidade Federal de Goiás, 2013.

[35] 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, 2020.

[36] SAKAI, T. **Metrics, statistics, tests**. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8173 LNCS:116–163, 2014.

[37] SENG, J. L. **A study on industry and synthetic standard benchmarks in relational and object databases**. *Industrial Management and Data Systems*, 103(7):516–532, 2003.

[38] SILVA, J.; KOWATA, E.; VINCENZI, A. **Extracting and exposing relational database metadata on the web**. In: *Proceedings of IADIS International Conference WWW/Internet*, p. 35–42, 2012.

[39] STAPENHURST, T. **The Benchmarking Book**, volume 91. Routledge, 2009.

[40] TURBYFILL, C.; ORJI, C. U.; BITTON, D. **As3ap: An ansi sql standard scaleable and portable benchmark for relational database systems.**, 1993.

[41] WANG, H.; AGGARWAL, C. C. **A SURVEY OF ALGORITHMS FOR KEYWORD SEARCH ON GRAPH DATA**. *Database*, 40:161–180, 2010.

[42] WAZLAWICK, R. **Metodologia de Pesquisa para Ciência da Computação**. Elsevier Editora Ltda., 2017.

APÊNDICE A

Consultas Propostas

Tabela A.1: *Conjunto de palavras-chaves para avaliação semântica quanto as características Stopwords (S), Expansão da Consulta (EX), Funções de Agregação (AG) e Segmentação e Frase (FR) para a Base de Dados IMDB.*

	Consultas	***S***	**EX**	**AG**	**FR**	***Semântica***
1	*marc forster director*		✓			Infomações sobre o *director Marc Forest*
2	*director movie list*		✓			Lista de Diretores
3	*finished animation series*		✓			Séries do gênero animação que terminaram
4	*movies action*		✓			Lista com todos os filmes de ação
5	*movies princess disney*		✓			Lista de nome de todos os filmes de princesa da *Disney*
6	*movies years 2020*		✓			Filmes cujo ano de lançamento seja 2020
7	*supernatural episodes list*		✓			Lista de episódios da série *Supernatural*
8	*"game of thrones"actors*		✓		✓	Atores da Série *Game of Thrones*
9	*christmas tv shows*		✓			*Shows* de TV de Natal
10	*"Steven Spielberg"directoring*		✓		✓	Todos os filmes dirigidos por *Steven Spielberg*
11	*movies "James Harkness"acting*		✓		✓	Filmes que o ator *James Harkness* atuou
12	*actor dead "Fast and Furious"*		✓		✓	Atores do filme Velozes e Furiosos que estão mortos
13	*movies with most actress*		✓	✓		Filmes com muitas atrizes
14	*best movies ratings 2020*		✓	✓		Filmes melhor classificados lançados em 2020
15	*actor died in 2020*	✓	✓			Atores que morreram em 2020
16	*movie short "based true story"*	✓	✓			Filmes curtos com título *based true story*
17	*movies episode in the sea*	✓	✓			Episódios de série com título *sea*
18	*avengers movies all actress name*	✓	✓			Todas as atrizes do filme *Avengers*
19	*episode of "Yes, Prime Minister"*	✓	✓		✓	Episódios da série *Yes, Prime Minister*

Tabela A.1: *Conjunto de palavras-chaves para avaliação semântica quanto as características Stopwords (S), Expansão da Consulta (EX), Funções de Agregação (AG) e Segmentação e Frase (FR) para a Base de Dados IMDB.*

	Consultas	***S***	**EX**	**AG**	**FR**	***Semântica***
20	*movies of drama and romance*	✓				Filmes que possuem como gênero drama e romance
21	*"good versus evil"*	✓				Filmes de nome *Good versus evil*
22	*comedy or documentary*	✓				Informações do gênero *comedy* e *documentary*
23	*movies about police*	✓				O nome de todos filmes sobre policiais
24	*number of movies genres*	✓		✓		O número de filmes por gêneros
25	*how many smallville episodes are there*	✓		✓		A quantidade de episódios de *Smallville*
26	*movies with more than one director*	✓		✓		Filmes que possuem mais de um diretor
27	*movies with "angelina jolie"and "brad pitt"*	✓			✓	Filmes que possuem como atores *Angelina Jolie* e *Brad Pitt*
28	*movie comedian of the 1920s and 1930s*	✓			✓	Filmes de comédia lançados entre 1920 e 1930
29	*directed a "star is born"*	✓			✓	Diretores do filme *Star is born*
30	*i rate "The Lion King"*	✓			✓	Classificação do filme *The Lion King*
31	*First episode of Warrior 2019*	✓			✓	Primeiro episódio de *Warrior* em 2019
32	*"Wayne Roberts"and "Roger Moore"*	✓			✓	Registros da pessoa *Wayne Roberts* e *Roger Moore* - pode ser verificado atores ou diretores
33	*cast of "Jack and Jill"*	✓			✓	Todos os participantes do filme *Jack and Jill*
34	*movie director Scorcese and actor "Di Caprio"*	✓			✓	Filme do diretor *Scorcede* e Ator *Di Caprio*
35	*movies bestrating*			✓		Filmes melhores classificados

Tabela A.1: *Conjunto de palavras-chaves para avaliação semântica quanto as características Stopwords (S), Expansão da Consulta (EX), Funções de Agregação (AG) e Segmentação e Frase (FR) para a Base de Dados IMDB.*

	Consultas	***S***	**EX**	**AG**	**FR**	***Semântica***
36	*year first "star wars"movie*			✓	✓	Ano do primeiro filme da *Star Wars*
37	*rating movies for genres*			✓		Classificação dos filmes por gênero
38	*number of films released per year*			✓		O número de filmes lançados por ano
39	*movies with two genres*			✓		Filmes que possuem dois gêneros
40	*vikings series start date*			✓		Ano de inicio da série *Viking*
41	*number movies types*			✓		Número de filmes por tipo
42	*how many "star wars"movies are there*			✓	✓	A quantidade de filmes da saga *Star Wars*
43	*movie musical the last 5 years*				✓	Os filmes do gênero musical dos últimos 5 anos
44	*horror movies after 2018*				✓	Os filmes cujo gênero seja "horror" e lançados a partir de 2018
45	*movies "husband wife relation"*				✓	Informações sobre o filme *Husband and Wife Relation*
46	*actors "The Last Word"*				✓	Todos os atores do filme *The last Word*
47	*acting and directing at the same time*				✓	Atores que atuaram e dirigiram o mesmo filme
48	*movies not released actresses "rachel mcadams"*				✓	Filmes não lançados da atriz *Rachel Mcadams*
49	*movies director "mel gibson"*				✓	Filmes do diretor *Mel Gilbson*
50	*movies without director*				✓	Filmes sem diretor

Tabela A.2: *Conjunto de palavras-chaves para avaliação semântica quanto as características Stopwords (S), Expansão da Consulta (EX), Funções de Agregação (AG) e Segmentação e Frase (FR) para a Base de Dados Mondial.*

	Consultas	***S***	**EX**	**AG**	**FR**	***Semântica***
1	*Countries border italy*		✓			Os países que fazem fronteira com a Itália
2	*Artificial lake Oceania*		✓			Lago Artificial na Oceania
3	*Lake Denmark*		✓			lago na Dinamarca
4	*California*		✓			Informações sobre o estado da Califórnia
5	*Madagaskar*		✓			Informações sobre a Ilha de Madagascar
6	*Countries*		✓			Lista de todos os países
7	*Asia*		✓			Informações sobre o Continente Asiático
8	*Christian*		✓			Informações sobre a religião Cristã
9	*Tocantins*		✓			Informações sobre o Rio Tocantins, Ou sobre o estado de Tocantins
10	*Brasil*		✓			Informações sobre o país Brasil
11	*Countries member wtro*		✓		✓	Lista de países membros da Organização Mundial do Comércio
12	*Countries language English*		✓		✓	países que falam a língua inglesa
13	*Oceans name*		✓		✓	Lista com nome de todos os oceanos
14	*Countries without sea border*		✓		✓	Países que não possuem o mar em uma de suas fronteiras
15	*"Arabian Desert"*		✓		✓	Informações sobre o Deserto da Arabia
16	*"New York"*		✓		✓	Informações sobre *New York*

Tabela A.2: *Conjunto de palavras-chaves para avaliação semântica quanto as características Stopwords (S), Expansão da Consulta (EX), Funções de Agregação (AG) e Segmentação e Frase (FR) para a Base de Dados Mondial.*

	Consultas	***S***	**EX**	**AG**	**FR**	***Semântica***
17	*Religion in Japan*	✓	✓			Informações sobre a religião no Japão
18	*River through Paris*	✓	✓			Lista de rios que cortam a cidade de *Paris*
19	*Canada name the provinces and territories*	✓	✓			Todos os estados do *Canadá*
20	*Countries and language spoken*	✓	✓			Países e os idiomas falados
21	*Country largest population in the world*	✓	✓	✓		O país mais populoso do mundo
22	*About ethnicity in philippines*	✓				Informações sobre etnia das *Phipippines*
23	*Cities in Georgia*	✓				Todas as cidades do estado da Geórgia
24	*Countries in Europe*	✓				Países que fazem parte do continente Europeu
25	*River in Cuzco*	✓				Rios que cortam *Cuzco*
26	*USA States name them*	✓				Lista com os estados dos Estados Unidos da América
27	*USA member of international organizations*	✓				As organizações internacionais que os Estados Unidos são membros
28	*Tributaries of Nile river*	✓				Afluentes do Rio Nilo
29	*Countries less than 1 million population*			✓		Lista de todos os países que possuem menos de 1 milhão de pessoas
30	*countries more smaller than area of Hawaii island*	✓		✓		Lista de países, cuja área é menor que a área total da *Ilha do Havaí*
31	*Most spoken language the world*	✓		✓		Língua mais falada no mundo
32	*Biggest river in the world*	✓		✓		Maior rio em extensão no mundo

Tabela A.2: *Conjunto de palavras-chaves para avaliação semântica quanto as características Stopwords (S), Expansão da Consulta (EX), Funções de Agregação (AG) e Segmentação e Frase (FR) para a Base de Dados Mondial.*

	Consultas	***S***	**EX**	**AG**	**FR**	***Semântica***
33	*Largest countries in territory*	✓		✓		O maior país em extensão territorial
34	*Number ethnic groups in Niger*	✓		✓		Os grupos étnicos encontrados na *Nigéria*
35	*Country with most borders*	✓		✓		País com a maior fronteira.
36	*Caribbean name the island countries*	✓			✓	Lista de países e ilhas que estão situados no Mar do *Caribe*
37	*Largest city in the world by population*	✓			✓	Cidade com a maior população no mundo
38	*Names of countries without rivers*	✓			✓	Países que não possuem rios catalogados em seu território
39	*International organizations based in london*	✓			✓	Organizações internacionais com sede em Londres
40	*Largest continent population in the world*	✓		✓	✓	O continente mais populoso do mundo
41	*Biggest island in the "pacific ocean"*	✓		✓	✓	Maior ilha do Oceano Pacífico
42	*Mountain more height*			✓		Montanha mais alta do mundo
43	*Name largest island*			✓		Maior ilha em território
44	*Population mundial*			✓		A quantidade de habitantes do mundo
45	*Ocean the largest number of countries*			✓	✓	Oceano que possuem a maior quantidade de países banhados
46	*Largest Continents*			✓	✓	Continente com a maior área territorial
47	*Country with most borders with other countries*			✓	✓	País que possuem o maior número de fronteiras com outros países

Tabela A.2: *Conjunto de palavras-chaves para avaliação semântica quanto as características Stopwords (S), Expansão da Consulta (EX), Funções de Agregação (AG) e Segmentação e Frase (FR) para a Base de Dados Mondial.*

	Consultas	***S***	**EX**	**AG**	**FR**	***Semântica***
48	*All language*				✓	Todos os idiomas falados no mundo
49	*Countries monarchy government*				✓	Países que possuem como forma de governo a Monarquia
50	*"Pico de Agulhas Negras"*				✓	Informações sobre a Montanha Pico de Agulhas Negras

Tabela A.3: *Conjunto de palavras-chaves para avaliação semântica quanto as características Stopwords (S), Expansão da Consulta (EX), Funções de Agregação (AG) e Segmentação e Frase (FR) para a Base de Dados DBLP.*

	Consultas	***S***	**EX**	**AG**	**FR**	***Semântica***
1	*Article trafficker*		✓			Os artigos que contenham no título a palavra *trafficker*
2	*Author profile*		✓			Lista de todos os autores
3	*Profile editors*			✓		Quantidade de editores em *inproceedings*, *articles* e *editores*
4	*proceedings database 2019*		✓			Os *proceedings* que contenham no título a palavra *database* e com ano de publicação 2019
5	*conference kdd 2019*		✓			As publicações de *proceedings* em livros que tenham como título *KDD* e que sejam publicados em 2019
6	*Conference proceedings publishers*		✓			As conferências de *proceedings* publicadas
7	*Workshorp ACM*		✓			Todos os *proceedings* que tenha como *publiser* a expressão *Workshop ACM*
8	*Springer proceedings in indexing*		✓			Todos os *proceedings* publicados pela *Springer*
9	*Last proceedings data mining*		✓			O mais recente *proceedings* de "*Data Mining*"
10	*"European Symposium"*		✓		✓	As publicações de *proceedings* em livros que tenham como título *European Symposium*
11	*Journal article without volume number*		✓		✓	Os *articles* que possuem jornais relacionados sem número de volume.
12	*"Computer science"book publishers*		✓		✓	Todos os *booktitles* e *publisers* da linha *computer science*
13	*Conference series proceedings "Information Security"*		✓		✓	Todos os *proceedings* da linha *Information Security*
14	*Article about bullying*	✓	✓			Os artigos que contenham no título a palavra *bullying*
15	*Conference on "Artificial Intelligence"*	✓	✓		✓	As conferencias de *Artificial Intelligence*

Tabela A.3: *Conjunto de palavras-chaves para avaliação semântica quanto as características Stopwords (S), Expansão da Consulta (EX), Funções de Agregação (AG) e Segmentação e Frase (FR) para a Base de Dados DBLP.*

	Consultas	***S***	**EX**	**AG**	**FR**	***Semântica***
16	*Conferences in Índia 2018*	✓	✓			Os *proceedings* que contenham como título a palavra *Índia* e publicados em 2018
17	*Proceedings in mathematical*	✓				Os *proceedings* que contenham no título a palavra *mathematical*
18	*The 2019 proceedings indexed ieee*	✓				Todos os *proceddings* de 2019 indexados pelo *IEEE*
19	*Title in the proceeding*	✓				Os títulos dos *proceedings*.
20	*Springer proceedings in indexing*	✓				Todos os *proceedings* indexados na *Springer*
21	*Conference proceedings with isbn*	✓				Todos os *proceedings* com número de ISBN.
22	*Article about technology*	✓	✓			Todos os artigos que possuem no título a palavra *technology*
23	*Publisher in article*	✓				Os *articles* associados a *publisher*
24	*Articles published on nanotechnology*	✓				Todos os *articles* que publicados que possuem no título a palavra *Nanotechnology*
25	*Booktitle publisered in 2020*	✓				Todos os *booktitles* e *books* publicados com versão de 2020
26	*Research on "open data"*	✓			✓	Os *Articles* ou *proceedings* que possuem no título ao expressão *"open data"*
27	*Journal of "information systems"2019*	✓			✓	Os *preoceedings* que contenha no título a expressão "*Information Systems*"e que publicados em 2019
28	*Proceedings of japan with isbn*	✓			✓	Todos *proceedings* realizados no *Japan* e com número de ISBN
29	*2019 ieee conference on "big data"*	✓			✓	Todos os *proceedings* de conferencia da *IEEE* e que possuem no título a expressão *big data*

Tabela A.3: *Conjunto de palavras-chaves para avaliação semântica quanto as características Stopwords (S), Expansão da Consulta (EX), Funções de Agregação (AG) e Segmentação e Frase (FR) para a Base de Dados DBLP.*

	Consultas	***S***	**EX**	**AG**	**FR**	***Semântica***
30	*Publisher proceedings about "computer vision"*	✓			✓	Todos os *proceedings* associados a tabela *publisher* e que possuem no título a expressão *Computer Vision*
31	*News article about computer journal education*	✓			✓	Todos os *articles* de computação publicados recentemente no jornal *education* do ano corrente - 2020
32	*Article in a journal reports*	✓			✓	todos os *articles* que possuem a palavra *reports* na especificação de *journal*
33	*Articles with more than one author*	✓		✓		Os *articles* que possuem mais de um autor
34	*Number of articles published per year*	✓		✓		O número de artigos publicados por ano
35	*how many paper have been published by ACM in 2019*	✓		✓	✓	Quantas publicações aconteceram em 2019 pelo *IEEE*
36	*Number of authors per article*			✓		A quantidade de autores por artigo.
37	*Number of articles published per journal*			✓		A quantidade de artigos por *Journal*
38	*Proceedings recent years indexed*			✓		O ano mais recente de arquivos de *proceedings* indexados
39	*Articles same author*			✓		O número de artigos de cada autor
40	*How many book editions are there*			✓	✓	A quantidade de edições possuem os *booktitles*
41	*Publishers 2017 sigmod*	✓				Os *proceedings* publicados que tenham no título *sigmod* e publicados em 2017
42	*Article not indexed*				✓	Os *articles* que não estão em um *booktitle* ou possuem *publisher* relacionado
43	*Book of proceedings 2019*				✓	Todos os livros de *proceedings* de 2019
44	*Book publisher "springer"*				✓	Todos os livros publicados pela *Springer*

Tabela A.3: *Conjunto de palavras-chaves para avaliação semântica quanto as características Stopwords (S), Expansão da Consulta (EX), Funções de Agregação (AG) e Segmentação e Frase (FR) para a Base de Dados DBLP.*

	Consultas	***S***	**EX**	**AG**	**FR**	***Semântica***
45	*Proceedings series "web"*				✓	Todos os *proceedings* da série que contenha a palavra *web*
46	*Inproceedings booktitle*				✓	Todos os *inproceedings* que possuem *booktitle* associados
47	*Article types "systematic review"*				✓	Todos os artigos que contenham no título a especificação de *systematic review*
48	*"Proceedings of the European Conference on Information Systems"*				✓	Todos os *proceedigngs* que possuem como título *Proceedings of the European Conference on Information Systems*
49	*Inproceeding "Semantic Web"*				✓	Todos os registros de *inproceedings* sobre *Semantic Web*
50	*Reviewed articles published*				✓	Todos os *articles* que possui revisão / editor publicado

APÊNDICE B

Experimentos Realizados

Tabela B.1: *Resultado da Execução da Técnica Banks-II e base de dados DBLP. A e C - Relevantes; B e D - Não Relevantes.*

	S	***EX***	**AG**	**FR**	**Realizou o processamento desejado?**	***Esperados***		Recuperados		Não Recuperados		*precision*	*recall*	*f-measure*
						Universo	Encontrados	A	B	C	D			
1		✓			Não	1318151	13881	0	0	13881	1304270	0,000000	0,000000	0,000000
2		✓			Não	8601	8601	0	0	8601	0	0,000000	0,000000	0,000000
3		✓			Não	8601	2	0	0	2	8599	0,000000	0,000000	0,000000
4		✓			Não	46622	41	0	0	41	46581	0,000000	0,000000	0,000000
5		✓			Não	46622	24	0	0	24	46598	0,000000	0,000000	0,000000
6		✓			Não	46622	44529	0	0	44529	2093	0,000000	0,000000	0,000000
7		✓			Não	46622	3200	0	0	3200	43422	0,000000	0,000000	0,000000
8		✓			Não	46622	15171	0	0	15171	31451	0,000000	0,000000	0,000000
9		✓			Não	46622	1	0	0	1	46621	0,000000	0,000000	0,000000
10		✓		✓	Não	46622	10	0	0	10	46612	0,000000	0,000000	0,000000
11		✓		✓	Não	1318651	244851	0	0	244851	1073800	0,000000	0,000000	0,000000
12		✓		✓	Não	838	67	0	0	67	771	0,000000	0,000000	0,000000
13		✓		✓	Não	46622	12	0	0	12	46610	0,000000	0,000000	0,000000
14	✓	✓			Não	1318151	31	0	0	31	1318120	0,000000	0,000000	0,000000
15	✓	✓			Não	46622	933	0	0	933	45689	0,000000	0,000000	0,000000
16	✓	✓			Não	46622	95	0	0	95	46527	0,000000	0,000000	0,000000
17	✓				Não	46622	3	0	0	3	46619	0,000000	0,000000	0,000000
18	✓				Não	46622	851	0	0	851	45771	0,000000	0,000000	0,000000
19	✓				Não	46622	46622	0	0	46622	0	0,000000	0,000000	0,000000

Tabela B.1: *Resultado da Execução da Técnica Banks-II e base de dados DBLP. A e C - Relevantes; B e D - Não Relevantes.*

	S	***EX***	**AG**	**FR**	**Realizou o processamento desejado?**	***Esperados***		Recuperados		Não Recuperados		*precision*	*recall*	*f-measure*
						Universo	Encontrados	A	B	C	D			
20	✓				Não	46622	15207	0	0	15207	31415	0,000000	0,000000	0,000000
21	✓				Não	46622	38561	0	0	38561	8061	0,000000	0,000000	0,000000
22	✓				Não	1318151	9786	0	0	9786	1308365	0,000000	0,000000	0,000000
23	✓				Não	1318151	1	0	0	1	1318150	0,000000	0,000000	0,000000
24	✓				Não	1318151	156	0	0	156	1317995	0,000000	0,000000	0,000000
25	✓				Não	838	2	0	0	2	836	0,000000	0,000000	0,000000
26	✓			✓	Não	1364773	1600	0	0	1600	1363173	0,000000	0,000000	0,000000
27	✓			✓	Não	46622	118	0	0	118	46504	0,000000	0,000000	0,000000
28	✓			✓	Não	46622	1210	0	0	1210	45412	0,000000	0,000000	0,000000
29	✓			✓	Não	46622	14	0	0	14	46608	0,000000	0,000000	0,000000
30	✓			✓	Não	46622	472	0	0	472	46150	0,000000	0,000000	0,000000
31	✓			✓	Não	1318151	2	0	0	2	1318149	0,000000	0,000000	0,000000
32	✓			✓	Não	1318151	702	0	0	702	1317449	0,000000	0,000000	0,000000
33	✓		✓		Não	1318151	830380	0	0	830380	487771	0,000000	0,000000	0,000000
34	✓		✓		Não	1318151	11	0	0	11	1318140	0,000000	0,000000	0,000000
35	✓		✓	✓	Não	46622	1	0	0	1	46621	0,000000	0,000000	0,000000
36			✓		Não	1318151	1153929	0	0	1153929	164222	0,000000	0,000000	0,000000
37			✓		Não	1318151	1629	0	0	1629	1316522	0,000000	0,000000	0,000000
38			✓		Não	46622	1	0	0	1	46621	0,000000	0,000000	0,000000

Tabela B.1: *Resultado da Execução da Técnica Banks-II e base de dados DBLP. A e C - Relevantes; B e D - Não Relevantes.*

	S	***EX***	**AG**	**FR**	**Realizou o processamento desejado?**	***Esperados***		Recuperados		Não Recuperados		*precision*	*recall*	*f-measure*
						Universo	Encontrados	A	B	C	D			
39			✓		Não	1318151	601812	0	0	601812	716339	0,000000	0,000000	0,000000
40			✓	✓	Não	8403	766	0	0	766	7637	0,000000	0,000000	0,000000
41				✓	Não	46622	20	0	0	20	46602	0,000000	0,000000	0,000000
42				✓	Não	1318651	1318650	0	0	1318650	1	0,000000	0,000000	0,000000
43			000000	✓	Não	46622	4	0	0	4	46618	0,000000	0,000000	0,000000
44			000000	✓	Não	8403	185	0	0	185	8218	0,000000	0,000000	0,000000
45				✓	Não	46622	4	0	0	4	46618	0,000000	0,000000	0,000000
46				✓	Não	1439011	2	0	0	2	1439009	0,000000	0,000000	0,000000
47				✓	Não	1318151	3	0	0	3	1318148	0,000000	0,000000	0,000000
48				✓	Não	46622	6	0	0	6	46616	0,000000	0,000000	0,000000
49				✓	Não	1439511	1826	0	0	1826	1437685	0,000000	0,000000	0,000000
50				✓	Não	1318151	14	0	0	14	1318137	0,000000	0,000000	0,000000
	22	16	8	23								0,000000	0,000000	0,000000

Tabela B.2: *Resultado da Execução da Técnica Banks-II e base de dados IMDB. A e C - Relevantes; B e D - Não Relevantes.*

	S	***EX***	**AG**	**FR**	**Realizou o processamento desejado?**	***Esperados***		Recuperados		Não Recuperados		*precision*	*recall*	*f-measure*
						Universo	Encontrados	A	B	C	D			
1		✓			Não	316156	12	0	0	12	316144	0,000000	0,000000	0,000000
2		✓			Não	316156	316156	0	0	316156	0	0,000000	0,000000	0,000000
3		✓			Não	3673131	2240	0	0	2240	3670891	0,000000	0,000000	0,000000
4		✓			Não	473410	27155	0	0	27155	446255	0,000000	0,000000	0,000000
5		✓			Não	473410	256	0	0	256	473154	0,000000	0,000000	0,000000
6		✓			Não	473410	7842	0	0	7842	465568	0,000000	0,000000	0,000000
7		✓			Não	473410	195	0	0	195	473215	0,000000	0,000000	0,000000
8		✓			Não	473410	9	0	0	9	473401	0,000000	0,000000	0,000000
9		✓			Não	473410	81	0	0	81	473329	0,000000	0,000000	0,000000
10		✓		✓	Não	473410	31	0	0	31	473379	0,000000	0,000000	0,000000
11		✓		✓	Não	473410	2	0	0	2	473408	0,000000	0,000000	0,000000
12		✓		✓	Não	473410	1	0	0	1	473409	0,000000	0,000000	0,000000
13		✓	✓		Não	473410	98851	0	0	98851	374559	0,000000	0,000000	0,000000
14		✓	✓		Não	473410	3	0	0	3	473407	0,000000	0,000000	0,000000
15	✓	✓			Não	225493	61	0	0	61	225432	0,000000	0,000000	0,000000
16	✓	✓			Não	473410	12	0	0	12	473398	0,000000	0,000000	0,000000
17	✓	✓			Não	473410	36	0	0	36	473374	0,000000	0,000000	0,000000
18	✓	✓		✓	Não	473410	26	0	0	26	473384	0,000000	0,000000	0,000000
19	✓	✓		✓	Não	2416136	12	0	0	12	2416124	0,000000	0,000000	0,000000

Tabela B.2: *Resultado da Execução da Técnica Banks-II e base de dados IMDB. A e C - Relevantes; B e D - Não Relevantes.*

					Realizou o processamento desejado?	***Esperados***		Recuperados		Não Recuperados		*precision*	*recall*	*f-measure*
	S	***EX***	**AG**	**FR**		Universo	Encontrados	A	B	C	D			
20	✓				Não	473410	12895	0	0	12895	460515	0,000000	0,000000	0,000000
21	✓				Não	473410	583	0	0	583	472827	0,000000	0,000000	0,000000
22	✓				Não	28	2	0	0	2	26	0,000000	0,000000	0,000000
23	✓				Não	473410	181	0	0	181	473229	0,000000	0,000000	0,000000
24	✓		✓		Não	473410	27	0	0	27	473383	0,000000	0,000000	0,000000
25	✓		✓		Não	473410	1	0	0	1	473409	0,000000	0,000000	0,000000
26	✓		✓		Não	473410	25195	0	0	25195	448215	0,000000	0,000000	0,000000
27	✓			✓	Não	473410	2	0	0	2	473408	0,000000	0,000000	0,000000
28	✓			✓	Não	473410	14667	0	0	14667	458743	0,000000	0,000000	0,000000
29	✓			✓	Não	316156	2	0	0	2	316154	0,000000	0,000000	0,000000
30	✓			✓	Não	473410	2	0	0	2	473408	0,000000	0,000000	0,000000
31	✓			✓	Não	2416136	2	0	0	2	2416134	0,000000	0,000000	0,000000
32	✓			✓	Não	392541	5	0	0	5	392536	0,000000	0,000000	0,000000
33	✓			✓	Não	2136884	12	0	0	12	2136872	0,000000	0,000000	0,000000
34	✓			✓	Não	473410	4	0	0	4	473406	0,000000	0,000000	0,000000
35			✓		Não	473410	10	0	0	10	473400	0,000000	0,000000	0,000000
36			✓		Não	473410	1	0	0	1	473409	0,000000	0,000000	0,000000
37			✓		Não	473410	27	0	0	27	473383	0,000000	0,000000	0,000000
38			✓		Não	473410	49	0	0	49	473361	0,000000	0,000000	0,000000

Tabela B.2: *Resultado da Execução da Técnica Banks-II e base de dados IMDB. A e C - Relevantes; B e D - Não Relevantes.*

	S	***EX***	**AG**	**FR**	**Realizou o processamento desejado?**	***Esperados***		Recuperados		Não Recuperados		*precision*	*recall*	*f-measure*
						Universo	Encontrados	A	B	C	D			
39			✓		Não	473410	83517	0	0	83517	389893	0,000000	0,000000	0,000000
40			✓		Não	473410	1	0	0	1	473409	0,000000	0,000000	0,000000
41			✓		Não	473410	2	0	0	2	473408	0,000000	0,000000	0,000000
42			✓	✓	Não	338511	1	0	0	1	338510	0,000000	0,000000	0,000000
43				✓	Não	338511	690	0	0	690	337821	0,000000	0,000000	0,000000
44				✓	Não	338511	3410	0	0	3410	335101	0,000000	0,000000	0,000000
45				✓	Não	473410	8	0	0	8	473402	0,000000	0,000000	0,000000
46				✓	Não	225493	11	0	0	11	225482	0,000000	0,000000	0,000000
47				✓	Não	392557	2	0	0	2	392555	0,000000	0,000000	0,000000
48				✓	Não	473410	2	0	0	2	473408	0,000000	0,000000	0,000000
49				✓	Não	473410	4	0	0	4	473406	0,000000	0,000000	0,000000
50				✓	Não	473410		0	0	0	473410	0,000000	0,000000	0,000000
	20	19	13	22								0,000000	0,000000	0,000000

Tabela B.3: *Resultado da Execução da Técnica Banks-II e base de dados Mondial. A e C - Relevantes; B e D - Não Relevantes.*

	S	***EX***	**AG**	**FR**	**Realizou o processamento desejado?**	***Esperados***		Recuperados		Não Recuperados		*precision*	*recall*	*f-measure*
						Universo	Encontrados	A	B	C	D			
1		✓			Não	238	6	0	0	6	232	0,000000	0,000000	0,000000
2		✓			Não	130	1	0	0	1	129	0,000000	0,000000	0,000000
3		✓			Não	130	1	0	0	1	129	0,000000	0,000000	0,000000
4		✓			Não	1450	1	1	86	0	1363	0,011494	1,000000	0,022727
5		✓			Não	276	1	1	13	0	262	0,071429	1,000000	0,133333
6		✓			Não	238	238	0	97	238	-97	0,000000	0,000000	0,000000
7		✓			Não	5	1	1	56	0	-52	0,017544	1,000000	0,034483
8		✓			Não	454	81	80	1	1	372	0,987654	0,987654	0,987654
9		✓			Não	1688	2	2	9	0	1677	0,181818	1,000000	0,307692
10		✓			Não	238	1	1	84	0	153	0,011765	1,000000	0,023256
11		✓		✓	Não	238	111	0	0	111	127	0,000000	0,000000	0,000000
12		✓		✓	Não	238	21	0	0	21	217	0,000000	0,000000	0,000000
13		✓		✓	Não	34	34	0	0	34	0	0,000000	0,000000	0,000000
14		✓		✓	Não	238	45	0	0	45	193	0,000000	0,000000	0,000000
15		✓		✓	Não	63	1	1	7	0	55	0,125000	1,000000	0,222222
16		✓		✓	Não	3111	1	1	9	0	3101	0,100000	1,000000	0,181818
17	✓	✓			Não	454	1	0	0	1	453	0,000000	0,000000	0,000000
18	✓	✓			Não	218	9	0	0	9	209	0,000000	0,000000	0,000000
19	✓	✓			Não	1450	12	0	0	12	1438	0,000000	0,000000	0,000000

Tabela B.3: *Resultado da Execução da Técnica Banks-II e base de dados Mondial. A e C - Relevantes; B e D - Não Relevantes.*

	S	***EX***	**AG**	**FR**	**Realizou o processamento desejado?**	***Esperados***		Recuperados		Não Recuperados		*precision*	*recall*	*f-measure*
						Universo	Encontrados	A	B	C	D			
20	✓	✓			Não	382	302	0	0	302	80	0,000000	0,000000	0,000000
21	✓	✓	✓		Não	238	1	0	0	1	237	0,000000	0,000000	0,000000
22	✓				Não	604	4	0	0	4	600	0,000000	0,000000	0,000000
23	✓				Não	3111	5	0	0	5	3106	0,000000	0,000000	0,000000
24	✓				Não	238	53	0	0	53	185	0,000000	0,000000	0,000000
25	✓				Não	61889	1	0	0	1	61888	0,000000	0,000000	0,000000
26	✓				Não	1450	51	0	0	51	1399	0,000000	0,000000	0,000000
27	✓				Não	8.008	72	0	0	72	7936	0,000000	0,000000	0,000000
28	✓				Não	218	3	0	0	3	215	0,000000	0,000000	0,000000
29	✓		✓	✓	Não	238	83	0	0	83	155	0,000000	0,000000	0,000000
30	✓		✓		Não	238	158	0	0	158	80	0,000000	0,000000	0,000000
31	✓		✓		Não	144	1	0	0	1	143	0,000000	0,000000	0,000000
32	✓		✓		Não	218	1	0	0	1	217	0,000000	0,000000	0,000000
33	✓		✓		Não	238	1	0	0	1	237	0,000000	0,000000	0,000000
34	✓		✓		Não	540	1	0	0	1	539	0,000000	0,000000	0,000000
35	✓		✓		Não	238	1	0	0	1	237	0,000000	0,000000	0,000000
36	✓			✓	Não	238	30	0	0	30	208	0,000000	0,000000	0,000000
37	✓			✓	Não	3111	1	0	0	1	3110	0,000000	0,000000	0,000000
38	✓			✓	Não	238	131	0	0	131	107	0,000000	0,000000	0,000000

Tabela B.3: *Resultado da Execução da Técnica Banks-II e base de dados Mondial. A e C - Relevantes; B e D - Não Relevantes.*

						Esperados		Recuperados		Não Recuperados		*precision*	*recall*	*f-measure*
	S	***EX***	**AG**	**FR**	**Realizou o processamento desejado?**	Universo	Encontrados	A	B	C	D			
39	✓			✓	Não	153	4	0	0	4	149	0,000000	0,000000	0,000000
40	✓		✓	✓	Não	5	1	0	0	1	4	0,000000	0,000000	0,000000
41	✓		✓	✓	Não	276	1	0	0	1	275	0,000000	0,000000	0,000000
42			✓		Não	240	1	0	0	1	239	0,000000	0,000000	0,000000
43			000000 x		Não	276	1	0	0	1	275	0,000000	0,000000	0,000000
44			000000 x		não	238	1	0	0	1	237	0,000000	0,000000	0,000000
45			✓	✓	Não	272	1	0	0	1	271	0,000000	0,000000	0,000000
46			✓	✓	Não	5	1	0	0	1	4	0,000000	0,000000	0,000000
47			✓	✓	Não	238	1	0	0	1	237	0,000000	0,000000	0,000000
48				✓	não	144	74	0	0	74	70	0,000000	0,000000	0,000000
49				✓	Não	238	28	0	0	28	210	0,000000	0,000000	0,000000
50				✓	Não	240	1	1	3	0	236	0,250000	1,000000	0,400000
	25	21	16	19								0,035134	0,000000	0,000000

Tabela B.4: *Resultado da Execução da Técnica Keymantic e base de dados DBLP. A e C - Relevantes; B e D - Não Relevantes.*

	S	***EX***	**AG**	**FR**	**Realizou o processamento desejado?**	***Esperados***		Map	Recuperados		Não Recuperados		*precision*	*recall*	*f-measure*
						Universo	Encontrados		A	B	C	D			
1		✓			Não	14	1318151	13881	0	0	13881	1304270	0,000000	0,000000	0,000000
2		✓			Não	6	8601	8601	0	0	8601	0	0,000000	0,000000	0,000000
3		✓			Não	E	333 8601	2	0	0	2	8599	0,000000	0,000000	0,000000
4		✓			Não	111	46622	41	0	0	41	46581	0,000000	0,000000	0,000000
5		✓			Não	100	46622	24	0	0	24	46598	0,000000	0,000000	0,000000
6		✓			Não	16	46622	44529	0	0	44529	2093	0,000000	0,000000	0,000000
7		✓			Não	99	46622	3200	0	0	3200	43422	0,000000	0,000000	0,000000
8		✓			Não	304	46622	15171	0	0	15171	31451	0,000000	0,000000	0,000000
9		✓			Não	303	46622	1	0	0	1	46621	0,000000	0,000000	0,000000
10		✓		✓	Não	E	46622	10	0	0	10	46612	0,000000	0,000000	0,000000
11		✓		✓	Não	168	1318651	244851	0	0	244851	1073800	0,000000	0,000000	0,000000
12		✓		✓	Não	10	838	67	0	0	67	771	0,000000	0,000000	0,000000
13		✓		✓	Não	303	46622	12	0	0	12	46610	0,000000	0,000000	0,000000
14	✓	✓			Não	128	1318151	31	0	0	31	1318120	0,000000	0,000000	0,000000
15	✓	✓			Não	100	46622	933	0	0	933	45689	0,000000	0,000000	0,000000
16	✓	✓			Não	100	46622	95	0	0	95	46527	0,000000	0,000000	0,000000
17	✓				Não	191	46622	3	0	0	3	46619	0,000000	0,000000	0,000000
18	✓				Não	609	46622	851	0	0	851	45771	0,000000	0,000000	0,000000
19	✓				Não	566	46622	46622	0	0	46622	0	0,000000	0,000000	0,000000

Tabela B.4: *Resultado da Execução da Técnica Keymantic e base de dados DBLP. A e C - Relevantes; B e D - Não Relevantes.*

	S	***EX***	**AG**	**FR**	**Realizou o processamento desejado?**	***Esperados***		Map	Recuperados		Não Recuperados		*precision*	*recall*	*f-measure*
						Universo	Encontrados		A	B	C	D			
20	✓				Não	303	46622	15207	0	0	15207	31415	0,000000	0,000000	0,000000
21	✓				Não	303	46622	38561	0	0	38561	8061	0,000000	0,000000	0,000000
22	✓				Não	128	1318151	9786	0	0	9786	1308365	0,000000	0,000000	0,000000
23	✓				Não	11	1318151	1	0	0	1	1318150	0,000000	0,000000	0,000000
24	✓				Não	128	1318151	156	0	0	156	1317995	0,000000	0,000000	0,000000
25	✓				Não	10	838	2	0	0	2	836	0,000000	0,000000	0,000000
26	✓			✓	PH	100	1364773	1600	0	0	1600	1363173	0,000000	0,000000	0,000000
27	✓			✓	PH	202	46622	118	0	0	118	46504	0,000000	0,000000	0,000000
28	✓			✓	Não	303	46622	1210	0	0	1210	45412	0,000000	0,000000	0,000000
29	✓			✓	Ph	100	46622	14	0	0	14	46608	0,000000	0,000000	0,000000
30	✓			✓	PH	606	46622	472	0	0	472	46150	0,000000	0,000000	0,000000
31	✓			✓	Não	404	1318151	2	0	0	2	1318149	0,000000	0,000000	0,000000
32	✓			✓	Não	404	1318151	702	0	0	702	1317449	0,000000	0,000000	0,000000
33	✓		✓		Não	404	1318151	830380	0	0	830380	487771	0,000000	0,000000	0,000000
34	✓		✓		Não	1362	1318151	11	0	0	11	1318140	0,000000	0,000000	0,000000
35	✓		✓	✓	Não	E	46622	1	0	0	1	46621	0,000000	0,000000	0,000000
36			✓		Não	425	1318151	1153929	0	0	1153929	164222	0,000000	0,000000	0,000000
37			✓		Não	694	1318151	1629	0	0	1629	1316522	0,000000	0,000000	0,000000
38			✓		Não	275	46622	1	0	0	1	46621	0,000000	0,000000	0,000000

Tabela B.4: *Resultado da Execução da Técnica Keymantic e base de dados DBLP. A e C - Relevantes; B e D - Não Relevantes.*

	S	***EX***	**AG**	**FR**	**Realizou o processamento desejado?**	***Esperados***		Map	Recuperados		Não Recuperados		*precision*	*recall*	*f-measure*
						Universo	Encontrados		A	B	C	D			
39			✓		Não	8	1318151	601812	0	0	601812	716339	0,000000	0,000000	0,000000
40			✓	✓	Não	101	8403	766	0	0	766	7637	0,000000	0,000000	0,000000
41				✓	Não	106	46622	20	0	0	20	46602	0,000000	0,000000	0,000000
42				✓	Não	128	1318651	1318650	0	0	1318650	1	0,000000	0,000000	0,000000
43	F00			✓	Não	58	46622	4	0	88	4	46530	0,000000	0,000000	0,000000
44				✓	Não	6	8403	185	0	0	185	8218	0,000000	0,000000	0,000000
45				✓	Não	9	46622	4	0	0	4	46618	0,000000	0,000000	0,000000
46				✓	Não	6	1439011	2	0	0	2	1439009	0,000000	0,000000	0,000000
47				✓	PH	202	1318151	3	0	0	3	1318148	0,000000	0,000000	0,000000
48				✓	PH	404	46622	6	0	0	6	46616	0,000000	0,000000	0,000000
49				✓	PH	90	1439511	1826	0	0	1826	1437685	0,000000	0,000000	0,000000
50				✓	Não	10	1318151	14	0	0	14	1318137	0,000000	0,000000	0,000000
	22	16	8	23									0,000000	0,000000	0,000000

Tabela B.5: *Resultado da Execução da Técnica Keymantic e base de dados IMDB. A e C - Relevantes; B e D - Não Relevantes.*

	S	***EX***	**AG**	**FR**	**Realizou o processamento desejado?**	***Esperados***		Map	Recuperados		Não Recuperados		*precision*	*recall*	*f-measure*
						Universo	Encontrados		A	B	C	D			
1		✓			Não	31	316156	12	0	0	12	316144	0,000000	0,000000	0,000000
2		✓			Não	6	316156	316156	0	0	316156	0	0,000000	0,000000	0,000000
3		✓			Não	30	3673131	2240	0	0	2240	3670891	0,000000	0,000000	0,000000
4		✓			Não	3	473410	27155	27155	5289	0	440966	0,836981	1,000000	0,911257
5		✓			Não	6	473410	256	0	4	256	473150	0,000000	0,000000	0,000000
6		✓			Não	6	473410	7842	0	0	7842	465568	0,000000	0,000000	0,000000
7		✓			Não	6	473410	195	0	0	195	473215	0,000000	0,000000	0,000000
8		✓			Não	2	473410	9	0	0	9	473401	0,000000	0,000000	0,000000
9		✓			Não	30	473410	81	0	0	81	473329	0,000000	0,000000	0,000000
10		✓		✓	PH	54	473410	31	0	0	31	473379	0,000000	0,000000	0,000000
11		✓		✓	PH	3	473410	2	0	0	2	473408	0,000000	0,000000	0,000000
12		✓		✓	Não	100	473410	1	0	0	1	473409	0,000000	0,000000	0,000000
13		✓	✓		Não	101	473410	98851	0	0	98851	374559	0,000000	0,000000	0,000000
14		✓	✓		Não	4	473410	3	0	0	3	473407	0,000000	0,000000	0,000000
15	✓	✓			Não	67	225493	61	0	0	61	225432	0,000000	0,000000	0,000000
16	✓	✓			Não	101	473410	12	0	0	12	473398	0,000000	0,000000	0,000000
17	✓	✓			Não	100	473410	36	0	0	36	473374	0,000000	0,000000	0,000000
18	✓	✓		✓	Não	125	473410	26	0	0	26	473384	0,000000	0,000000	0,000000
19	✓	✓		✓	PH	6	2416136	12	0	0	12	2416124	0,000000	0,000000	0,000000

Tabela B.5: *Resultado da Execução da Técnica Keymantic e base de dados IMDB. A e C - Relevantes; B e D - Não Relevantes.*

	S	***EX***	**AG**	**FR**	**Realizou o processamento desejado?**	***Esperados***		Map	Recuperados		Não Recuperados		*precision*	*recall*	*f-measure*
						Universo	Encontrados		A	B	C	D			
20	✓				Não	101	473410	12895	0	0	12895	460515	0,000000	0,000000	0,000000
21	✓				Não	30	473410	583	0	0	583	472827	0,000000	0,000000	0,000000
22	✓				Não	30	28	2	0	0	2	26	0,000000	0,000000	0,000000
23	✓				Não	6	473410	181	0	0	181	473229	0,000000	0,000000	0,000000
24	✓		✓		Não	12	473410	27	0	0	27	473383	0,000000	0,000000	0,000000
25	✓		✓		Não	101	473410	1	0	0	1	473409	0,000000	0,000000	0,000000
26	✓		✓		Não	101	473410	25195	0	0	25195	448215	0,000000	0,000000	0,000000
27	✓			✓	Não	101	473410	2	0	0	2	473408	0,000000	0,000000	0,000000
28	✓			✓	Não	101	473410	14667	0	0	14667	458743	0,000000	0,000000	0,000000
29	✓			✓	Não	101	316156	2	0	0	2	316154	0,000000	0,000000	0,000000
30	✓			✓	PH	30	473410	2	0	0	2	473408	0,000000	0,000000	0,000000
31	✓			✓	Não	101	2416136	2	0	0	2	2416134	0,000000	0,000000	0,000000
32	✓			✓	PH	30	392541	5	0	0	5	392536	0,000000	0,000000	0,000000
33	✓			✓	PH	30	2136884	12	0	0	12	2136872	0,000000	0,000000	0,000000
34	✓			✓	Não	101	473410	4	0	0	4	473406	0,000000	0,000000	0,000000
35			✓		Não	3	473410	10	0	0	10	473400	0,000000	0,000000	0,000000
36			✓		Não	101	473410	1	0	0	1	473409	0,000000	0,000000	0,000000
37			✓		Não	4	473410	27	0	0	27	473383	0,000000	0,000000	0,000000
38			✓		Não	101	473410	49	0	0	49	473361	0,000000	0,000000	0,000000

Tabela B.5: *Resultado da Execução da Técnica Keymantic e base de dados IMDB. A e C - Relevantes; B e D - Não Relevantes.*

	S	***EX***	**AG**	**FR**	**Realizou o processamento desejado?**	***Esperados***		Map	Recuperados		Não Recuperados		*precision*	*recall*	*f-measure*
						Universo	Encontrados		A	B	C	D			
39			✓		Não	13	473410	83517	0	0	83517	389893	0,000000	0,000000	0,000000
40			✓		Não	65	473410	1	0	0	1	473409	0,000000	0,000000	0,000000
41			✓		Não	3	473410	2	0	0	2	473408	0,000000	0,000000	0,000000
42			✓	✓	Não	101	338511	1	0	0	1	338510	0,000000	0,000000	0,000000
43				✓	Não	101	338511	690	0	0	690	337821	0,000000	0,000000	0,000000
44				✓	Não	11	338511	3410	0	0	3410	335101	0,000000	0,000000	0,000000
45				✓	PH	3	473410	8	0	0	8	473402	0,000000	0,000000	0,000000
46				✓	PH	54	225493	11	0	0	11	225482	0,000000	0,000000	0,000000
47				✓	Não	202	392557	2	0	0	2	392555	0,000000	0,000000	0,000000
48				✓	Não	101	473410	2	0	0	2	473408	0,000000	0,000000	0,000000
49				✓	Não	101	473410	4	0	0	4	473406	0,000000	0,000000	0,000000
50				✓	Não	6	473410		0	0	0	473410	0,000000	0,000000	0,000000
													0,016740	0,020000	0,018225

Tabela B.6: *Resultado da Execução da Técnica Keymantic e base de dados Mondial. A e C - Relevantes; B e D - Não Relevantes.*

	S	***EX***	**AG**	**FR**	**Realizou o processamento desejado?**	***Esperados***		Map	Recuperados		Não Recuperados		*precision*	*recall*	*f-measure*
						Universo	Encontrados		A	B	C	D			
1		✓			Não	238	6	13	0	0	6	232	0,000000	0,000000	0,000000
2		✓			Não	130	1	13	0	0	1	129	0,000000	0,000000	0,000000
3		✓			Não	130	1	2	0	0	1	129	0,000000	0,000000	0,000000
4		✓			Não	1450	1	67	1	2	0	1447	0,333333	1,000000	0,500000
5		✓			Não	276	1	67	1	0	0	275	1,000000	1,000000	1,000000
6		✓			Não	238	238	67	0	0	238	0	0,000000	0,000000	0,000000
7		✓			Não	5	1	67	1	14	0	-10	0,066667	1,000000	0,125000
8		✓			Não	454	81	66	81	1	0	372	0,987805	1,000000	0,993865
9		✓			Não	1688	2	66	2	0	0	1686	1,000000	1,000000	1,000000
10		✓			Não	238	1	66	1	2	0	235	0,333333	1,000000	0,500000
11		✓		✓	Não	238	111	13	0	0	111	127	0,000000	0,000000	0,000000
12		✓		✓	Não	238	21	13	0	0	21	217	0,000000	0,000000	0,000000
13		✓		✓	Não	34	34	2	0	1	34	-1	0,000000	0,000000	0,000000
14		✓		✓	Não	238	45	12	0	0	45	193	0,000000	0,000000	0,000000
15		✓		✓	PH	63	1	67	1	0	0	62	1,000000	1,000000	1,000000
16		✓		✓	PH	3111	1	67	1	1	0	3109	0,500000	1,000000	0,666667
17	✓	✓			Não	454	1	11	0	0	1	453	0,000000	0,000000	0,000000
18	✓	✓			Não	218	9	101	0	0	9	209	0,000000	0,000000	0,000000
19	✓	✓			Não	1450	12	116	0	0	12	1438	0,000000	0,000000	0,000000

Tabela B.6: *Resultado da Execução da Técnica Keymantic e base de dados Mondial. A e C - Relevantes; B e D - Não Relevantes.*

	S	***EX***	**AG**	**FR**	**Realizou o processamento desejado?**	***Esperados***		Map	Recuperados		Não Recuperados		*precision*	*recall*	*f-measure*
						Universo	Encontrados		A	B	C	D			
20	✓	✓			Não	382	302	100	0	0	302	80	0,000000	0,000000	0,000000
21	✓	✓	✓		Não	238	1	100	0	0	1	237	0,000000	0,000000	0,000000
22	✓				Não	604	4	100	0	0	4	600	0,000000	0,000000	0,000000
23	✓				Não	3111	5	100	0	0	5	3106	0,000000	0,000000	0,000000
24	✓				Não	238	53	100	0	0	53	185	0,000000	0,000000	0,000000
25	✓				Não	61889	1	100	0	0	1	61888	0,000000	0,000000	0,000000
26	✓				Não	1450	51	202	0	0	51	1399	0,000000	0,000000	0,000000
27	✓				Não	8.008	72	32	0	0	72	7936	0,000000	0,000000	0,000000
28	✓				Não	218	3	202	0	0	3	215	0,000000	0,000000	0,000000
29			✓	✓	Não	238	83	404	0	0	83	155	0,000000	0,000000	0,000000
30	✓		✓		Não	238	158	1819	0	0	158	80	0,000000	0,000000	0,000000
31	✓		✓		Não	144	1	101	0	0	1	143	0,000000	0,000000	0,000000
32	✓		✓		Não	218	1	202	0	0	1	217	0,000000	0,000000	0,000000
33	✓		✓		Não	238	1	100	0	0	1	237	0,000000	0,000000	0,000000
34	✓		✓		Não	540	1	100	0	0	1	539	0,000000	0,000000	0,000000
35	✓		✓		Não	238	1	11	0	0	1	237	0,000000	0,000000	0,000000
36	✓			✓	Não	238	30	2184	0	0	30	208	0,000000	0,000000	0,000000
37	✓			✓	Não	3111	1	404	0	0	1	3110	0,000000	0,000000	0,000000
38	✓			✓	Não	238	131	101	0	0	131	107	0,000000	0,000000	0,000000

Tabela B.6: *Resultado da Execução da Técnica Keymantic e base de dados Mondial. A e C - Relevantes; B e D - Não Relevantes.*

	S	***EX***	**AG**	**FR**	**Realizou o processamento desejado?**	***Esperados***		Map	Recuperados		Não Recuperados		*precision*	*recall*	*f-measure*
						Universo	Encontrados		A	B	C	D			
39	✓			✓	Não	153	4	101	0	0	4	149	0,000000	0,000000	0,000000
40	✓		✓	✓	Não	5	1	404	0	0	1	4	0,000000	0,000000	0,000000
41	✓		✓	✓	Não	276	1	202	0	0	1	275	0,000000	0,000000	0,000000
42			✓		Não	240	1	16	0	0	1	239	0,000000	0,000000	0,000000
43			✓		Não	276	1	94	0	0	1	275	0,000000	0,000000	0,000000
44			✓		Não	238	1	20	0	0	1	237	0,000000	0,000000	0,000000
45			✓	✓	Não	272	1	100	0	0	1	271	0,000000	0,000000	0,000000
46			✓	✓	Não	5	1	2	0	0	1	4	0,000000	0,000000	0,000000
47			✓	✓	Não	238	1	101	0	0	1	237	0,000000	0,000000	0,000000
48				✓	Não	144	74	2	0	0	74	70	0,000000	0,000000	0,000000
49				✓	Não	238	28	7	0	0	28	210	0,000000	0,000000	0,000000
50				✓	PH	240	1	100	0	0	1	239	0,000000	0,000000	0,000000
	24	21	16	19									0,104423	0,160000	0,126371

Tabela B.7: *Resultado da Execução da Técnica Ramada et al.e base de dados DBLP. A e C - Relevantes; B e D - Não Relevantes.*

	S	*EX*	AG	FR	Realizou o processamento desejado?	*Esperados*			Recuperados		Não Recuperados		*precision*	*recall*	*f-measure*
						Universo	Encontrados		A	B	C	D			
1		✓			EX	1318151	13881	F	0	0	13881	1304270	0,000000	0,000000	0,000000
2		✓			EX	8601	8601	F	0	0	8601	0	0,000000	0,000000	0,000000
3		✓			EX	8601	2	F	0	0	2	8599	0,000000	0,000000	0,000000
4		✓			EX	46622	41	E	0	0	41	46581	0,000000	0,000000	0,000000
5		✓			EX	46622	24	E	0	0	24	46598	0,000000	0,000000	0,000000
6		✓			EX	46622	44529	E	0	0	44529	2093	0,000000	0,000000	0,000000
7		✓			EX	46622	3200	F	0	0	3200	43422	0,000000	0,000000	0,000000
8		✓			EX	46622	15171	E	0	0	15171	31451	0,000000	0,000000	0,000000
9		✓	✓		EX	46622	1	E	0	0	1	46621	0,000000	0,000000	0,000000
10		✓		✓	EX	46622	10	F	0	0	10	46612	0,000000	0,000000	0,000000
11		✓		✓	EX	1318651	244851	F	755	1322	244096	1072478	0,363505	0,003084	0,006115
12		✓		✓	EX	838	67	E	0	0	67	771	0,000000	0,000000	0,000000
13		✓		✓	EX	46622	12	E	0	0	12	46610	0,000000	0,000000	0,000000
14	✓	✓			EX	1318151	31	F	0	0	31	1318120	0,000000	0,000000	0,000000
15	✓	✓			EX	46622	933	E	0	0	933	45689	0,000000	0,000000	0,000000
16	✓	✓			EX	46622	95	E	0	0	95	46527	0,000000	0,000000	0,000000
17	✓				Não	46622	3	E	0	0	3	46619	0,000000	0,000000	0,000000
18	✓				Não	46622	851	E	0	0	851	45771	0,000000	0,000000	0,000000
19	✓				Não	46622	46622	E	0	0	46622	0	0,000000	0,000000	0,000000

Tabela B.7: *Resultado da Execução da Técnica Ramada et al.e base de dados DBLP. A e C - Relevantes; B e D - Não Relevantes.*

					Realizou o processamento desejado?	***Esperados***			Recuperados		Não Recuperados		*precision*	*recall*	*f-measure*
	S	***EX***	**AG**	**FR**		Universo	Encontrados		A	B	C	D			
20	✓				Não	46622	15207	E	0	0	15207	31415	0,000000	0,000000	0,000000
21	✓				Não	46622	38561	E	0	0	38561	8061	0,000000	0,000000	0,000000
22	✓				Não	1318151	9786	F	1	0	9785	1308365	1,000000	0,000102	0,000204
23	✓				Não	1318151	1	F	1	980948	0	337202	0,000001	1,000000	0,000002
24	✓				Não	1318151	156	E	0	0	156	1317995	0,000000	0,000000	0,000000
25	✓				Não	838	2	E	0	0	2	836	0,000000	0,000000	0,000000
26	✓			✓	Não	1364773	1600	E	0	0	1600	1363173	0,000000	0,000000	0,000000
27	✓			✓	Não	46622	118	E	0	0	118	46504	0,000000	0,000000	0,000000
28	✓			✓	Não	46622	1210	E	0	0	1210	45412	0,000000	0,000000	0,000000
29	✓			✓	Não	46622	14	E	0	0	14	46608	0,000000	0,000000	0,000000
30	✓			✓	Não	46622	472	E	0	0	472	46150	0,000000	0,000000	0,000000
31	✓			✓	Não	1318151	2	E	0	0	2	1318149	0,000000	0,000000	0,000000
32	✓			✓	Não	1318151	702	E	0	0	702	1317449	0,000000	0,000000	0,000000
33	✓		✓		Não	1318151	830380	E	0	0	830380	487771	0,000000	0,000000	0,000000
34	✓		✓		Não	1318151	11	E	0	0	11	1318140	0,000000	0,000000	0,000000
35	✓		✓	✓	Não	46622	1	E	0	0	1	46621	0,000000	0,000000	0,000000
36			✓		Não	1318151	1153929	E	0	0	1153929	164222	0,000000	0,000000	0,000000
37			✓		Não	1318151	1629	E	0	0	1629	1316522	0,000000	0,000000	0,000000
38			✓		Não	46622	1	E	0	0	1	46621	0,000000	0,000000	0,000000

Tabela B.7: *Resultado da Execução da Técnica Ramada et al.e base de dados DBLP. A e C - Relevantes; B e D - Não Relevantes.*

	S	***EX***	**AG**	**FR**	**Realizou o processamento desejado?**	***Esperados***			Recuperados		Não Recuperados		*precision*	*recall*	*f-measure*
						Universo	Encontrados		A	B	C	D			
39			✓		Não	1318151	601812	F	0	2	601812	716337	0,000000	0,000000	0,000000
40			✓	✓	Não	8403	766	E	0	0	766	7637	0,000000	0,000000	0,000000
41				✓	Não	46622	20	E	0	0	20	46602	0,000000	0,000000	0,000000
42				✓	Não	1318651	1318650	F	0	0	1318650	1	0,000000	0,000000	0,000000
43				✓	Não	46622	4	E	0	0	4	46618	0,000000	0,000000	0,000000
44				✓	Não	8403	185	F	185	310	0	7908	0,373737	1,000000	0,544118
45				✓	Não	46622	4	F	0	0	4	46618	0,000000	0,000000	0,000000
46				✓	Não	1439011	2	E	0	0	2	1439009	0,000000	0,000000	0,000000
47				✓	Não	1318151	3	E	0	0	3	1318148	0,000000	0,000000	0,000000
48				✓	Não	46622	6	E	0	0	6	46616	0,000000	0,000000	0,000000
49				✓	Não	1439511	1826	E	0	0	1826	1437685	0,000000	0,000000	0,000000
50				✓	Não	1318151	14	F	0	0	14	1318137	0,000000	0,000000	0,000000
	22	16	9	23									0,034745	0,040064	0,037215

Tabela B.8: *Resultado da Execução da Técnica Ramada et al.e base de dados IMDB. A e C - Relevantes; B e D - Não Relevantes.*

	S	*EX*	AG	FR	Realizou o processamento desejado?	*Esperados*			Recuperados		Não Recuperados		*precision*	*recall*	*f-measure*
						Universo	Encontrados		A	B	C	D			
1		✓			EX	316156	12	E	0	0	12	316156	0,000000	0,000000	0,000000
2		✓			EX	316156	316156	F	8	233	316148	315915	0,033195	0,000025	0,000051
3		✓			EX	3673131	2240	E	0	0	2240	3673131	0,000000	0,000000	0,000000
4		✓			EX	473410	27155	F	20	200	27135	473190	0,090909	0,000737	0,001461
5		✓			EX	473410	256	F	4	38	252	473368	0,095238	0,015625	0,026846
6		✓			EX	473410	7842	F	25	213	7817	473172	0,105042	0,003188	0,006188
7		✓			EX	473410	195	F	0	0	195	473410	0,000000	0,000000	0,000000
8		✓			EX	473410	9	F	9	36926	0	436475	0,000244	1,000000	0,000487
9		✓			EX	473410	81	F	0	0	81	473410	0,000000	0,000000	0,000000
10		✓		✓	EX	473410	31	F	0	0	31	473410	0,000000	0,000000	0,000000
11		✓		✓	EX	473410	2	F	0	0	2	473410	0,000000	0,000000	0,000000
12		✓		✓	EX	473410	1	F	0	0	1	473410	0,000000	0,000000	0,000000
13		✓	✓		EX	473410	98851	F	407	5901	98444	467102	0,064521	0,004117	0,007741
14		✓	✓		EX	473410	3	F	0	132	3	473278	0,000000	0,000000	0,000000
15	✓	✓			EX	225493	61	E	0	0	61	225493	0,000000	0,000000	0,000000
16	✓	✓			EX	473410	12	E	0	0	12	473410	0,000000	0,000000	0,000000
17	✓	✓			EX	473410	36	E	0	0	36	473410	0,000000	0,000000	0,000000
18	✓	✓		✓	EX	473410	26	E	0	0	26	473410	0,000000	0,000000	0,000000
19	✓	✓		✓	EX	2416136	12	E	0	0	12	2416136	0,000000	0,000000	0,000000

Tabela B.8: *Resultado da Execução da Técnica Ramada et al.e base de dados IMDB. A e C - Relevantes; B e D - Não Relevantes.*

	S	*EX*	AG	FR	Realizou o processamento desejado?	*Esperados*			Recuperados		Não Recuperados		*precision*	*recall*	*f-measure*
						Universo	Encontrados		A	B	C	D			
20	✓				Não	473410	12895	E	0	0	12895	473410	0,000000	0,000000	0,000000
21	✓				Não	473410	583	F	0	0	583	473410	0,000000	0,000000	0,000000
22	✓				Não	28	2	E	0	0	2	28	0,000000	0,000000	0,000000
23	✓				Não	473410	181	F	0	129	181	473281	0,000000	0,000000	0,000000
24	✓		✓		Não	473410	27	E	0	0	27	473410	0,000000	0,000000	0,000000
25	✓		✓		Não	473410	1	E	0	0	1	473410	0,000000	0,000000	0,000000
26	✓		✓		Não	473410	25195	E	0	0	25195	473410	0,000000	0,000000	0,000000
27	✓			✓	Não	473410	2	E	0	0	2	473410	0,000000	0,000000	0,000000
28	✓			✓	Não	473410	14667	E	0	0	14667	473410	0,000000	0,000000	0,000000
29	✓			✓	Não	316156	2	E	0	0	2	316156	0,000000	0,000000	0,000000
30	✓			✓	Não	473410	2	E	0	0	2	473410	0,000000	0,000000	0,000000
31	✓			✓	Não	2416136	2	E	0	0	2	2416136	0,000000	0,000000	0,000000
32	✓			✓	Não	392541	5	F	0	0	5	392541	0,000000	0,000000	0,000000
33	✓			✓	Não	2136884	12	E	0	0	12	2136884	0,000000	0,000000	0,000000
34	✓			✓	Não	473410	4	E	0	0	4	473410	0,000000	0,000000	0,000000
35			✓		Não	473410	10	F	0	103	10	473307	0,000000	0,000000	0,000000
36			✓		Não	473410	1	E	0	0	1	473410	0,000000	0,000000	0,000000
37			✓		Não	473410	27	F	0	87	27	473323	0,000000	0,000000	0,000000
38			✓		Não	473410	49	E	0	0	49	473410	0,000000	0,000000	0,000000

Tabela B.8: *Resultado da Execução da Técnica Ramada et al.e base de dados IMDB. A e C - Relevantes; B e D - Não Relevantes.*

	S	*EX*	AG	FR	Realizou o processamento desejado?	*Esperados*			Recuperados		Não Recuperados		*precision*	*recall*	*f-measure*
						Universo	Encontrados		A	B	C	D			
39			✓		Não	473410	83517	F	0	0	83517	473410	0,000000	0,000000	0,000000
40			✓		Não	473410	1	E	0	0	1	473410	0,000000	0,000000	0,000000
41			✓		Não	473410	2	F	0	214	2	473196	0,000000	0,000000	0,000000
42			✓	✓	Não	338511	1	E	0	0	1	338511	0,000000	0,000000	0,000000
43				✓	Não	338511	690	E	0	0	690	338511	0,000000	0,000000	0,000000
44				✓	Não	338511	3410	E	0	0	3410	338511	0,000000	0,000000	0,000000
45				✓	Não	473410	8	F	0	3	8	473407	0,000000	0,000000	0,000000
46				✓	Não	225493	11	E	0	0	11	225493	0,000000	0,000000	0,000000
47				✓	Não	392557	2	E	0	0	2	392557	0,000000	0,000000	0,000000
48				✓	Não	473410	2	E	0	0	2	473410	0,000000	0,000000	0,000000
49				✓	Não	473410	4	F	0	7	4	473403	0,000000	0,000000	0,000000
50		✓		✓	Não	473410		F	0	3	0	473407	0,000000	0,000000	0,000000
	20	20	13	22									0,007783	0,020474	0,011279

Tabela B.9: *Resultado da Execução da Técnica Ramada et al.e base de dados Mondial. A e C - Relevantes; B e D - Não Relevantes.*

	S	***EX***	**AG**	**FR**	**Realizou o processamento desejado?**	***Esperados***			Recuperados		Não Recuperados		*precision*	*recall*	*f-measure*
						Universo	Encontrados		A	B	C	D			
1		✓			EX	238	6	E	0	0	6	238	0,000000	0,000000	0,000000
2		✓			EX	130	1	E	0	0	1	130	0,000000	0,000000	0,000000
3		✓			EX	130	1	E	0	0	1	130	0,000000	0,000000	0,000000
4		✓			EX	1450	1	E	0	0	1	1450	0,000000	0,000000	0,000000
5		✓			EX	276	1	F	0	0	1	276	0,000000	0,000000	0,000000
6		✓			EX	238	238	F	0	243	238	-5	0,000000	0,000000	0,000000
7		✓			EX	5	1	F	0	0	1	5	0,000000	0,000000	0,000000
8		✓			EX	454	81	F	0	0	81	454	0,000000	0,000000	0,000000
9		✓			EX	1688	2	F	0	0	2	1688	0,000000	0,000000	0,000000
10		✓			EX	238	1	F	0	0	1	238	0,000000	0,000000	0,000000
11		✓		✓	EX	238	111	E	0	0	111	238	0,000000	0,000000	0,000000
12		✓		✓	EX	238	21	E	0	0	21	238	0,000000	0,000000	0,000000
13		✓		✓	EX	34	34	E	0	0	34	34	0,000000	0,000000	0,000000
14		✓		✓	EX	238	45	E	0	0	45	238	0,000000	0,000000	0,000000
15		✓		✓	EX	63	1	E	0	0	1	63	0,000000	0,000000	0,000000
16		✓		✓	EX	3111	1	F	0	0	1	3111	0,000000	0,000000	0,000000
17	✓	✓			EX	454	1	E	0	0	1	454	0,000000	0,000000	0,000000
18	✓	✓			EX	218	9	E	0	0	9	218	0,000000	0,000000	0,000000
19	✓	✓			EX	1450	12	E	0	0	12	1450	0,000000	0,000000	0,000000

Tabela B.9: *Resultado da Execução da Técnica Ramada et al.e base de dados Mondial. A e C - Relevantes; B e D - Não Relevantes.*

	S	***EX***	**AG**	**FR**	**Realizou o processamento desejado?**	***Esperados***			Recuperados		Não Recuperados		*precision*	*recall*	*f-measure*
						Universo	Encontrados		A	B	C	D			
20	✓	✓			EX	382	302	E	0	0	302	382	0,000000	0,000000	0,000000
21	✓	✓	✓		EX	238	1	E	0	0	1	238	0,000000	0,000000	0,000000
22	✓				Não	604	4	E	0	0	4	604	0,000000	0,000000	0,000000
23	✓				Não	3111	5	E	0	0	5	3111	0,000000	0,000000	0,000000
24	✓				Não	238	53	E	0	0	53	238	0,000000	0,000000	0,000000
25	✓				Não	61889	1	E	0	0	1	61889	0,000000	0,000000	0,000000
26	✓				Não	1450	51	E	0	0	51	1450	0,000000	0,000000	0,000000
27	✓				Não	8.008	72	E	0	0	72	8008	0,000000	0,000000	0,000000
28	✓				Não	218	3	E	0	0	3	218	0,000000	0,000000	0,000000
29			✓	✓	Não	238	83	E	0	0	83	238	0,000000	0,000000	0,000000
30	✓		✓		Não	238	158	E	0	0	158	238	0,000000	0,000000	0,000000
31	✓		✓		Não	144	1	E	0	0	1	144	0,000000	0,000000	0,000000
32	✓		✓		Não	218	1	E	0	0	1	218	0,000000	0,000000	0,000000
33	✓		✓		Não	238	1	E	0	0	1	238	0,000000	0,000000	0,000000
34	✓		✓		Não	540	1	E	0	0	1	540	0,000000	0,000000	0,000000
35	✓		✓		Não	238	1	E	0	0	1	238	0,000000	0,000000	0,000000
36	✓			✓	Não	238	30	E	0	0	30	238	0,000000	0,000000	0,000000
37	✓			✓	Não	3111	1	E	0	0	1	3111	0,000000	0,000000	0,000000
38	✓			✓	Não	238	131	E	0	0	131	238	0,000000	0,000000	0,000000

Tabela B.9: *Resultado da Execução da Técnica Ramada et al.e base de dados Mondial. A e C - Relevantes; B e D - Não Relevantes.*

									Recuperados		Não Recuperados		*precision*	*recall*	*f-measure*
	S	***EX***	**AG**	**FR**	**Realizou o processamento desejado?**	***Esperados*** Universo	Encontrados		A	B	C	D			
39	✓			✓	Não	153	4	E	0	0	4	153	0,000000	0,000000	0,000000
40	✓		✓	✓	Não	5	1	E	0	0	1	5	0,000000	0,000000	0,000000
41	✓		✓	✓	Não	276	1	E	0	0	1	276	0,000000	0,000000	0,000000
42			✓		Não	240	1	E	0	0	1	240	0,000000	0,000000	0,000000
43			x		Não	276	1	E	0	0	1	276	0,000000	0,000000	0,000000
44			✓		Não	238	1	E	0	0	1	238	0,000000	0,000000	0,000000
45			✓	✓	Não	272	1	E	0	0	1	272	0,000000	0,000000	0,000000
46			✓		AG	5	1	F	0	2	1	3	0,000000	0,000000	0,000000
47			✓	✓	Não	238	1	E	0	0	1	238	0,000000	0,000000	0,000000
48				✓	Não	144	74	E	0	0	74	144	0,000000	0,000000	0,000000
49				✓	Não	238	28	E	0	0	28	238	0,000000	0,000000	0,000000
50				✓	Não	240	1		0	0	1	240	0,000000	0,000000	0,000000
	24	21	16	18									0,000000	0,000000	0,000000

Anexo I

Código I.1 Trecho código bidirectional

```
for (Map.Entry<Integer,Node> entry : interestSet.entrySet()) {
    n = entry.getValue();
    queries = sqlKey.get(n.getTableName());
    if (queries != null && !queries.isEmpty()) {
        for (String q : queries) {
            try {
                q += " WHERE t1.__search_id = " + n.getSearchID();
                statement = conn.createStatement();
                resultSet = statement.executeQuery(q);
                while (resultSet.next()) {
                    Node connected = set.get(resultSet.getInt(2));
                    if (forwardMap.containsKey(n)) {
                        if (forwardMap.get(n) != null
                        && !forwardMap.get(n).contains(connected))
                            forwardMap.get(n).add(connected);
                    } else {
                        ArrayList<Node> list = new ArrayList<>();
                        list.add(connected);
                        forwardMap.put(n,list);
                    }
                }
            } catch (SQLException e) {
                e.printStackTrace();
            }
        }
    }
}
```

Código I.2 *Script* criação sequência

```
1 CREATE SEQUENCE __search_id_seq
2     INCREMENT BY 1
3     NO MAXVALUE
4     NO MINVALUE
5     CACHE 1;
```

Código I.3 Resultado para a Consulta "Tocantins"da base de dados Mondial usando a técnica *Banks-II*.

```
1  Connected
2  Creating graph from DB. Please Wait...
3  Database Created in: 46 millis
4  Query started at: 11:28
5  Creating Interest Set for "Tocantins". Please Wait...
6  Matched: Tocantins
7  Matched: Tocantins
8  Matched: Tocantins
9  Matched: Tocantins
10 Matched: Tocantins
11 Matched: Tocantins
12 Matched: Tocantins
13 Matched: Tocantins
14 Matched: Tocantins
15 Matched: Tocantins
16 Matched: Tocantins
17 Interest Set Built in: 36 seconds
18 Path built
19 Results:
20 1
21 ----------------------------
22 Tree
23 Score 1.0
24 Root 17319
25 2
26 ----------------------------
27 Tree
28 Score 0.8186218808782963
29 Root 19089
30 3
31 ----------------------------
32 Tree
33 Score 0.8186218808782963
34 Root 19089
35 Father : 19089 Sons : 5415
36 Father : 5415 Sons : Hasn't sons
37
38
39 Global Time: 80 seconds
```

Código I.4 Resultado para a Consulta "Oceans name"da base de dados Mondial usando a técnica *Banks-II*.

```
1  Connected
2  -----------------
3  Creating graph from DB. Please Wait...
4  -----------------
5  Database Created in: 47 millis
6  -----------------
7  Query started at: 12:03
8
9  Creating Interest Set for "Oceans name". Please Wait...
10 -----------------
11 Interest Set Built in: 26 seconds
12
13 Path built
14
15 Results:
16
17 Global Time: 27 seconds
```

Anexo II

Código II.1 *XML criação grafo esquema*

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<xmljgraph name= "xmlimdb">
        <nodes>
                <node index = "0" name = "actors" pk= "id" x="" y = "" length = "" width = ""
                connected = "roles" />
                <node index = "1" name = "directors" pk= "id" x="" y = "" length = "" width = ""
                connected = "directors_genres;movies_directors" />
                <node index = "2" name = "directors_genres" pk= "null" x="" y = "" length = ""
                width = "" connected = "directors" />
                <node index = "3" name = "movies" pk= "id" x="" y = "" length = "" width = ""
                connected = "movies_directors;movies_genres;roles" />
                <node index = "4" name = "movies_directors" pk= "null" x="" y = "" length = ""
                width = "" connected = "directors;movies" />
                <node index = "5" name = "movies_genres" pk= "null" x="" y = "" length = ""
                width = "" connected = "movies" />
                <node index = "6" name = "roles" pk= "null" x="" y = "" length = "" width = ""
                connected = "actors;movies" />
        </nodes>
        <edges>
                <edge index = "7" startindex = "0" endindex = "6" startname = "actors"
                endname = "roles" startkeys = "id" endkeys = "actor_id" />
                <edge index = "8" startindex = "1" endindex = "2" startname = "directors"
                endname = "directors_genres" startkeys = "id" endkeys = "director_id" />
                <edge index = "9" startindex = "1" endindex = "4" startname = "directors"
                endname = "movies_directors" startkeys = "id" endkeys = "director_id" />
                <edge index = "10" startindex = "3" endindex = "4" startname = "movies"
                endname = "movies_directors" startkeys = "id" endkeys = "movie_id" />
                <edge index = "11" startindex = "3" endindex = "5" startname = "movies"
                endname = "movies_genres" startkeys = "id" endkeys = "movie_id" />
                <edge index = "12" startindex = "3" endindex = "6" startname = "movies"
                endname = "roles" startkeys = "id" endkeys = "movie_id" />
        </edges>
</xmljgraph>
```

Código II.2 XML das tabelas da base de dados

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<schema name = "imdb">
        <table name = "actors" synonyms = "" hyponyms = "" hypernyms = "" semanticid = "" foreignKeys = "roles;movies">
                <attribute name = "id" datatype = "integer" regex = "[0-9]+" synonyms = "" hyponyms = "" hypernyms = "" />
                <attribute name = "first_name" datatype = "string" regex = "[a-z,A-Z,0-9, ]+" synonyms = "" hyponyms = "" hypernyms = "" />
                <attribute name = "last_name" datatype = "string" regex = "[a-z,A-Z,0-9, ]+" synonyms = "" hyponyms = "" hypernyms = "" />
                <attribute name = "gender" datatype = "char" regex = "[a-z,A-Z,0-9, ]+" synonyms = "" hyponyms = "" hypernyms = "" />
        </table>
        <table name = "directors" synonyms = "" hyponyms = "" hypernyms = "" semanticid = "" foreignKeys = "directors_genres;movies_directors;movies">
                <attribute name = "id" datatype = "integer" regex = "[0-9]+" synonyms = "" hyponyms = "" hypernyms = "" />
                <attribute name = "first_name" datatype = "string" regex = "[a-z,A-Z,0-9, ]+" synonyms = "" hyponyms = "" hypernyms = "" />
                <attribute name = "last_name" datatype = "string" regex = "[a-z,A-Z,0-9, ]+" synonyms = "" hyponyms = "" hypernyms = "" />
        </table>
        <table name = "directors_genres" synonyms = "" hyponyms = "" hypernyms = "" semanticid = "" foreignKeys = "directors">
                <attribute name = "genre" datatype = "string" regex = "[a-z,A-Z,0-9, ]+" synonyms = "" hyponyms = "" hypernyms = "" />
                <attribute name = "prob" datatype = "" regex = "[a-z,A-Z,0-9, ]+" synonyms = "" hyponyms = "" hypernyms = "" />
        </table>
        <table name = "movies" synonyms = "" hyponyms = "" hypernyms = "" semanticid = "" foreignKeys = "movies_directors;movies_genres;roles;directors;actors">
                <attribute name = "id" datatype = "integer" regex = "[0-9]+" synonyms = "" hyponyms = "" hypernyms = "" />
                <attribute name = "name" datatype = "string" regex = "[a-z,A-Z,0-9, ]+" synonyms = "" hyponyms = "" hypernyms = "" />
                <attribute name = "year" datatype = "integer" regex = "[0-9]+" synonyms = "" hyponyms = "" hypernyms = "" />
                <attribute name = "rank" datatype = "" regex = "[a-z,A-Z,0-9, ]+" synonyms = "" hyponyms = "" hypernyms = "" />
        </table>
        <table name = "movies_directors" synonyms = "" hyponyms = "" hypernyms = "" semanticid = "" foreignKeys = "directors;movies">
        </table>
        <table name = "movies_genres" synonyms = "" hyponyms = "" hypernyms = "" semanticid = "" foreignKeys = "movies">
                <attribute name = "genre" datatype = "string" regex = "[a-z,A-Z,0-9, ]+" synonyms = "" hyponyms = "" hypernyms = "" />
        </table>
        <table name = "roles" synonyms = "" hyponyms = "" hypernyms = "" semanticid = "" foreignKeys = "actors;movies">
                <attribute name = "role" datatype = "string" regex = "[a-z,A-Z,0-9, ]+" synonyms = "" hyponyms = "" hypernyms = "" />
        </table>
</schema>
```

Código II.3 Resultado para a Consulta "Tocantins"da base de dados Mondial usando a técnica *Keymantic*.

```
1
2  SELECT city.name, city.Population, city.Latitude, city.Longitude
3  FROM city
4  WHERE  lower(city.name) like lower('%Tocantins%')
5
6  SELECT city.name, city.Population, city.Latitude, city.Longitude
7  FROM city
8  WHERE  lower(city.Latitude) like lower('%Tocantins%')
9
10 SELECT city.name, city.Population, city.Latitude, city.Longitude
11 FROM city
12 WHERE  lower(city.Longitude) like lower('%Tocantins%')
13
14 .
15 .
16 .
17
18 SELECT country.Name, country.Code, country.Capital, country.Area,
19 country.Population FROM country
20 WHERE  lower(country.Name) like lower('%Tocantins%')
21
22 SELECT country.Name, country.Code, country.Capital, country.Area,
23 country.Population FROM country
24 WHERE lower(country.Code) like lower('%Tocantins%')
25
26 SELECT country.Name, country.Code, country.Capital, country.Area,
27 country.Population FROM country
28 WHERE  lower(country.Capital) like lower('%Tocantins%')
29
30 SELECT country.Name, country.Code, country.Capital, country.Area,
31 country.Population FROM country
32 WHERE  lower(country.Area) like lower('%Tocantins%')
33 .
34 .
35 .
```

ANEXO III

Anexo III

Código III.1 *Script* criação tabela TME

```
1  --
2  -- Estrutura da Tabela de Metadados para Exposição
3  --
4
5  CREATE TABLE IF NOT EXISTS TME (
6      serial int(11) NOT NULL AUTO_INCREMENT,
7      provider varchar(255) DEFAULT NULL,
8      url varchar(255) DEFAULT NULL,
9      email varchar(255) DEFAULT NULL,
10     oai_set varchar(255) DEFAULT NULL,
11     dispquery enum('false','true') NOT NULL
12     COMMENT 'Disponível para consulta externa',
13     dc_title varchar(255) DEFAULT NULL,
14     dc_creator text,
15     dc_subject varchar(255) DEFAULT NULL,
16     dc_description text,
17     dc_contributor varchar(255) DEFAULT NULL,
18     dc_publisher varchar(255) DEFAULT NULL,
19     dc_date date DEFAULT NULL,
20     dc_type varchar(255) DEFAULT NULL,
21     dc_format varchar(255) DEFAULT NULL,
22     dc_identifier varchar(255) DEFAULT NULL,
23     dc_source varchar(255) DEFAULT NULL,
24     dc_language varchar(255) DEFAULT NULL,
25     dc_relation varchar(255) DEFAULT NULL,
26     dc_coverage varchar(255) DEFAULT NULL,
27     dc_rights varchar(255) DEFAULT NULL,
28     loginbd varchar(15) NOT NULL
29     COMMENT 'Login do banco de dados',
30     passwordbd varchar(15) NOT NULL
31     COMMENT 'Senha de acesso ao banco de dados',
32     PRIMARY KEY (serial)
33  ) ENGINE=MyISAM DEFAULT CHARSET=latin1 AUTO_INCREMENT=1;
```