

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

AUDIR DA COSTA OLIVEIRA FILHO

**Benchmark para Métodos de Consultas
por Palavras-Chave a Bancos de Dados
Relacionais**

Goiânia
2018

**TERMO DE CIÊNCIA E DE AUTORIZAÇÃO 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 nº 9610/98, 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.

1. Identificação do material bibliográfico: ☒ **Dissertação** ☐ **Tese**

2. Identificação da Tese ou Dissertação:

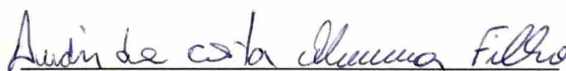
Nome completo do autor: Audir da Costa Oliveira Filho

Título do trabalho: Benchmark para Métodos de Consultas por Palavras-Chave a Bancos de Dados Relacionais

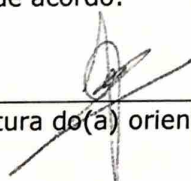
3. Informações de acesso ao documento:

Concorda com a liberação total do documento ☒ **SIM** ☐ **NÃO**¹

Havendo concordância com a disponibilização eletrônica, torna-se imprescindível o envio do(s) arquivo(s) em formato digital PDF da tese ou dissertação.


Assinatura do(a) autor(a)²

Ciente e de acordo:


Assinatura do(a) orientador(a)²

Data: 02 / 08 / 18

¹ Neste caso o documento será embargado por até um ano a partir da data de defesa. A extensão deste prazo suscita justificativa junto à coordenação do curso. Os dados do documento não serão disponibilizados 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

²A assinatura deve ser escaneada.

AUDIR DA COSTA OLIVEIRA FILHO

Benchmark para Métodos 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 Ciência da Computação.

Área de concentração: Ciência da Computação.

Orientador: Prof. Dr. João Carlos da Silva

Goiânia
2018

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

da Costa Oliveira Filho, Audir
Benchmark para Métodos de Consultas por Palavras-Chave
aBancos de Dados Relacionais [manuscrito] / Audir da Costa Oliveira
Filho. - 2018.
90 f.

Orientador: Prof. Dr. João Carlos da Silva.
Dissertação (Mestrado) - Universidade Federal de Goiás, Instituto
de Informática (INF), Programa de Pós-Graduação em Ciência da
Computação, Goiânia, 2018.

1. Consultas por Palavras-chave. 2. Banco de Dados Relacionais.
3. Benchmark. I. da Silva, João Carlos, orient. II. Título.

CDU 004



ATA Nº 04/2018

**ATA DA SESSÃO DE JULGAMENTO DA DISSERTAÇÃO
DE MESTRADO DE AUDIR DA COSTA OLIVEIRA FILHO**

Aos vinte e um dias do mês de junho de dois mil e dezoito, às dez horas, na sala 151 do Instituto de Informática da Universidade Federal de Goiás, Campus Samambaia, reuniu-se a banca examinadora designada na forma regimental pela Coordenação do Curso para julgar a dissertação de mestrado intitulada “**Benchmark para Métodos de Consultas por Palavras-Chave a Bancos de Dados Relacionais**”, apresentada pelo aluno Audir da Costa Oliveira Filho como parte dos requisitos necessários à obtenção do grau de Mestre em Ciência da Computação, área de concentração Ciência da Computação. A banca examinadora foi presidida pelo orientador do trabalho de dissertação, Professor Doutor João Carlos da Silva (INF/UFG), tendo como membros os Professores Doutores Leonardo Andrade Ribeiro (INF/UFG) e Auri Marcelo Rizzo Vincenzi (DC/UFSCar). O prof. Auri participou a distância por webconferência. Aberta a sessão, o candidato expôs seu trabalho. Em seguida, o aluno foi arguido pelos membros da banca e:

(☒) tendo demonstrado suficiência de conhecimento e capacidade de sistematização do tema de sua dissertação, a banca concluiu pela **aprovação** do candidato, sem restrições.

(☐) não tendo demonstrado suficiência de conhecimento e capacidade de sistematização do tema de sua dissertação, a banca concluiu pela **reprovação** do candidato.

Os trabalhos foram encerrados às 13:00 horas. Nos termos do Regulamento Geral dos Cursos de Pós-Graduação desta Universidade, lavrou-se a presente ata que, lida e julgada conforme, segue assinada pelos membros da banca examinadora.

Prof. Dr. João Carlos da Silva _____

Prof. Dr. Leonardo Andrade Ribeiro _____

Prof. Dr. Auri Marcelo Rizzo Vincenzi _____

[Assinaturas manuscritas]

AUDIR DA COSTA OLIVEIRA FILHO

Benchmark para Métodos de Consultas por Palavras-Chave a Bancos de Dados Relacionais

Dissertação defendida no Programa de Pós-Graduação do Instituto de Informática da Universidade Federal de Goiás como requisito parcial para obtenção do título de Mestre em Ciência da Computação, aprovada em 21 de Maio de 2018, pela Banca Examinadora constituída pelos professores:

Prof. Dr. João Carlos da Silva
Instituto de Informática – UFG
Presidente da Banca

Prof. Dr. Leonardo Andrade Ribeiro
Instituto de Informática – UFG

Prof. Dr. Auri Marcelo Rizzo Vincenzi
Departamento de Computação – UFSCAR

Todos os direitos reservados. É proibida a reprodução total ou parcial do trabalho sem autorização da universidade, do autor e do orientador(a).

Audir da Costa Oliveira Filho

Graduou-se em Ciência da Computação na Universidade Católica de Goiás (2008). Durante a graduação, participou de dois projetos de iniciação científica, sendo bolsista do CNPq em um deles no departamento de Computação. Especializou-se em Qualidade e Gestão de Software pela Pontifícia Universidade Católica de Goiás (2010). Atualmente é professor efetivo no Instituto Federal de Educação, Ciência e Tecnologia de Goiás (IFG), lotado no campus Luziânia.

Aos meus pais.

Agradecimentos

Agradeço ao professor Dr. João Carlos da Silva pela orientação, disponibilidade, paciência e pela confiança depositada em mim. Sou muito grato por essa grande contribuição dada durante esse período. Ao senhor, meu mais sincero obrigado.

À meu pai e minha mãe por todo apoio. Vocês acreditam e sempre investem em mim com todo amor. O suporte e educação dado por vocês sempre foi sem medida. Agradeço minha irmã por todo carinho em todas minhas decisões e desafios.

Ao colega de mestrado Walisson Pereira de Souza, pela ajuda em toda a fase do mestrado, pelas discussões, questionamentos levantados e pelos desabafos. Sou muito grato a todo apoio que me deu durante a escrita desta dissertação.

À colega Mariana Soller Ramada pela atenção e disponibilidade para esclarecer minhas dúvidas e me auxiliar nas etapas do projeto.

Aos colegas do curso, que, sempre presentes no laboratório 254, me motivaram para alcançar esse objetivo. Obrigado por partilhar de forma carinhosa esses momentos de estudos, alegrias e tristezas.

Aos colegas docentes do Instituto Federal de Goiás do campus Luziânia, em especial aos professores Leonardo, Luiz Loja, Christiane, Daniel Lucena e Mariângela pelo companheirismo e por acreditarem em mim.

Aos amigos Lilia Moraes, Fernando Maciel e Wagner pela amizade, pelos conselhos, pelo apoio, e por sempre estarem presentes ouvindo meus desabafos durante esse período.

A amiga sempre presente Larissa, por todo carinho e atenção o por compartilhar seu tempo ouvindo meus desabafos e angústias.

A todos os meus amigos e familiares que mesmo distantes sempre torceram e me apoiaram dando força nos momentos difíceis.

Agradeço a todos que contribuíram, de forma direta e indireta, para a conclusão deste trabalho.

“A força não provém da capacidade física. Provém de uma vontade indomável.”

Mahatma Gandhi,

.

Resumo

Oliveira Filho, Audir da Costa. **Benchmark para Métodos de Consultas por Palavras-Chave a Bancos de Dados Relacionais**. Goiânia, 2018. 90p. Dissertação de Mestrado. Instituto de Informática, Universidade Federal de Goiás.

Técnicas de consultas por palavras-chave se mostraram muito eficazes devido à sua facilidade de utilização por usuário na *Web*. Contudo, grande parte dos dados estão armazenados em bancos de dados relacionais, sendo necessário conhecimento de uma linguagem estruturada para acesso a esses dados. Nesse sentido, durante a última década alguns trabalhos foram propostos com intuito de realizar consultas por palavras-chaves a bancos de dados relacionais. No entanto, os sistemas que implementam essa abordagem foram validados utilizando métodos *ad hoc* com bancos de dados que podem não refletir as cargas utilizadas no mundo real. O presente trabalho propõe um *benchmark* para avaliação dos métodos de consultas por palavras-chave a bancos de dados relacionais definindo uma forma padronizada com cargas de trabalhos condizentes com a do mundo real. Esta proposta auxilia na avaliação de eficácia dos sistemas atuais e futuros. Os resultados obtidos com a aplicação do *benchmark* sugerem que ainda existe muitas lacunas a serem tratadas pelas técnicas de consultas por palavras-chave.

Palavras-chave

Consulta com palavras-chave, Banco de dados relacionais, Benchmark.

Abstract

Oliveira Filho, Audir da Costa. **Benchmark for Query Methods by Keywords to Relational Databases**. Goiânia, 2018. 90p. MSc. Dissertation. Instituto de Informática, Universidade Federal de Goiás.

Keyword query techniques have been proven to be very effective due of their user-friendliness on the Web. However, much of the data is stored in relational databases, being necessary knowledge of a structured language to access this data. In this sense, during the last decade some works have been proposed with the intention of performing keyword queries to relational databases. However, systems that implement this approach have been validated using ad hoc methods that may not reflect real-world workloads. The present work proposes a benchmark for evaluation of the methods of keyword queries to relational databases defining a standardized form with workloads that are consistent with the real world. This proposal assists in assessing the effectiveness of current and future systems. The results obtained with the benchmark application suggest that there are still many gaps to be addressed by keyword query techniques.

Keywords

Keyword Search, Relational Database, Benchmark

Sumário

Lista de Figuras	14
Lista de Tabelas	15
1 Introdução	16
1.1 Contexto e Motivação	16
1.2 Principais Problemas	18
1.3 Objetivos	19
1.4 Metodologia	19
1.5 Organização do Texto	21
2 Fundamentação Teórica	22
2.1 Técnicas de Buscas por Palavras-chave a Bancos de Dados Relacionais	22
2.1.1 Grafo de Dados	23
2.1.2 Grafo de Esquema	25
2.1.3 Acesso Prévio aos Dados	27
2.2 Técnicas de <i>Benchmark</i>	30
2.3 Considerações Finais	32
3 Trabalhos Relacionados	33
3.1 Métodos de Avaliação	33
3.2 <i>Benchmarks</i> para Consultas por Palavras-Chave	40
3.3 Considerações Finais	42
4 Proposta de <i>Benchmark</i>	43
4.1 Bancos de Dados	44
4.2 Consultas	46
4.3 Avaliação de Relevância	49
4.4 Métricas	50
4.5 Comparação com Outra Abordagem	52
4.6 Considerações Finais	54
5 Resultados da Avaliação	56
5.1 Etapas da Avaliação	56
5.2 Implementação e Configuração do Experimento	57
5.3 Resultados	58
5.3.1 Quantidade de Configurações	59
5.3.2 TOP-1 Resultados	60
5.3.3 <i>Mean Reciprocal Rank (MRR)</i>	60

5.3.4	Precisão em $n(P@n)$	62
5.3.5	Average Precision (AP)	65
5.3.6	Mean Average Precision (MAP)	66
5.4	Discussão dos Resultados	67
5.5	Considerações Finais	68
6	Conclusão	69
6.1	Contribuições	70
6.2	Trabalhos Futuros	70
	Referências Bibliográficas	72
A	Consultas e Sentidos Pretendidos	77
B	Precisão $n(P@n)$ das Técnicas Avaliadas	82

Lista de Figuras

1.1	Gerenciadores de bancos de dados mais populares, extraído de [40]	17
2.1	Banco de dados IMDB [35]	23
2.2	Grafo de Dados gerado a partir do banco de dados IMDB da Figura 2.1	24
2.3	Grafo de Dados do banco de dados da Figura 2.1 com termos incididos da consulta “actor name Hugo movies”	24
2.4	Steiner Tree gerada para a consulta “actor name Hugo movies”	25
2.5	Grafo do esquema gerado a partir do banco de dados IMDB da Figura 2.1	25
2.6	Grafo do esquema do banco de dados IMDB da Figura 2.1 com termos incididos da consulta “actor name Hugo movies”	26
2.7	Candidate Networks para a consulta “actor name Hugo movies”	26
2.8	Etapas de conversão da consultas por palavras-chave do Keymantic (inspirado em [7])	29
2.9	Matriz de pesos extraído de [7])	29
3.1	Esquema do banco de dados TPC-H extraído de [43]	37
3.2	Tempo de execução de Keymantic para diferentes tamanhos de bancos de dados e diferentes tipos de consultas. Gráfico extraído de [7]	39
4.1	Esquema do banco de dados IMDB, adaptado de [2]	45
4.2	Esquema do banco de dados DBLP, adaptado de [2]	46
4.3	Esquema do banco de dados Northwind, adaptado de [2]	47
4.4	Esquema do banco de dados Mondial, adaptado de [1]	48
4.5	Esquema do banco de dados Wikipedia porposto por Coffman e Weaver[12]	53
5.1	Etapas de Avaliação do Benchmark	56
5.2	Quantidade média das configurações geradas	59
5.3	Percentual de Top-1 Resultados relevantes (%)	60
5.4	Mean Reciprocal Rank	63
5.5	Precisão @ 5	63
5.6	Precisão @ 10	64
5.7	Variação dos resultados de Average Precision aplicadas Técnica Keymantic[7]	65
5.8	Variação dos resultados de Average Precision aplicadas a Técnica de Ramada et al.[39]	66
5.9	Mean Average Precision (MAP)	66

Lista de Tabelas

3.1	Avaliações Realizadas pelos Métodos	34
3.2	Bancos de dados utilizados na avaliação de DBXplorer [3]	35
3.3	Características dos bancos de dados utilizados no trabalho de Coffman e Weaver [12]	41
4.1	Características dos bancos de dados utilizados.	44
4.2	Estatísticas sobre as consultas	49
4.3	Mapeamentos considerados relevantes para consulta "movies genres drama"	50
4.4	Comparação do método proposto com outra abordagem	52
5.1	Síntese da quantidade de consultas finalizadas de um total de 50 consultas executadas	58
5.2	Reciprocal Rank de cada consulta da Técnica Ramada et al [39] (Nº = número da consulta, D = DBLP, I = IMDB, M = Mondial e N = Northwind)	61
5.3	Reciprocal Rank de cada consulta da Técnica Keymantic [7] (Nº = número da consulta, D = DBLP, I = IMDB, M = Mondial e N = Northwind)	62
A.1	Consultas para o Banco de Dados DBLP	78
A.2	Consultas para o Banco de Dados IMDB	79
A.3	Consultas para o Banco de Dados Mondial	80
A.4	Consultas para o Banco de Dados Northwind	81
B.1	Precisão $n(P@n)$ da técnica Keymantic[7] no banco de dados DBLP	83
B.2	Precisão $n(P@n)$ da técnica Keymantic[7] no banco de dados DBLP	84
B.3	Precisão $n(P@n)$ da técnica Keymantic[7] no banco de dados Mondial	85
B.4	Precisão $n(P@n)$ da técnica Keymantic[7] no banco de dados Northwind	86
B.5	Precisão $n(P@n)$ da técnica Ramada[39] no banco de dados DBLP	87
B.6	Precisão $n(P@n)$ da técnica Ramada[39] no banco de dados IMDB	88
B.7	Precisão $n(P@n)$ da técnica Ramada[39] no banco de dados Mondial	89
B.8	Precisão $n(P@n)$ da técnica Ramada[39] no banco de dados Northwind	90

Introdução

Este capítulo apresenta uma visão introdutória a respeito do problema a ser tratado por este trabalho, destacando a abordagem proposta e os objetivos a serem alcançados. Na Seção 1.1 são apresentados o contexto e a motivação que deram origem a este trabalho. Em seguida, a Seção 1.2 enumera os problemas encontrados durante a pesquisa. A Seção 1.3 destaca os objetivos definidos, a Seção 1.4 especifica a metodologia empregada para que fossem alcançados os objetos propostos e, por fim, a Seção 1.5 descreve a organização do texto.

1.1 Contexto e Motivação

Com a evolução da indústria de *hardware* e *software*, uma quantidade imensurável de dados são gerados e armazenadas nos mais diversos bancos de dados [18]. Isso representa uma mudança expressiva nos meios de comunicação devido à enorme possibilidade de disseminação do conhecimento. Dados que antes eram mantidos em espaços físicos restritos passaram a ser guardados virtualmente em fontes distintas disseminadas pelo globo.

Uma parte significativa destes dados estão armazenados em documentos (páginas Web, vídeos ou outros arquivos digitais) para os quais não existe uma definição do modelo e de uma estrutura logicamente coerente dos dados com um determinado significado inerente. Outra parte, não menos significativa, está armazenada em bancos de dados predominantemente de natureza relacional, em que os dados estão organizados em coleções de dados inter-relacionadas.

Com a *Internet*, surge então o paradigma de consultas por palavras-chave, visto que este mecanismo facilita a obtenção dos dados, pois não requer conhecimento sobre como as informações estão organizadas logicamente. No entanto, essas técnicas devem ser adaptadas para que possam ser utilizadas para retornar dados de bancos de dados relacionais, considerando que a obtenção dos dados armazenados em sistemas relacionais exigem o conhecimento de uma linguagem de consulta específica, por exemplo a linguagem SQL (*Structured Query Language*).

Apesar do surgimento de diversos outros tipos de repositórios de dados, grande parte dos dados existentes nas organizações são armazenados ainda hoje em repositórios relacionais [40]. Na Figura 1.1 é possível observar que dos 10 (dez) gerenciadores de bancos de dados mais utilizados a maioria deles implementam a abordagem relacional. Vale destacar inclusive que, os três primeiros bancos de dados mais utilizados, que são relacionais, possuem uma utilização significativamente maior que os demais.

Rank			DBMS	Database Model	Score		
Feb 2018	Jan 2018	Feb 2017			Feb 2018	Jan 2018	Feb 2017
1.	1.	1.	Oracle +	Relational DBMS	1303.28	-38.66	-100.55
2.	2.	2.	MySQL +	Relational DBMS	1252.47	-47.24	-127.83
3.	3.	3.	Microsoft SQL Server +	Relational DBMS	1122.04	-26.03	-81.42
4.	4.	4.	PostgreSQL +	Relational DBMS	388.38	+2.19	+34.70
5.	5.	5.	MongoDB +	Document store	336.42	+5.47	+0.92
6.	6.	6.	DB2 +	Relational DBMS	189.97	-0.30	+2.07
7.	7.	8.	Microsoft Access	Relational DBMS	130.07	+3.37	-3.32
8.	9.	10.	Redis +	Key-value store	127.02	+3.88	+12.98
9.	10.	11.	Elasticsearch +	Search engine	125.32	+2.76	+17.01
10.	8.	7.	Cassandra +	Wide column store	122.78	-1.10	-11.60

Figura 1.1: Gerenciadores de bancos de dados mais populares, extraído de [40]

Devido a popularidade dos bancos de dados relacionais, é altamente relevante pensar em mecanismos que facilitam as consultas nessas bases de informações, sendo assim, surgem várias pesquisas propondo técnicas para realizar consultas por palavras-chave a bancos de dados relacionais. Essas técnicas realizam o mapeamento de cada palavra-chave da consulta com algum termo do banco de dados, por exemplo, nomes de relações, atributos ou tuplas. Contudo, as técnicas propostas ainda não obtiveram um nível aceitável de eficácia para serem utilizadas em bancos de dados reais.

Existem alguns motivos que podem ser elencados que responderiam o porquê da não implantação dos sistemas propostos por essas pesquisas, dentre as barreiras encontradas, destacam-se as avaliações *ad hoc* incompletas realizadas. Como não existem uma forma padrão de realizar a validação dos experimentos, as pesquisas realizadas nessa área utilizam consultas e bancos de dados que não representam o contexto dos dados do mundo real [12].

Dois tipos de técnicas de consultas por palavras-chave foram propostas para bancos de dados relacionais, sendo elas: 1) técnicas que requerem acesso aos dados do banco de dados na etapa de interpretação da consulta. Nessas técnicas são construídos índices sobre o conjunto de dados para que seja possível identificar as tuplas que se relacionam a cada palavra-chave. 2) técnicas que não necessitam de acesso aos dados na etapa de interpretação das consultas. Para essas técnicas o processo de tradução das consultas é feito utilizando-se somente os metadados do banco de dados.

Coffman e Weaver [12] propuseram um *benchmark* para avaliar o primeiro grupo, porém, tal avaliação não se adéqua completamente às técnicas que não possuem acesso aos dados. No entanto, esse segundo grupo possui um papel importante, visto que, por meio deles é possível realizar consultas a bancos de dados relacionais na *Web*, preservando a segurança dos dados armazenados assim como é feito atualmente pelos motores de busca da *Web*. Ou seja, o proprietário de um banco de dados, por meio desse método, pode disponibilizar sua base dando acesso somente ao esquema do banco de dados e só quando solicitado, responder a uma determinada consulta mantendo com isso o sigilo dos dados.

Sendo assim, este trabalho tem por objetivo apresentar uma forma de avaliação para sistemas de consultas por palavras-chave a banco de dados relacionais que não possuem acesso aos dados durante a fase de interpretação da consulta, buscando ressaltar os pontos que impedem a utilização destas técnicas.

1.2 Principais Problemas

Apesar da atual mudança na forma de realizar consultas de dados pela Internet e do sucesso dos motores de busca em tornar informações acessíveis, esta técnica não evoluiu tanto para busca por dados estruturados [13].

Muitas pesquisas foram propostas nesta área de consultas por palavras-chave a bancos de dados relacionais durante a última década, porém tais pesquisas foram validadas utilizando experimentos *ad hoc* que podem, além de não refletir as cargas de trabalho do mundo real, impedir a repetibilidade experimental devido aos escassos detalhes fornecidos na literatura [13].

Os sistemas que possibilitam realizar consultas com palavras-chave a bancos de dados relacionais possuem características que os diferem dos demais sistemas, visto que, para obter o resultado de uma consulta precisam criar caminhos de junções entre relações distintas dentro do esquema do banco de dados, que podem gerar muitas combinações de resultados relevantes.

Inexiste na literatura *benchmarks* exclusivos para avaliação de sistemas que utilizam somente os metadados para formulação das consultas. Desta forma, trabalhos publicados nesta área não possuem uma padronização com relação a bases de dados, consultas e métricas utilizadas para avaliação desses sistemas que possuem características específicas.

Outro problema encontrado durante a pesquisa consiste na granularidade correta dos resultados de pesquisa. Em sistemas que realizam consultas em dados não estruturados, um documento de texto pode conter um único elemento que é pertinente para determinada consulta. Em bancos de dados relacionais uma consulta pode possuir muitos dados

relacionados. Por exemplo, uma pesquisa simples a um artigo no banco de dados DBLP deve retornar apenas a informação deste artigo e não a bibliografia completa relacionada.

Visto que os dados em um banco de dados relacional são armazenados em tabelas inter-relacionadas, onde, em alguns casos, os dados podem estar normalizados para eliminar redundâncias, a identificação de um dado relevante para uma determinada consulta se torna mais complicada devido ao fato que a obtenção de uma informação depende da junção de relações distintas.

1.3 Objetivos

Considerando as técnicas de consultas por palavras-chave a bancos de dados relacionais e os problemas identificados sobre as formas de avaliações realizadas pelos autores de cada técnica, o objetivo geral deste trabalho consiste em definir um *benchmark* capaz de avaliar a eficácia dos sistemas que possibilitam consultas com palavras-chave a bancos de dados relacionais que não necessitam de acesso *a priori* aos dados para traduzir as consultas para a linguagem SQL.

Com o objetivo geral definido, buscando atingi-los, foram destacados os seguintes objetivos específicos:

- Identificar e analisar as abordagens de avaliação utilizadas pelos autores para aferir a eficácia de suas técnicas de consultas por palavras-chaves a bancos relacionais;
- Identificar e avaliar bancos de dados, consultas e métricas que consigam melhor extrair os pontos fortes e fracos das técnicas a serem avaliadas.
- Criar um meio para disponibilizar a comunidade científica todos os dados necessários para utilização do método proposto;

1.4 Metodologia

Visando atingir os objetivos dessa pesquisa, foi definido uma metodologia de trabalho composta pelas seguintes etapas: fundamentação teórica, revisão bibliográfica, proposição da abordagem, execução e obtenção dos dados e documentação. As etapas apresentadas serão melhores descritas a seguir.

- **Fundamentação Teórica**

Esta etapa envolve a busca e estudo, para compreensão, de conceitos que norteiam a pesquisa. Tem como objetivo reunir um conjunto de referências sobre os principais conceitos envolvidos na pesquisa, suas questões e habilitar o entendimento para o

estado da arte. Sendo assim, assuntos como bancos de dados relacionais, interpretação de consultas, recuperação de informação, métricas para avaliação e *benchmarks* foram estudados nessa fase dando um melhor embasamento para a pesquisa.

- **Revisão Bibliográfica**

Nesta etapa do trabalho, foi realizada uma revisão bibliográfica tendo em vista reunir e avaliar os dados disponíveis para obtenção de evidências e formalizar o conhecimento sobre o assunto proposto por esse projeto.

Foram consultados trabalhos publicados, em periódicos e conferências, nas línguas inglesa e portuguesa. As bases de dados consultadas foram : *Portal da Capes*¹, *ACM Digital Library*², *IEEEExplore*³, *Web of Science*⁴, *ScienceDirect*⁵ e *Scopus*⁶.

Visando a busca pelos artigos pertinentes ao estudo deste trabalho, foram definidas as seguintes palavras-chave da pesquisa: “*keyword search*”, “*relational database*”, “*benchmark*” e “*evaluation*”. Foram incluídas termos sinônimos e plurais dessas palavras como: “*relational databases*”, “*relational data source*”, “*RDBMS*”, “*DBMS*” e “*benchmarks*” e “*evaluations*”

- **Proposição da Abordagem**

Conforme a revisão bibliográfica realizada, foi possível identificar limitações e aspectos positivos tanto nos trabalhos que apresentavam novas técnicas quanto nos trabalhos que propuseram métodos de avaliação destas técnicas. Desta forma, nessa etapa, foi definido uma nova proposta de avaliação que melhor destacasse as características das técnicas a serem avaliadas.

- **Execução e Obtenção dos Dados**

Com a abordagem proposta, esta etapa teve como foco a execução da avaliação das técnicas definidas conforme o objetivo deste trabalho. O experimento foi conduzido conforme planejado e a coleta de dados foi minuciosamente registrada para posterior análise dos resultados.

- **Documentação**

Em meio as etapas definidas, foi desenvolvida a documentação da pesquisa, por meio da escrita de textos acadêmicos como artigos e a própria dissertação, visando a publicação destes em meios científicos.

¹<http://www.periodicos.capes.gov.br>

²<http://dl.acm.org>

³<http://ieeexplore.ieee.org>

⁴<http://webofknowledge.com>

⁵<http://sciencedirect.com>

⁶<http://www.scopus.com>

1.5 Organização do Texto

Além deste capítulo introdutório, o presente trabalho está organizado em outros 5 capítulos. O Capítulo 2 apresenta uma fundamentação teórica sobre técnicas de consultas por palavras-chave a bancos de dados relacionais, seguido pelos trabalhos relacionados que são descritos no Capítulo 3. O Capítulo 4 discute uma proposta de *benchmark*. Posteriormente, o Capítulo 5 traz detalhes da execução da avaliação feita juntamente com resultados obtidos. Finalmente, o Capítulo 6 expõe as conclusões e contribuições obtidas com este trabalho e os possíveis trabalhos futuros.

Fundamentação Teórica

Tendo como base auxiliar o entendimento sobre o pontos tratados neste trabalho, este capítulo apresenta uma fundamentação teórica dos principais conceitos no âmbito de avaliação de sistemas de consultas por palavras-chave.

Desta forma, a Seção 2.1 apresenta uma revisão bibliográfica das principais técnicas de consultas por palavras-chave. A Seção 2.2 conceitua as técnicas de *benchmark* no âmbito da computação. Por fim, a Seção 2.3 apresenta um resumo das discussões sobre as técnicas apresentadas.

2.1 Técnicas de Buscas por Palavras-chave a Bancos de Dados Relacionais

Consulta por palavras-chave é uma técnica que se mostrou muito eficaz e eficiente para busca de informações na Internet. Com a disseminação de motores de busca, surgiram nos últimos anos várias pesquisas que propuseram a utilização dessa técnica para consultas a bancos de dados relacionais. Porém, não existe ainda uma única definição formal das técnicas, ou seja, existem algumas vertentes distintas que buscam solucionar o problema, assim como também inexistente uma definição dos resultados válidos para a consulta [12].

Nessas técnicas, as consultas por palavras-chave são representadas como um conjunto ($n \geq 1$) de termos separados por espaço(s) em branco, onde cada termo é uma palavra-chave e expressa alguma informação a ser obtida sobre valores de tuplas, atributos, relações ou alguma função de agregação em um banco de dados [41].

As consultas submetidas a um banco de dados passam por um processo de tradução, em que são geradas consultas SQL correspondentes. Esse processo de tradução pode gerar muitas interpretações distintas inclusive com consultas que não condizem com a intenção do usuário [38].

Na consulta “*actor name Hugo movies*”, submetida ao banco de dados IMDB da Figura 2.1, o usuário tem como objetivo que seja retornado a lista dos filmes que o ator

com nome Hugo atuou. Na construção dessa consulta, as expressões SQLs geradas devem possuir junções entre as tabelas *Actors*, *Roles* e *Movies*. Dessa forma, o termo “actor” deve ser relacionado com a relação *Actors*, o termo “name” deve ser relacionado ao atributo *Actors.first_name*, o termo “Hugo” é um termo de dado que deve ser relacionado ao atributo *Actors.first_name* e o termo “movies” deve ser relacionado a relação *movies*. No caso, por exemplo, do termo “name” ser relacionado ao atributo *Movies.name* o mapeamento estaria errado.

Actors				Roles			Movie_directors	
id	first_name	Last_name	Gender	actor_id	movie_id	role	director_id	movie_id
a1	Allen 'Farina'	Hoskins	M	a1	m1	Palace Guard	d1	m1
a2	Carrie-Anne	Moss	F	a2	m2	Trinity	d2	m2
a3	Hugo	Weaving	M	a3	m2	Agent Smith	d3	m2

Directors			Directors_genres			Movies_genre	
id	first_name	last_name	directors_id	genre	prob	movie_id	genre
d1	Larry (I)	Semon	d1	Comedy	1	m1	Comedy
d2	Andy	Wachowski	d2	Thriller	0.833333	m1	Family
d3	Larry	Wachowski	d3	Sci-Fi	0.833333	m2	Action
						m2	Thriller
						m2	Sci-Fi

Movies			
id	name	year	rank
m1	Wizard of Oz	1925	4.9
m2	The Matrix	1999	8.5

Figura 2.1: Banco de dados IMDB [35]

Várias propostas adotam estratégias distintas buscando melhorar o desempenho. Dentre as técnicas de consultas por palavras-chave a bancos de dados relacionais pesquisadas, é possível destacar três importantes características, sendo elas técnicas que constroem grafos de dados para mapeamento das tuplas do banco, técnicas que utilizam grafos do esquema e técnicas que possuem ou não acesso prévio aos dados. Essas serão apresentadas nas próximas seções.

2.1.1 Grafo de Dados

Na abordagem de grafo de dados, o banco de dados é modelado em um grafo $G = (V, E)$, em que cada vértice ($v \in V$) corresponde a uma tupla ($t \in T$), conectado a outro vértice por uma aresta ($e \in E$) [41]. Nesse grafo, as arestas correspondem a relacionamentos entre as chaves primárias e estrangeiras das relações envolvidas [30].

A Figura 2.2 apresenta o grafo gerado a partir do banco de dados exibido na Figura 2.1. Nesse grafo cada vértice é formado por uma tupla do banco de dados, desta forma, o primeiro vértice “a1” corresponde a tupla de *id* igual a “a1” da relação *Actors* apresentado na Figura 2.1. Da mesma forma que o vértice “r1” corresponde a primeira tupla da relação *Roles*, que possui o valor do atributo *actors_id* igual a “a1” e

movie_id igual a “*m1*”. A aresta existente entre esses dois vértices é definida conforme o relacionamento entre chave-primaria e estrangeira entre eles.

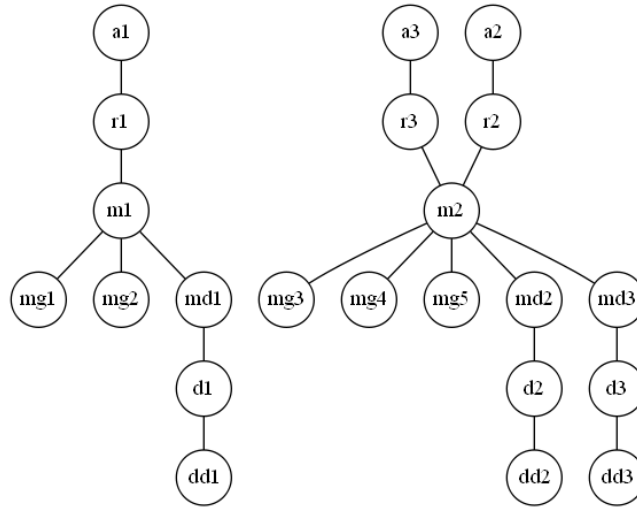


Figura 2.2: Grafo de Dados gerado a partir do banco de dados IMDB da Figura 2.1

Nas técnicas que utilizam grafos de dados, durante o processo de tradução da consulta, são identificadas as *Steiner Trees* que correspondem às subárvores que conectam os nós contendo as palavras-chave da consulta. Uma *Steiner Tree* pode ser definida como uma subárvore T de um grafo $G(V, E)$ contendo todos os vértices V' , onde V' corresponde a um conjunto de vértices $V' \subseteq V$ que representa as palavras-chave da consulta [31].

Dessa forma, ao realizar a consulta “actor name Hugo movies” em um sistema que utiliza grafo de dados, cada palavra-chave é mapeada a um vértice do grafo assim como é apresentado na Figura 2.3.

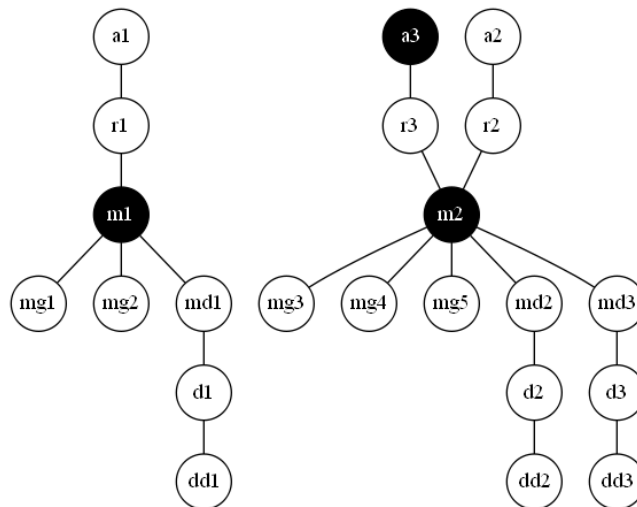


Figura 2.3: Grafo de Dados do banco de dados da Figura 2.1 com termos incididos da consulta “actor name Hugo movies”

A palavra-chave “*Hugo*”, que é um termo de dados, incide sobre o vértice **a3** e o termo “*movie*” acaba gerando incidência sobre o vértice **m1** e **m2**. Devido ao fato de **m1** não possuir ligação com o vértice **a3** esse não pode ser considerado uma *steiner tree*, sendo então desconsiderado. Dessa forma, a árvore apresentada na Figura 2.4 possui um único caminho de junção possível para responder a consulta efetuada. A partir desse caminho são geradas as interpretações possíveis.

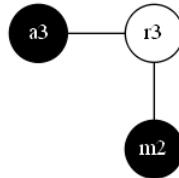


Figura 2.4: *Steiner Tree gerada para a consulta “actor name Hugo movies”*

2.1.2 Grafo de Esquema

Nessa abordagem, as diversas técnicas modelam o esquema do banco de dados em um grafo $G = (V, E)$, nos quais cada vértice $v \in V$ corresponde a uma relação do esquema e o relacionamento entre chaves primária e estrangeira dessas relações formam as arestas $e \in E$ de G [41]. A Figura 2.5 apresenta o grafo gerado a partir do esquema do banco de dados exibido pela Figura 2.1.

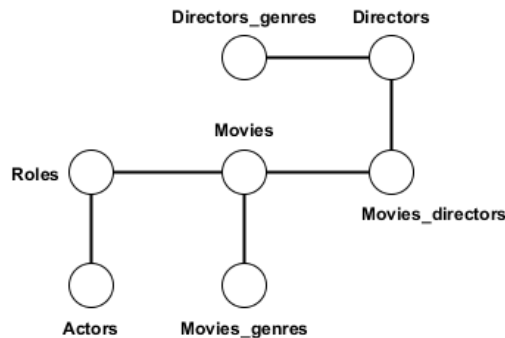


Figura 2.5: *Grafo do esquema gerado a partir do banco de dados IMDB da Figura 2.1*

Uma diferença entre essa abordagem e o grafo de dados consiste no tamanho do grafo gerado. Grafos de esquema são menores e utilizam um espaço menor de memória RAM durante o processamento. Nos grafos de dados, o consumo de memória é elevado se considerar a construção de um grafo de uma base de dados com uma grande quantidade de tuplas o que poderia vir a inviabilizar a utilização da técnica.

Técnicas que utilizam grafos de esquemas, durante o processo de tradução da consulta, identificam Redes Candidatas (*Candidate Networks (CN)*). Dado um grafo

$G(V,E)$, uma CN pode ser definida como uma subárvore contendo um conjunto de vértices $V' \subseteq V$, em que cada vértice de V' corresponde a uma relação em que uma palavra-chave incidiu sobre o grafo G .

Sistemas que utilizam essa abordagem realizam o processo de tradução em três etapas: etapa de identificação dos vértices que incidem com as palavras-chave da consulta, etapa de geração das CNs e etapa de avaliação das CNs. Desta forma, tendo como base a consulta “*actor name Hugo movies*”, na primeira etapa são identificados os atributos, tabelas ou tuplas correspondentes a cada palavra-chave, sendo que esse procedimento utiliza da construção de índices sobre o conjunto de dados ou outros métodos que possuem informações sobre o esquema.

Ainda na primeira etapa, os sistemas identificam os vértices no grafo em que cada palavra-chave incide. A Figura 2.6 apresenta o grafo do esquema com os vértices *movies* e *actor* destacados, pois são eles que correspondem as palavras-chave da consulta.

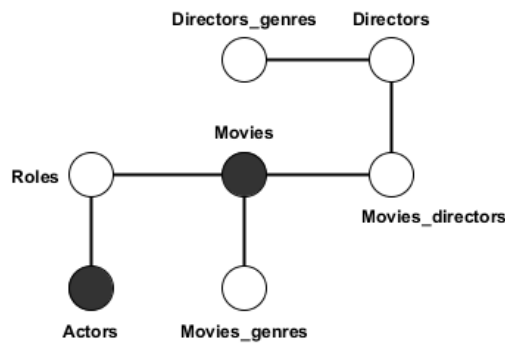


Figura 2.6: Grafo do esquema do banco de dados IMDB da Figura 2.1 com termos incididos da consulta “*actor name Hugo movies*”

Na segunda etapa o sistema identifica, caso exista, os caminhos existentes entre os vértices incididos para formarem as árvores de junção (*Join Trees*), ou redes candidatas. O restante dos vértices que não fazem parte da consulta e não fazem parte das junções são desconsiderados. A Figura 2.7 apresenta as árvores de junção geradas.

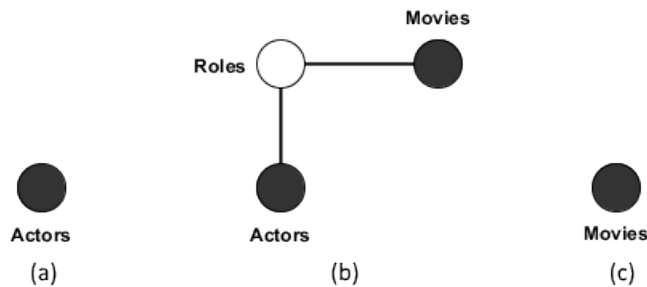


Figura 2.7: Candidate Networks para a consulta “*actor name Hugo movies*”

Na etapa de avaliação, cada *Candidate Network* gerada é convertida em consulta SQL correspondente. As consultas são ranqueadas de acordo com o comprimento do caminho de junções, sendo, na sequência, executadas e apresentadas aos usuários.

2.1.3 Acesso Prévio aos Dados

Técnicas de consultas por palavras-chave com acesso aos dados adotam uma estratégia que necessita de acesso *a priori* aos dados durante a interpretação da consulta. Essas técnicas constroem índices especializados sobre a instância do banco de dados para identificar as tuplas correspondentes às palavras-chave da consulta.

BANKS [8] e DISCOVER [25] constroem índices que referenciam tuplas do banco de dados. DBSemSXplorer [17] utiliza índice invertido que referencia os termos das tuplas e dos metadados. DBXplorer [3] e MeanKS [28] utilizam tabelas auxiliares para armazenar informações das tuplas antes do processo de conversão da consulta. BLINKS [24] constrói um grafo de dados particionado em blocos e realiza o processamento da consulta utilizando um índice de dois níveis para acelerar a pesquisa.

Em contraponto, existem situações onde não é possível a obtenção de acesso aos dados para realização das consultas, sendo um exemplo dados armazenados na Web invisível [6]. Sendo assim, surgem pesquisas que propõem técnicas que não necessitam de acesso aos dados durante a interpretação da consulta.

Tais técnicas propõem a realização de consultas por palavras-chave a bancos de dados relacionais com acesso somente aos metadados do banco de dados. O método extrai os metadados e gera diferentes mapeamentos das palavras-chave da consulta para o vocabulário de um banco de dados. Os trabalhos de Bergamaschi et al. [7] e Ramada et al. [39] seguem essa abordagem.

Um banco de dados relacional D é constituído de um conjunto de tabelas $R(A_1, A_2, \dots, A_n)$, onde R é o nome da tabela e A_1, A_2, \dots, A_n são seus atributos. Utilizando-se dessa designação Bergamaschi et al. [7] apresenta as seguintes definições para demonstrar o problema de consulta por palavras-chave.

Definição 2.1. Um *termo* do banco de dados k constitui um elemento de um vocabulário do banco de dados D , definido como V_d .

Definição 2.2. Uma *consulta* com palavras-chave q é composta por um conjunto de palavras, onde cada uma delas pode ser mapeada para um elemento de interesse do banco de dados.

Definição 2.3. Um *mapeamento* M é definido como a correlação de uma palavra-chave da consulta q com o respectivo *termo* do vocabulário do banco de dados D .

Durante o processo de mapeamento de uma consulta por palavras-chave, cada palavra da consulta pode ser mapeada para um nome de relação, atributo e valores de tuplas ou alguma função do banco de dados. Esse mapeamento classifica as palavras-chave em termo de esquema ou termo de valor.

Quando uma palavra-chave é mapeada para um nome de relação ou nome de atributo do banco de dados, esta palavra-chave é classificada como termo de esquema, caso contrário, se a palavra-chave não possui um mapeamento de esquema, ela passa a ser relacionada a domínio de atributo como um termos de valor. O mapeamento dos termos de valor é realizado em função dos termos de esquema próximos, presentes na consulta.

Seguindo essas definições, consultas para sistemas que não possuem acesso aos dados durante o processo de interpretação da consulta devem possuir pelo menos uma palavra-chave que seja mapeada para um termo de esquema, visto que existe a necessidade de mapeamento e contextualização desses termos. A existência de somente termos de valor na consulta impossibilita a associação aos termos de esquema do banco de dados.

Definição 2.4. Uma *configuração* C é um conjunto de *mapeamentos* definidos entre as palavras-chaves da consulta q para cada termo do banco de dados.

Uma consulta por palavras-chave pode gerar muitas configurações distintas, isso acontece porque uma palavra-chave pode ser mapeada para nomes de relações, atributos ou tuplas diferentes dependendo da estrutura do esquema do banco de dados utilizado na consulta.

Definição 2.5. Uma *interpretação* é uma consulta SQL criada a partir de uma configuração da consulta por palavras-chave.

Cada configuração gerada de uma determinada consulta pode criar muitas interpretações dependendo do caminho de junções existentes entre as relações mapeadas pela consulta. Desta forma, a quantidade de interpretações depende de quantos caminhos de junções diferentes existem entre as relações mapeadas.

O processo de conversão da consulta por palavras-chave proposta pelo sistema Keymantic[7] é realizado em 5 etapas assim como descrito na Figura 2.8. Na primeira etapa o sistema utiliza uma *matriz* $[m][n]$ de pesos intrínsecos, sendo que m corresponde à quantidade de termos da consulta e n o total de termos do esquema mapeados. Sendo assim definidas a matriz SW para cálculo de pesos dos termos de esquema, e a matriz VW para termos de valor. Um exemplo de uma matriz de pesos é apresentado na Figura 2.9.

Os valores dos pesos da matriz SW podem variar de 0 a 100 e são atribuídos conforme o mapeamento realizado, sendo que o termo é mapeado com valor 100 quando houver exatidão na correlação da palavra-chave com o termo do banco de dados, caso o termo não seja exato é atribuído uma pontuação menor. Para a matriz VW, os valores dos

pesos são definidos verificando o domínio do atributo, ou seja, se o termo pertence ou não ao domínio do atributo.

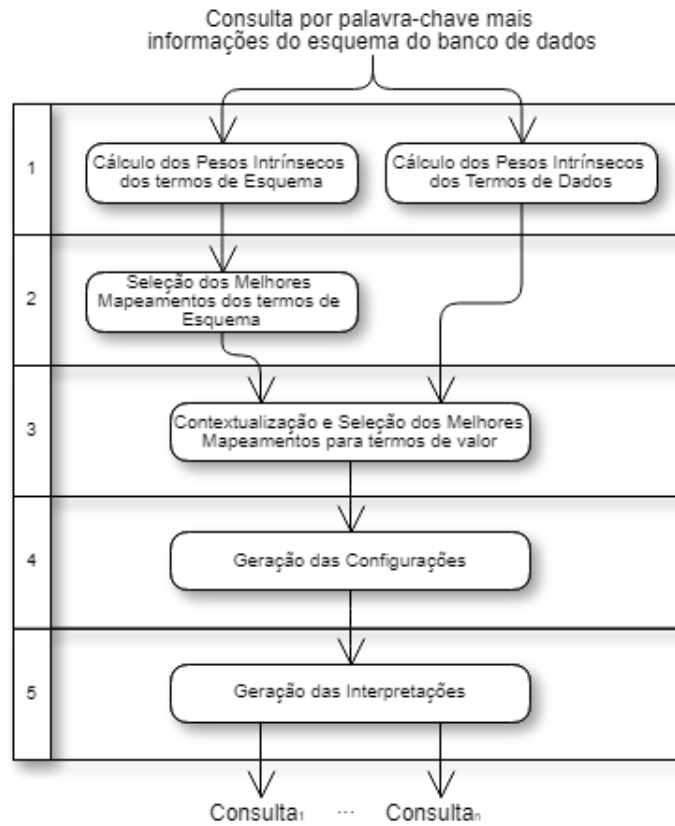


Figura 2.8: Etapas de conversão da consultas por palavras-chave do Keymantic (inspirado em [7])

Na segunda etapa é realizada a seleção dos melhores mapeamentos para os termos de esquema, sendo que os termos com maiores pontuações são provavelmente os que possuem o melhor resultado conforme a consulta. Na sequência, durante a terceira etapa, os termos que não foram mapeados para termos de esquema são mapeados como termos de valor. Como o usuário, geralmente, utiliza termos próximos para definir um sentido para a consulta, os termos de valores são contextualizados conforme o termo de esquema próximos a ele.

	R_1	...	R_r	$A_1^{R_1}$...	$A_{n_1}^{R_1}$...	$A_{n_n}^{R_n}$	$A_1^{R_1}$...	$A_{n_1}^{R_1}$...	$A_{n_n}^{R_n}$
$keyword_1$													
$keyword_2$													
...													
$keyword_k$													

Figura 2.9: Matriz de pesos extraído de [7])

Com os mapeamentos criados, na quarta etapa são geradas as respectivas configurações da consulta. Por fim, na quinta etapa as consultas SQL são geradas para cada

mapeamento produzido pelo método.

Ramada et al. [39] propõem um sistema com base no Keymantic [7], mas adiciona algumas melhorias ao processo de tradução das consultas. Dentre as melhorias estão: reconhecimento de operadores de agregação, realização de ordenação dos resultados utilizando o operador *order by* e melhorias com base nas técnicas de similaridade de *strings*. O método proposto permite a recuperação simplificada de informações presentes em múltiplos bancos de dados.

2.2 Técnicas de *Benchmark*

Um *benchmark*, de forma genérica, pode ser definido como um conjunto de atividades utilizadas para comparar produtos, serviços e processos de trabalho, a fim de identificar, adaptar e adotar práticas para melhorar o desempenho [42]. Na área da computação, é definido como o ato de executar um conjunto de avaliações sobre determinadas aplicações, de modo a obter o desempenho relativo.

Sendo uma técnica padronizada e com testes bem definidos, é amplamente usado para avaliação de desempenho comparando com precisão e eficiência diferentes tipos de produtos [9]. Desta forma, muitos *benchmarkings* são propostos para avaliar não só desempenho de hardware, assim como software, como compiladores ou sistemas de gerenciamento de bancos de dados.

Segundo Gray [20], para ser útil, um *benchmark* deve atender a quatro critérios importantes. Desta forma, deve ser:

- **Relevante:** deve medir o desempenho máximo dos sistemas ao executar operações típicas dentro do domínio avaliado.
- **Portátil:** deve ser fácil implementar o *benchmark* em muitos sistemas e arquiteturas diferentes.
- **Escalável:** o *benchmark* deve se aplicar aos sistemas de computadores pequenos e grandes. Deve ser possível escalar o *benchmark* até sistemas maiores e sistemas informáticos paralelos à medida que o desempenho e a arquitetura do computador evoluem.
- **Simples:** o *benchmark* deve ser compreensível, caso contrário, não terá credibilidade.

Algumas organizações definem *benchmarks* padrões para alguns tipos de domínios específicos. Para a área de banco de dados um dos mais conhecido é o TPC (*Transaction Processing Performance Council*). O TPC é um comitê sem fins lucrativos fundado para definir *benchmarks* para o processamento de transações e avaliação de desempenho de banco de dados [44].

Os *benchmarks* TPC são especialmente interessantes porque definem rigorosamente como os testes devem ser executados, como o preço do sistema deve ser medido e como os resultados devem ser relatados[20]. Sendo assim, foram definidos pelo TPC algumas especificações de *benchmarks* com objetivos distintos dentro da avaliação de sistemas gerenciadores de bancos de dados. Algumas das principais avaliações são apresentadas, de forma detalhada a seguir.

- TPC-C é um *benchmark* para OLTP (*Online Transaction Processing*) que tem como objetivo avaliar as transações realizadas pelo sistema em um ambiente de entrada de pedidos. Este *benchmark* simula um ambiente de computação completo onde vários usuários executam transações em um banco de dados. Dentre as principais métricas dessa técnica estão o **tmpC**, que mede as transações por minuto, e o **\$/tpmC** que aferi o custo associado por transação.
- TPC-DI é um *benchmark* para integração de dados que avalia a análise, combinação e transformação de dados de uma variedade de fontes e formatos em uma representação do modelo de dados unificado.
- TPC-DS é um *benchmark* para medir o desempenho de sistemas de suporte à decisão. Essa avaliação inclui sistemas de *Big Data*, porém não se limita a eles.
- TPC-E é um *benchmark* que simula um OLTP de um ambiente empresarial de uma corretora de clientes que geram transações relacionadas as negociações, pesquisa de conta e pesquisas de mercado. Com essa avaliação é possível configurar um número de cliente para simular cargas de trabalho de diferentes tipos de empresa.
- TPC-H é um *benchmark* de suporte à decisão. Essa avaliação compara sistemas de suporte à decisão que manipulam grandes volumes de dados, executando consultas com um alto grau de complexidade e dão resposta a questões críticas de negócio.
- TPC-VMS é um *benchmark* que avalia o desempenho de bancos de dados virtualizados. Esse método tem por objetivo representar um ambiente de virtualização no qual três cargas de trabalho de bancos de dados são consolidadas em um servidor.
- TPC-BB é um *benchmark* para medir o desempenho dos sistemas *Big Data* baseados em *Hadoop*. Esse método mede o desempenho executando um conjunto de consultas no contexto varejista.

Apesar da quantidade de modelos de *benchmark* definidos pela TPC, nenhum deles atende a especificidades dos sistemas que realizam consultas por palavras-chave a bancos de dados relacionais.

2.3 Considerações Finais

Esse capítulo apresentou uma breve descrição das pesquisas realizadas sobre métodos de consultas por palavras-chave a bancos de dados relacionais. Foram destacadas três características importantes, tais como técnicas que constroem grafos de dados, técnicas que constroem grafos de esquema e diferenças entre técnicas que utilizam ou não acesso prévio aos dados durante a interpretação da consulta.

Para orientação na construção do *benchmark* foram demonstradas as principais características apresentadas por esses métodos de avaliação para que o proposto por este trabalho atenda a seu objetivo. Os *benchmarks* do TPC foram apresentados, porém nenhum deles é utilizado para avaliar a eficácia de sistemas de consultas por palavras-chave.

Trabalhos Relacionados

Este capítulo está dividido em duas partes, sendo que, a Sub-seção 3.1 discute os métodos de avaliação utilizados para aferir cada proposta e na Sub-seção 3.2 são descritos os trabalhos que propõem técnicas de avaliação para métodos de consultas por palavra-chave a bancos de dados relacionais.

3.1 Métodos de Avaliação

Com o objetivo de conhecer como as técnicas realizam a tradução de consultas por palavras-chave em consultas SQL e, principalmente, como esses sistemas aferem a qualidade dos resultados retornados, foi realizada uma revisão bibliográfica com intuito de mapear tais características. A Tabela 3.1 faz um comparativo com as informações de bancos de dados, métricas e quantidade de consultas apresentadas por cada trabalho para realização das avaliações de suas propostas. De acordo com o estudo, notou-se que os sistemas não seguem um padrão para a aferição da qualidade dos resultados retornados.

O sistema BANKS [8] avalia a qualidade dos resultados retornados utilizando a métrica denominada Pontuação por Erro (*error score*). Na avaliação são utilizadas um total de 7 (sete) consultas, executadas em cada um dos seguintes bancos de dados: DBLP, que consiste em um banco de dados com informações acerca de publicações em sua maioria da área de Ciência da Computação, sendo composto com 100.000 (cem mil) tuplas; e IIT Bombay (Instituto Indiano de Tecnologia de Bombay), que armazena informações sobre Dissertações e Teses deste instituto. Neste último banco de dados, os autores não são claros a respeito da quantidade de informações utilizadas nos testes. Além disso, cada consulta foi definida com aproximadamente 4 (quatro) sentidos válidos, sendo que esses sentidos foram definidos pelos próprios autores. Nesta métrica, é definido um valor de *ranking* ideal para cada uma, sendo este considerado o melhor valor, que afere resultados mais promissores. Em seguida, é calculada a diferença entre os valores de *ranking* dos resultados retornados e os valores de *ranking* ideais. O valor retornado, *error score*, pode chegar até 100, considerado o pior resultado para a consulta retornada. Dessa maneira, quanto menor é o *error score*, mais relevante é o resultado. Contudo, a métrica

utilizada nesse sistema é simples e, com as informações fornecidas, não é possível afirmar se pode aferir a qualidade de resultados retornados em sistemas reais. Além de que os conjuntos de testes utilizados para avaliação do sistema são pequenos.

Tabela 3.1: *Avaliações Realizadas pelos Métodos*

Sistema	Banco de Dados	Métricas	Consultas
BANKS [8]	DBLP, IIT Bombay	Pontuação por erro	7 consultas
DBXplorer [3]	TPC-H, USR, ML, KB	Tempo de publicação Tempo de tradução	100 consultas
Discover [25]	TPC-H	Speedup Tempo de Execução	100 Consultas
Ding et al. [14]	DBLP, Mondial	Tempo de Processamento Consumo de memória Custo do peso da Aresta	20 consultas
Nandi e Jagadish [36]	IMDB, MiMI	-	-
Gu e Kitagawa [21]	IMDB	Precisão Top-1 Resultados Relevantes	10 consultas
SQAK [43]	University TPCH	Tempo Tradução Consulta Sensibilidade à disparidade dos termos	15 consultas
Frisk [37]	Northwind, IMDB DBLP, FoodMart	-	-
Keymantic [7]	University IMDB	Top-1 Resultados Relevantes Quantidade Configurações Tempo de Execução	99 consultas 44 consultas
Ham et al. [22]	IMDB	Mean Reciprocal Rank (MRR)	10 Consultas
DBSemSXplorer [17]	IMDB	Precisão / Revocação Mean Reciprocal Rank (MRR)	20 Consultas
MeanKS [28]	TPC-E	Coeficiente de Kendall Tamanho da Sobreposição Top-k	10 Consultas
ExpressQ [45]	TPC-H, ACMDL	Tempo Geração Consultas Tempo Execução Consultas	7 Consultas
Ramada et al. [39]	Company	Quantidade de Configurações Quantidade de Interpretações Mean Reciprocal Rank (MRR) Precisão	10 Consultas

A tradução das consultas com palavras-chave, no sistema DBXplorer [3], é realizada por meio da criação de estruturas externas ao banco de dados. De maneira simplificada, são criadas tabelas que são utilizadas para determinar a localização das palavras-chave da consulta no banco de dados. Essas informações armazenadas determinam o local em que os termos da consulta serão combinados, como nome de tabelas, colunas e linhas. Assim, o processo de conversão é dividido em duas etapas: publicação (*publish*) e busca

(*search*), sendo a primeira etapa responsável por criar estruturas externas para referenciar informações do banco de dados e a etapa de busca, realizada após a submissão da consulta. Além disso, esse sistema realiza a avaliação de três técnicas de consulta com palavras-chave, sendo elas:

- **Pub-Col**, em que são criadas estruturas para armazenar informações a respeito das possíveis combinações de cada palavra-chave da consulta em consonância com as colunas (atributos) das tabelas do banco de dados.
- **Pub-Cell**, em que são mantidas as informações a respeito das possíveis combinações de cada palavra-chave da consulta em relação às células das tabelas do banco de dados.
- **Pub-Prefix** esta técnica permite que sejam realizadas diferentes combinações de *tokens*, por meio da utilização de índices *B+ Trees*.

Na avaliação de DBXplorer [3], conforme visto na Tabela 3.1, foram utilizados os bancos de dados TPC-H, USR, ML e KB sendo esses, melhor detalhados na Tabela 3.2. USR é um banco de dados de endereços de funcionários, ML contém dados sobre listas de discussões e KB armazenam de informações de artigos e manuais de ajuda acerca de diversos produtos. Para cada banco de dados, foram criadas e executadas 100 (cem) consultas com palavras-chave, com quantidade de termos variando de 1 (um) a 5 (cinco), sorteadas aleatoriamente. Além disso, foram verificados o tempo de publicação e o tempo de tradução das consultas. A primeira métrica diz respeito ao tempo necessário para armazenar informações referentes ao esquema do banco de dados, devido à necessidade da técnica do sistema. Ainda, foi analisado o crescimento das estruturas auxiliares, de acordo com cada técnica e verificado o tempo de execução das consultas, inclusive em relação à quantidade de termos.

Tabela 3.2: Bancos de dados utilizados na avaliação de DBXplorer [3]

Banco de Dados	Tipo	Tamanho (MB)	Quantidade de Relações
TPC-H	Sintético	100 a 500	-
USR	Real	130	19
ML	Real	375	38
KB	Real	365	84

Discover [25] utiliza o *benchmark* TPC-H (com tamanho de 100 MB) para realização dos testes em três técnicas de consultas com palavras-chave (Discover, simples e ótimo). Nesse trabalho, foram apresentadas as seguintes definições:

- **Joining network of tuples** - uma árvore de tuplas em que para cada par de tuplas adjacentes existe uma aresta interligando-as. O tamanho dessa árvore é calculado de acordo com o número de caminhos de junção da consulta.
- **Keyword Query** - Conjunto de termos que compõe a consulta com palavras-chave, sendo que o sistema DISCOVER considera como resposta a essa consulta o conjunto de todas as possíveis redes de junção de tuplas, desde que sejam: mínimas, quando cada palavra-chave está contida em pelo menos uma tupla da rede de junção, e totais, quando não for possível remover qualquer tupla da rede de junção e ainda ter uma rede de junção total de tuplas.
- **Joining Network of Tuple Sets** - é uma árvore formada por conjunto de tuplas, em que para cada par de conjuntos de tuplas adjacentes, existe uma aresta interligando-os.
- **Candidate Network** - é uma rede de junção formada por conjuntos de tuplas, em que existam correspondência com os termos da consulta com palavras-chave.

Na avaliação de Discover [25] é realizado a construção das consultas com palavras-chave selecionando aleatoriamente termos advindos do vocabulário do banco de dados TPC (dados, nomes de tabelas ou atributos), assim como a quantidade de termos da consulta gerada. No geral, são realizadas 100 (cem) execuções para cada uma das técnicas avaliadas. Além disso, são comparados o tempo de execução das consultas das técnicas avaliadas e também o *speedup*. Esta última métrica avalia a evolução do tempo gasto em detrimento do aumento dos termos da consulta com palavras-chave.

Ding et al. [14] define um conjunto de 20 consultas por palavras-chave para cada banco de dados. Foram utilizados os bancos de dados DBLP e MDB, sendo que, o DBLP consiste 1,9 milhões registros de publicações científicas até o ano de 2004, e o MDB consiste em 1 milhão de registros de um sistema de recomendações de filmes. Como forma de avaliar a escalabilidade, os testes foram realizados em 10 subconjuntos de tamanhos menores do banco de dados DBLP, sendo em cada um deles realizados testes e avaliado o tempo de processamento, o consumo de memória baseado no número de nós do grafo gerado e o custo baseado no peso das arestas.

Outro teste realizado pelos autores consiste na variação do número de palavras-chave entre 2 a 6 termos. Para cada quantidade de termos, nesse segundo experimento os autores geraram aleatoriamente um conjunto de 100 consultas. É observado que o tempo de processamento não é alterado significativamente para a quantidade de termos, exceto um salto que existe quando passa de 2 para 3 termos na consulta.

Nandi e Jagadish [36] propõem um sistema que utiliza o recurso de completar a palavra-chave que está sendo digitada. Esse recurso preditivo reduz a possibilidade de erros de digitação, ao apresentar informações exatas de tabelas, atributos e valores de tuplas. Para avaliar o sistema em questão, os autores utilizaram 2(dois) bancos de dados:

o IMDb, contendo aproximadamente 100.000 (cem mil) registros de informações sobre filmes, séries, documentários e similares; e o MiMI, que armazena informações acerca de interações moleculares, com tamanho que ultrapassa 1GB e mais de 48 milhões de registros. No entanto, não foram encontradas informações no trabalho sobre quantidade de consultas e quais métricas foram utilizadas e nem como os experimentos foram conduzidos.

Os autores Gu e Kitagawa [21] apresentaram um sistema que permite realizar consultas com palavras-chave em bancos de dados relacionais através da construção de um grafo de dados ponderado. Para avaliarem o sistema proposto, os autores utilizaram o IMDb (Internet Movie Database) através dos arquivos de dados disponíveis no *site* desse repositório, sendo criadas 9 tabelas para submissão dos testes. Além disso, as consultas foram concebidas tendo como base as informações do banco de dados, de acordo com os autores, para evitar que sejam retornados resultados vazios. Foram utilizados dois conjuntos de consultas nos testes, sendo o primeiro consistido de 10 consultas sem informações do esquema, ou seja, sem termos compatíveis com nomes de tabelas e atributos do banco de dados, e o segundo contendo 10 consultas, só que, dessa vez, contendo informações do esquema do banco de dados. A métrica utilizada para aferir a qualidade do sistema compara os Top-10 resultados de cada consulta com palavras-chave submetida e informa o percentual de resultados relevantes de cada uma delas.

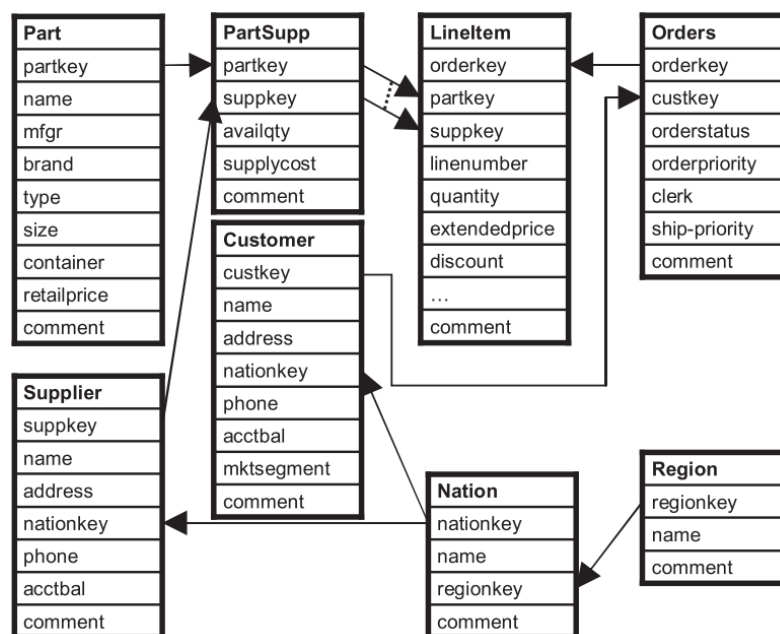


Figura 3.1: Esquema do banco de dados TPC-H extraído de [43]

A técnica SQAK [43] realiza alguns experimentos demonstrando o custo da construção e precisão da tradução das consultas. Para a avaliação, são usados um banco de dados simples de Universidade e o TPC-H, conforme o esquema apresentado na Figura 3.1. São definidos pelos autores 15 consultas para cada banco de dados, contendo

de 3 a 5 termos. Os experimentos mostram se uma consulta retornou ou não resultado relevante, ou seja, realiza a avaliação através da relevância binária. Além disso, são utilizadas outras duas métricas: *Mismatch Sensitivity*, que avalia a capacidade do sistema em retornar informações em relação às combinações dos termos da consulta com palavras-chaves com informações do banco de dados. Isso permite que sejam recuperados dados, mesmo com pequenos erros de digitação. A outra métrica utilizada foi o tempo de tradução da consulta que avalia o tempo que o sistema leva para realizar a tradução das consultas.

Os autores do SQAK [43] realizaram testes em um banco de dados com mais de 600 tabelas contendo informações sobre ativos de Tecnologia da Informação de uma grande empresa e em um *data warehouse* de dados de vendas no varejo com um esquema estrela contendo 14 tabelas. No entanto, apesar de comentar que o sistema possui uma precisão de quase 100% não são apresentados os resultados dos testes realizados.

FRISK [37] propõe uma técnica que modela um grafo de esquema onde são recuperados redes de junção de tuplas *Join Network of Tuples* conforme a incidência das palavras-chave. Um ponto de destaque nessa proposta é a utilização de um algoritmo que elimina palavras indesejadas, identifica erros de ortografia e sinônimas, ou seja, palavras distintas que possuem o mesmo significado.

Os autores apresentam quatro bases de dados para demonstrar a utilização do sistema, sendo eles Nortwind, IMDB contendo 9.839.026 tuplas, DBLP com 881.867 tuplas sobre publicações, autores, títulos e editores. Por último, utiliza o banco de dados FoodMart, que é uma base com 428.049 tuplas com informações sobre produtos e clientes. No trabalho não são apresentados nenhum tipo de informação sobre avaliação do sistema.

Keymantic [7] realizou experimentos para validar a eficiência e a eficácia utilizando dois bancos de dados reais, sendo eles, o IMDB e o segundo um banco de dados universitário que contém informações sobre cursos, professores, alunos, publicações, funcionários e outras informações acadêmicas. Para definição das consultas, os autores solicitaram consultas de usuários reais, tais usuários não tinham conhecimento do esquema do banco de dados. Na sequência, as consultas foram repassadas para um especialista técnica que gerou consultas SQL destas consultas definidas pelos usuário. Essas consultas SQL foram utilizadas para avaliar os resultados retornados pelo sistema. Ao todo foram definidos um total de 99 consultas para o banco universitário e 44 consultas para o IMDB. Para medir a eficácia, foi verificado os resultados pretendidos para cada consulta, com isso, são apresentados a quantidade de configurações geradas. Outro ponto de destaque foi se a resposta esperada estava em primeiro lugar. Os autores destacam que o sistema teve um desempenho pior no banco de dados IMDB devido a estrutura do seu esquema.

Visando medir a eficiência de Keymantic [7], os autores sugerem a definição de dois parâmetros: as consultas e o tamanho do banco de dados. Com relação às consultas,

o tempo de execução foi aferido considerando tanto a quantidade de termos da consulta quanto o tipo, ou seja, se a palavra-chave corresponde a termos de esquema ou valor. No parâmetro tamanho do banco, os autores criam instancias de tamanhos diferentes dos bancos de dados. Essa avaliação é demonstrada através de gráfico como o apresentado na Figura 3.2, sendo que, cada barra do gráfico corresponde a frações de tamanhos diferentes do banco de dados de avaliação original. Foram definidas consultas com palavras-chave de esquema, representado no grafo com “s”, e termos de dados, representados no gráfico com a letra “v”.

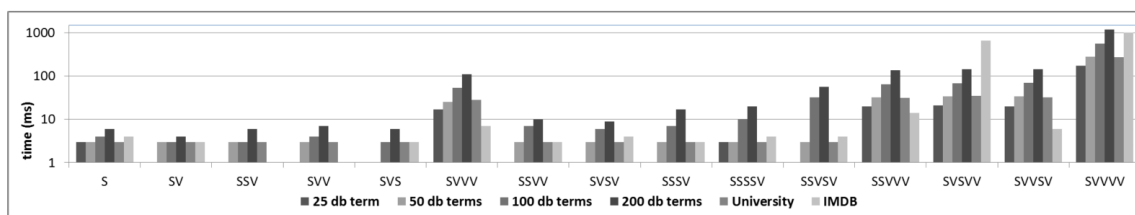


Figura 3.2: Tempo de execução de Keymantic para diferentes tamanhos de bancos de dados e diferentes tipos de consultas. Gráfico extraído de [7]

Haam et al. [22] utilizou o banco de dados de filmes IMDB, na realidade, um subconjunto dos dados desse banco com um total de 4.021.016 tuplas. Nessa avaliação foram definidas 10 consultas, formadas por 2 a 5 termos, conforme interpretações do usuário sobre um determinado assunto. Como métrica de avaliação foi utilizado *mean reciprocal rank* (MRR).

DBSemSXplorer [17] busca eliminar qualquer tendência em relação a avaliação do sistema. Para isso define um conjunto de 20 consultas, onde estas são formuladas por 3 usuários não técnicos e que não conheciam o funcionamento da abordagem proposta. Essas consultas são definidas visando uma variedade de informações de filmes do banco de dados IMDB. Para avaliar o sistema foi utilizado as métricas precisão/revocação e Mean Reciprocal Rank (MRR) dos resultados.

O sistema MeanKS [28] utiliza o TPC-E para validar seu método. O TPC-E é um *benchmark* que simula uma carga de trabalho de processamento de transações *on-line* (OLTP) de uma corretora, que possui um esquema razoavelmente complexo contendo 33 relações.

Os autores não são claros quanto à quantidade de consultas geradas para a avaliação da proposta. No entanto, definem 10 conjuntos de palavras-chave padrão que são usadas como modelo de consultas de entrada. Desta forma, para exemplificar, o primeiro modelo apresentado é “Customer, Company” que, para esse caso, pode resultar da consulta por palavras-chave “Joseph, Andersen” sendo “Joseph” um cliente e “Andersen” uma empresa.

Para validar MeanKS [28] os autores verificam a lista produzida por cada consulta e a compara com uma lista definida como correta. Para verificar a proximidade das listas são utilizados as métricas do coeficiente de Kendall e o tamanho da sobreposição entre os itens top- K . Foram usados cinco e dez como os valores de K sendo a pontuação de desempenho de um método medida pela pontuação média das dez consultas.

A técnica ExpressQ [45] é avaliada por seus autores com a utilização de banco de dados TPC-H e da biblioteca digital de publicações da ACM. Para cada banco de dados defini-se um conjunto de 7 consultas, sendo que as palavras-chave das consultas correspondem a nomes de relações, atributos ou valores de tuplas. Visando verificar a técnica proposta, a avaliação realiza uma comparação com Spark [33], em que as métricas tempo de geração e tempo execução das consultas SQL são utilizadas e os devidos resultados apresentados.

Ramada et al. [39], para validar seu trabalho, utiliza o banco de dados Company proposto por Elmasri e Navathe [16] com um tamanho de esquema de 508 KB. Esse banco de dados armazena informações sobre funcionários e seus respectivos dependentes, departamentos e os projetos nos quais cada funcionário atua. Para esse banco de dados, os autores definiram dez consultas onde a quantidade de termos varia de 1 a 5 termos.

Com esse cenário, são apresentados comparativos com a técnica Keymantic [7]. Para isso os autores utilizam as métricas quantidade de configurações e interpretações geradas por cada técnica, precisão e *Mean Reciprocal Rank*. Conforme resultados, demonstra-se que a proposta de Ramada et al. [39] proporciona um conjunto de resultados menor e mais significativo.

Conforme visto, as técnicas de consultas por palavras-chave a bancos de dados relacionais propostas são validadas por seus autores por meio de métodos não padronizados. Não existe um conjunto de dados padronizado com cargas de dados que reflitam as existentes no mundo real. Mesmo trabalhos que utilizam o mesmo banco de dados estes são subconjuntos de tamanhos diferentes. As consultas utilizadas não atendem a uma quantidade mínima para aferir todas as possibilidades de consultas de um banco de dados.

3.2 Benchmarks para Consultas por Palavras-Chave

Analisando as avaliações utilizadas pelos trabalhos que realizam consultas por palavras-chave a banco de dados relacionais, conforme apresentado na Seção 3.1, foi possível constatar que não existem um modelo padronizado de avaliação. Comparando as avaliações de trabalhos que utilizam o mesmo banco de dados identificam-se que os resultados gerados entre elas não são parecidos, pois os pesquisadores criam subconjuntos aleatórios dos bancos de dados originais [12].

Coffman e Weaver [12] foi o único trabalho encontrado que propõem um *benchmark* para avaliar a eficácia de técnicas que realizam consultas por palavras-chave a bancos de dados relacionais. Os mesmos autores, no trabalho [11], apresentam uma avaliação de desempenho de tempo de execução das técnicas. Nesses trabalhos foram avaliadas apenas técnicas de consultas por palavras-chave que tinham acesso aos dados do banco de dados para a tradução da consulta. Os autores propõem a utilização de três bancos de dados: IMDB, Mondial e Wikipedia, sendo os detalhes de cara um deles apresentados na Tabela 3.3. Para cada banco de dados foi definido um conjunto de 50 consultas juntamente com a definição de uma avaliação de relevância binária para análise dos resultados das consultas. Visando medir a eficácia dos sistemas, os pesquisadores utilizam quatro métricas, sendo elas: o número de resultados relevante do top-1, o *reciprocal rank*, a precisão média de uma consulta e a média da precisão média (MAP).

Tabela 3.3: Características dos bancos de dados utilizados no trabalho de Coffman e Weaver [12]

Banco de Dados	Tamanho (MB)	Quantidade Relações	Quantidade Tuplas
Mondial	9	28	17.115
IMDB	516	6	1.673.074
Wikipedia	550	6	206.318

Esse *benchmark*, foi utilizado na avaliação das técnicas BANKS [8], Discover [25], Efficient [26], Bidirectional [27], Effective [32], DPBF [14], Blinks [24], Spark [33], CD [13]. Os resultados apresentados por Coffman e Weaver com essa avaliação demonstram que a eficácia dos sistemas avaliados é consideravelmente pior que as apresentadas pelo autores em suas propostas.

Por fim, podem ser destacadas duas questões importantes a serem consideradas em trabalhos futuros de consultas por palavras-chave. Primeiramente, a ponderação dos nós dos grafos desempenha um importante papel na obtenção dos resultados. Técnicas como BANKS [8] e Bidirectional [27] obtiveram melhores resultados no banco de dados Wikipedia, enquanto que abordagens como DPBF [14] obtiveram piores resultados devido ao fato de minimizarem a utilização de pesos nos nós das *steiner trees*. A segunda questão destaca problemas de escalabilidade dos sistemas que utilizam grafos de dados. Com exceção da implementação de BANKS [8], nenhuma outra técnica obteve resultados satisfatórios para o banco de dados IMDB, mesmo sendo este um subconjunto reduzido da base de dados original. Isso ocorreu porque, as técnicas avaliadas propõem o armazenamento do grafo de dados completo na memória.

Em um segundo trabalho, Coffman e Weaver [11] propõem a avaliação da eficiência dos sistemas de consultas por palavras-chave a bancos de dados relacionais, utilizando das mesmas definições propostas em [12], porém fazendo uso das métricas

tempo de execução e tempo de resposta. Tempo de execução é o tempo decorrido desde a emissão de uma consulta até o término do algoritmo e tempo de resposta foi definido como sendo o tempo decorrido desde a emissão da consulta até que o resultado tenha sido retornado.

Para essa segunda proposta de avaliação, foram experimentados as técnicas BANKS [8], Discover [25], Discover [25], BANKS-II [27], DPBF [14], Blinks [24], Star [29] para avaliação da eficiência. Os resultados apresentados por Coffman e Weaver demonstram que o desempenho em tempo de execução dos sistemas é ruim devido à grande quantidade de consultas não concluídas com sucesso.

Os resultados dessa avaliação questionam a escalabilidade das técnicas de busca. Os autores destacam que os conjuntos de dados utilizados para a avaliação são pequenos pelos padrões atuais e mesmo assim os sistemas tiveram um consumo elevado de memória principal durante a execução dos testes. Coffman e Weaver [11] afirmam ainda que as técnicas de pesquisas por palavras-chave seriam incapazes de pesquisar em bancos de dados de tamanho moderado ou mesmo inviáveis para base de dados de redes sociais ou registros de saúde médica.

Os trabalhos de Coffman e Weaver [12, 11] não avaliaram sistemas que não possuem acesso prévio aos dados. Devido aos bancos de dados utilizados, e ao conjunto de consultas definidas pelos autores, que não abrangem todos as situações pois não diferenciam termos de esquema e valores, foi possível definir que esse *benchmark* não se adéqua completamente as técnicas de consultas por palavras-chave sem acesso prévio aos dados. Sendo assim, o presente trabalho propõe alterações que consiga avaliar da melhor forma a eficácia desses sistemas.

3.3 Considerações Finais

Este capítulo apresentou as avaliações realizadas em alguns trabalhos para validarem suas técnicas de consultas por palavras-chaves a bancos de dados relacionais. Foi possível destacar por meio desta revisão que as técnicas não seguem um padrão para aferir seus trabalhos. O único trabalho que faz uma definição de um método de avaliação foi aquele proposto por Coffman e Weaver [12, 11]. No entanto, em sua avaliação não são considerados técnicas que não possuem acesso aos dados durante a interpretação das consultas.

Proposta de *Benchmark*

Visando avaliar técnicas de consultas por palavras-chave a banco de dados relacionais que não possuem acesso aos dados durante a interpretação das consultas, este trabalho propõe um *benchmark* com pontos que permitam identificar características positivas e negativas de cada sistema a ser avaliado.

O processo sistemático de avaliação de sistemas que recuperam informações em bancos de dados deve associar uma métrica quantitativa aos resultados produzidos pelo sistema em resposta ao conjunto de consultas [4].

De acordo com Manning et al.[34], para avaliar a eficácia de sistemas de recuperação de informação é necessário uma coleção de testes compostas por: 1) uma coleção de documentos, representado por bancos de dados; 2) um conjunto de informações de teste, representado como consultas; 3) Um conjunto de julgamentos de relevância, de forma padrão, uma avaliação binária de ambos, relevante ou não relevante, para cada par consulta-resultado.

Diferente da coleção de documentos de Recuperação de Informação (IR), para o presente trabalho, serão utilizados bancos de dados relacionais que organizam seus dados em relações, atributos e tuplas que devem ser combinadas para responder uma necessidade de informação.

Desta forma, este capítulo define e descreve de forma detalhada a estrutura do *benchmark* proposto conforme a coleção de testes definida por Manning et al.[34]. A Seção 4.1 apresenta os bancos de dados utilizados na avaliação, destacando as características de cada um e o motivo de sua escolha. A Seção 4.2 menciona as consultas que foram criadas. A Seção 4.4 descreve as métricas utilizadas para este método de avaliação. A Seção 4.5 faz uma comparação entre esta e outras abordagens relacionadas, conforme foram introduzidas no Capítulo 3. Por fim, a Seção 4.6 sintetiza o capítulo evidenciando pontos fundamentais da proposta.

4.1 Bancos de Dados

Os sistemas a serem avaliados por este trabalho realizam processos de conversão das consultas por palavra-chave em consultas SQL, sendo que cada palavra-chave é modelada para um nome de atributo, ou nome de relação ou um valor de um atributo. Nestes sistemas, a junção em bancos de dados é realizada por meio de relacionamentos entre chave primária e chave estrangeira ou por meio do relacionamento tabela-atributo-valor, ou seja, cada palavra-chave é associada a uma tabela, atributo ou um valor de um determinado atributo conforme seu tipo de dado [39].

A definição das bases de dados utilizadas no *benchmark* proposto seguiu alguns critérios de seleção de forma a atender características que consigam extrair o máximo de cada técnica a ser avaliada. Os pontos observados para a escolha dos bancos de dados foram: tamanho, quantidade de relações e estrutura do banco de dados. O tamanho é importante para avaliar como os sistemas se comportam com cargas maiores e menores de dados. Quantidade de relações e a estrutura do banco de dados são importantes pois esses sistemas durante o processo de interpretação da consulta fazem uma combinação das palavras-chaves com relações, atributos e dados do banco de dados, sendo que após essa combinação são mapeados os caminhos de junções possíveis, desta forma, o *benchmark* busca verificar o comportamento dos sistemas em bancos de dados com características distintas.

Sendo assim, para o presente trabalho, propõe-se a utilização dos bancos de dados Mondial¹, IMDB², DBLP³ e Northwind. A Tabela 4.1 fornece estatísticas sobre os quatro bancos de dados utilizados. Os esquemas dos bancos de dados diferem principalmente na quantidade de relações e em seu tamanho. Este é um ponto importante, pois estruturas diferentes destacam características distintas na avaliação.

Tabela 4.1: Características dos bancos de dados utilizados.

Banco de Dados	Tamanho (MB)	Quantidade Tuplas	Quantidade Relações	Quantidade Atributos	Média Atributos
Mondial	11	27.210	33	142	4,3
IMDB	429	5.518.540	7	21	3
DBLP	58	881.876	6	22	3,66
Northwind	1,1	3.308	13	89	6,84

As bases de dados selecionadas são melhores descritas a seguir:

- **IMDB**

Foi criado em 1990 por um grupo de fãs de filmes e programas de TV. Atualmente

¹<http://www.dbis.informatik.uni-goettingen.de/Mondial/>

²<http://www.imdb.com/>

³<http://dblp.uni-trier.de/>

contém um enorme banco de dados com mais de 185 milhões de itens de dados, que vão desde filmes a programas de TV, entretenimento, elenco e membros de equipes. O banco de dados utilizado para este *benchmark* é um subconjunto do original, ou seja, ele não contém todas as tuplas da versão que está em produção, mesmo assim, como é apresentado na Tabela 4.1, este é o banco de dados com maior tamanho. Caso seja necessário, com a evolução dos sistemas é possível ampliar o tamanho desse conjunto de dados para melhor atender futuras avaliações. A Figura 4.1 apresenta o diagrama do esquema do banco de dados IMDB.

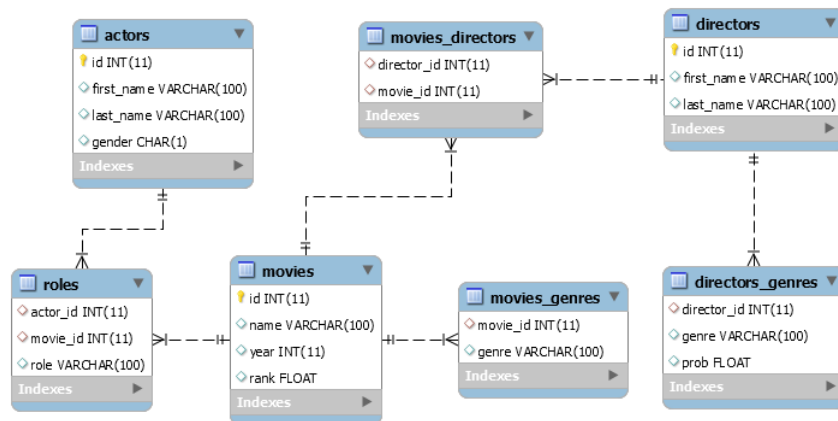


Figura 4.1: Esquema do banco de dados IMDB, adaptado de [2]

- **DBLP**

O banco de dados Digital Bibliography & Library Project (DBLP) é uma biblioteca aberta de informações bibliográficas de publicações na área de Ciência da Computação. O DBLP indexa mais de 3,3 milhões de trabalhos, publicados por mais de 1,7 milhões de autores. Esta base além de ser um dos bancos de dados mais utilizados nas avaliações realizadas pelos autores [8, 7], possui um contexto distinto dos outros bancos de dados selecionados, podendo assim contribuir com a avaliação. A Figura 4.2 apresenta o diagrama do esquema do banco de dados DBLP.

- **Northwind**

É um banco de dados sintético que contém dados de vendas para uma empresa fictícia chamada Northwind Traders, que importa e exporta alimentos especiais de todo o mundo. O esquema e cargas de trabalho foram obtidas em <https://relational.fit.cvut.cz/dataset/Northwind>. A escolha por essa base de dados deve-se ao fato de possuir uma quantidade de relações e atributos intermediária em relação aos demais bancos de dados e pela estrutura do esquema. A Figura 4.3 apresenta o diagrama do esquema do banco de dados Northwind.

- **Mondial**

É um banco de dados de informações obtidas de fontes geográficas de dados da Web. O esquema e cargas de trabalhos foram obtidas em <http://www.dbis.>

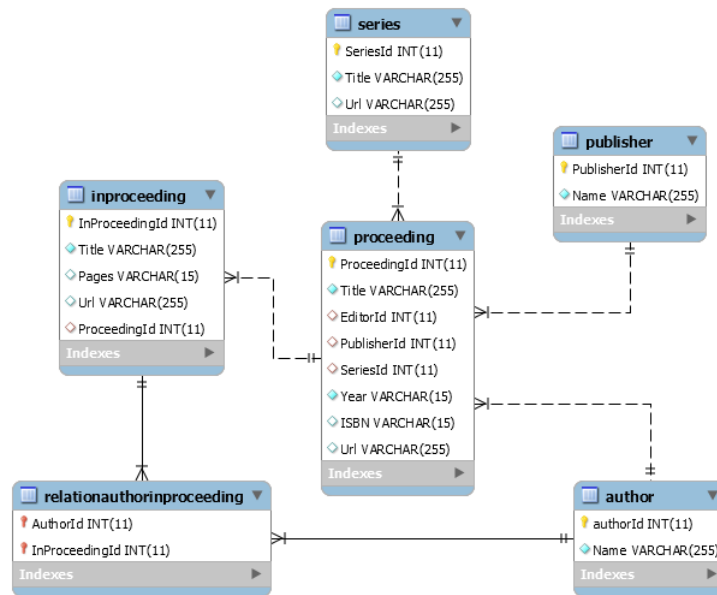


Figura 4.2: Esquema do banco de dados DBLP, adaptado de [2]

informatik.uni-goettingen.de/Mondial. Como pode ser observado na Tabela 4.1, o Mondial possui um tamanho médio, comparados com os outros bancos de dados selecionados, contudo, possui uma quantidade maior de relações e atributos, o que, para os sistemas a serem avaliados, o torna uma base de dados que exigirá mais das técnicas no momento do mapeamento das palavras-chave.

A Figura 4.4 apresenta o diagrama do esquema do banco de dados Mondial. O diagrama fornece informações como chaves primárias e estrangeiras assim como os atributos de cada relação, além dos relacionamentos entre as relações.

4.2 Consultas

Para avaliar a eficácia dos sistemas, foram definidas 50 consultas com as necessidades de informações para cada banco de dados. Esse número consegue abranger vários cenários distintos de avaliação. As consultas foram elaboradas por especialistas que possuem conhecimento das bases de dados. Não foram utilizadas consultas de usuários reais devido ao curto tempo do projeto.

Nos sistemas [39, 7], quando uma palavra-chave é mapeada para uma relação ou atributo do banco de dados, esta é classificada como termo de esquema, caso contrário, se a palavra-chave não possui um mapeamento de esquema, ela passa a ser relacionada a domínios de atributos e é considerada como termo de valor.

Desta forma, além da quantidade de palavras-chave de cada consulta, foi considerada para a definição das consultas a quantidade de termos de esquema e valor de cada uma das 50 consultas. A Tabela 4.2 fornece o intervalo do número de palavras-chave

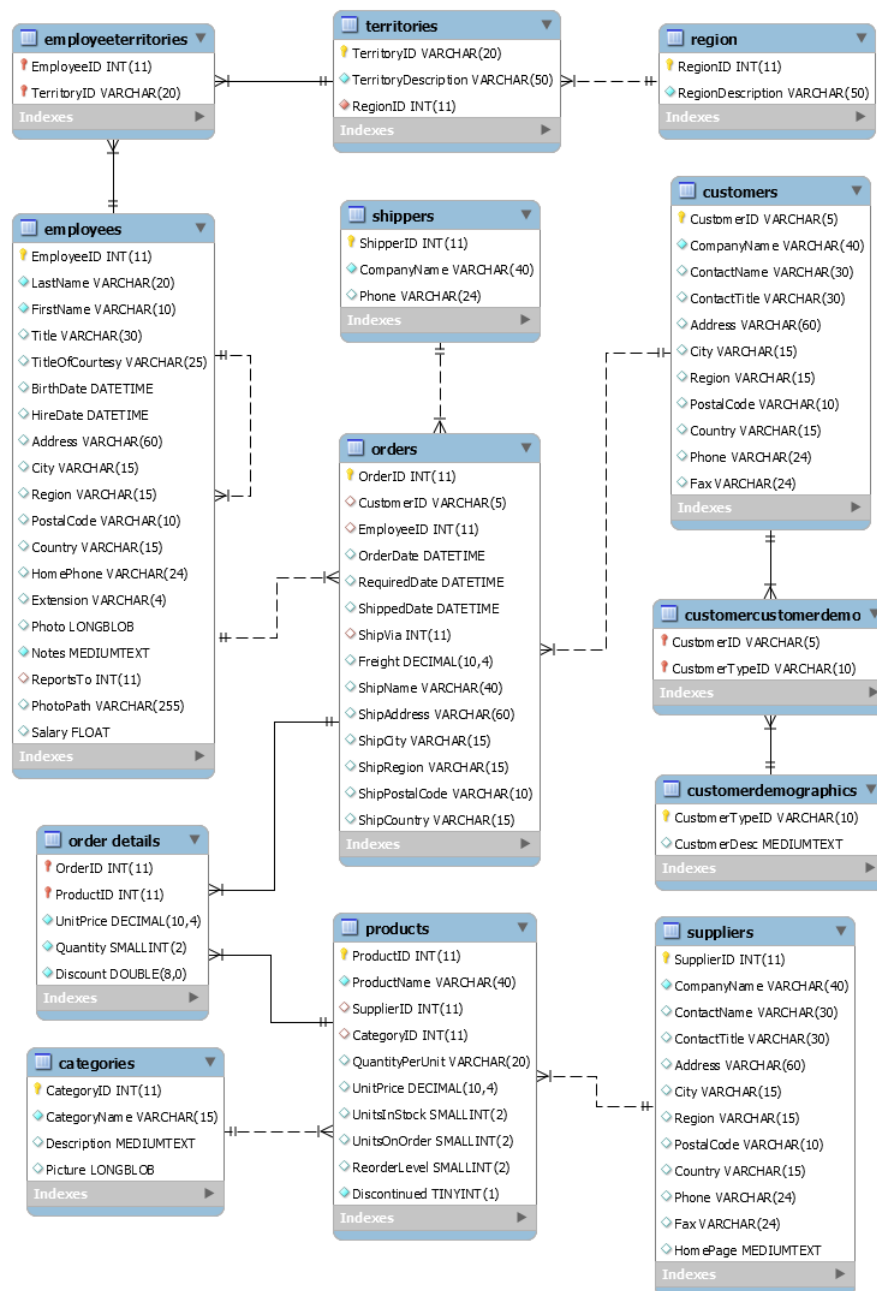


Figura 4.3: Esquema do banco de dados Northwind, adaptado de [2]

para cada conjunto de consultas, sendo que, para cada grupo foi definido uma distribuição uniforme de termos de esquema e de valor.

Para a construção das consultas, foram levados em consideração alguns pontos destacados por Ramada et al. [39]. Esses itens são importantes, pois representam melhor a real necessidade de usuários ao realizarem consultas a banco de dados relacionais.

Primeiramente, uma consulta pode conter uma palavra-chave que não desempenha um papel específico no esquema do banco de dados e esse termo não seria mapeado para uma estrutura correspondente. Considerando o diagrama do banco de dados apresen-

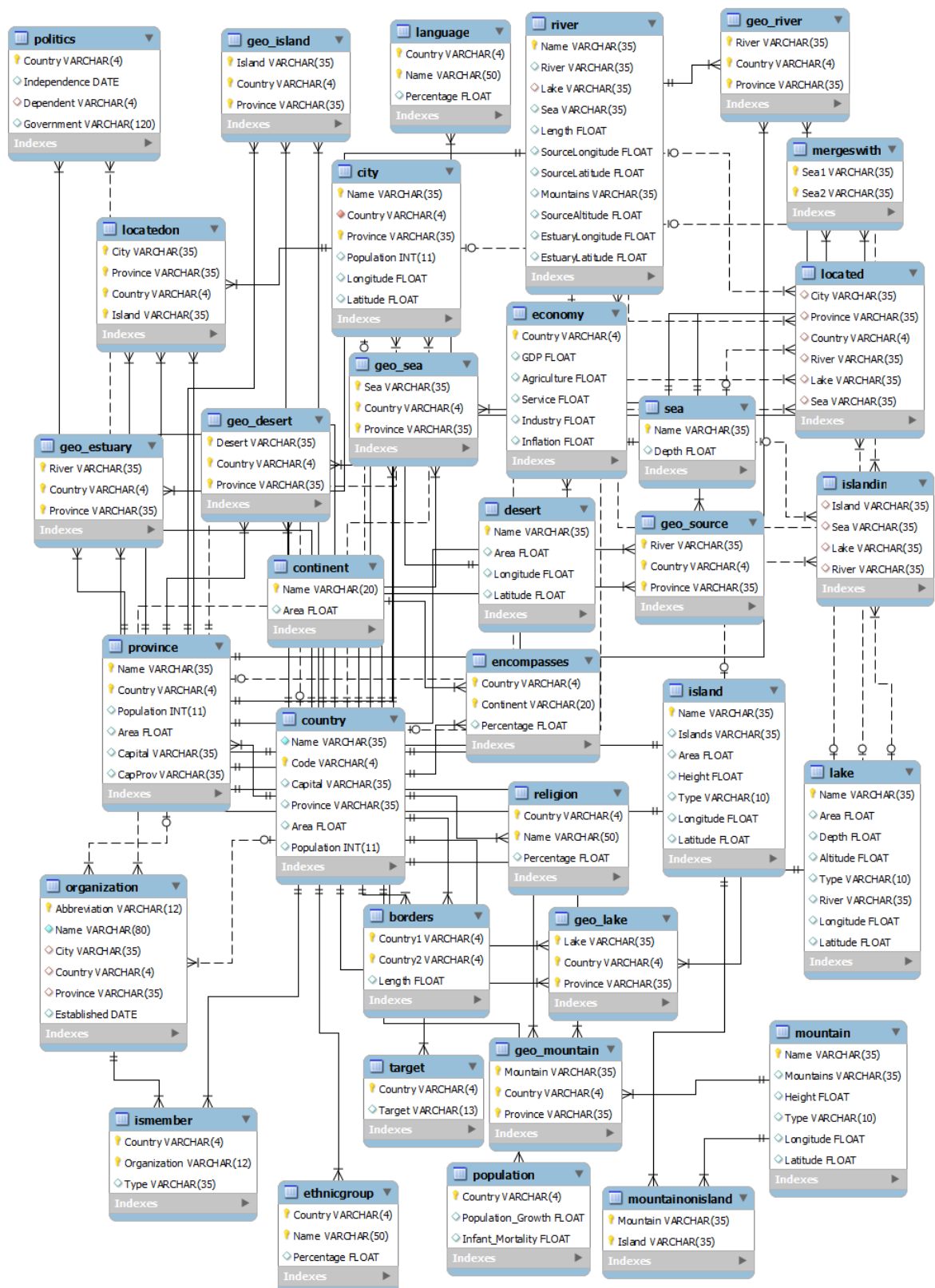


Figura 4.4: Esquema do banco de dados Mondial, adaptado de [1]

tado na Figura 4.1, na consulta *movie name highest rank* espera-se como resultado uma estatística, o nome do filme com maior classificação conforme avaliações feitas para esse

Tabela 4.2: *Estatísticas sobre as consultas*

Banco de Dados	Total Consultas	Intervalo número de Termos da consulta	Média número de termos da consulta
Mondial	50	1-7	3,24
IMDb	50	1-8	3,40
DBLP	50	1-8	3,44
Northwind	50	1-6	3,38
Total	200	1-8	3,36

filme.

O segundo ponto destaca que uma palavra-chave pode ter um significado em conjunto com outras palavras-chave. Por exemplo, na consulta *movie name “independence day”*, as palavras-chave entre aspas, “*independence day*”, devem ser mapeada conjuntamente. Por último, o significado de uma palavra-chave não é independente do significado das demais, desta forma, representam coletivamente o conceito pretendido.

Todas as consultas que foram definidas para este *benchmark*, com seus respectivos sentidos para cada banco de dados, estão disponíveis nas Tabelas [A.1](#), [A.2](#), [A.3](#) e [A.4](#) do Apêndice [A](#).

4.3 Avaliação de Relevância

A relevância do resultado de uma consulta é subjetiva e, desta forma, é sempre avaliada com relação à necessidade de informação do usuário. Como essa necessidade é aberta, várias consultas em SQL retornadas podem identificar um resultado que satisfaça a necessidade de informação, restando assim ao avaliador julgar cada um desses resultados.

Para esta proposta, foi definido uma avaliação de relevância binária para julgar os resultados de cada consulta. Esse formato de avaliação está em conformidade com o paradigma de Cranfield [10], da mesma forma que, está em conformidade com as avaliações feitas nos trabalhos descritos no Capítulo 3.

Buscando estabelecer padronização da avaliação, foi definido que, para cada palavra-chave da consulta, o avaliador deve verificar a correspondência correta das configurações geradas. Desta forma, para se definir que um resultado é relevante para uma consulta, a configuração gerada deve condizer com os metadados correspondentes para cada palavra-chave.

Para exemplificar, dada a consulta “*movies genres drama*”, que tem como objetivo obter a relação de filmes que possui gênero drama, as palavras-chave “*movies*” e “*genres*” deverão ser mapeadas como termos de esquema, visto que possuem correspondências com nomes de relações e/ou atributos, conforme o Banco de dados do IMDB da

Figura 4.1 e a palavra-chave “*drama*”, como não existe nenhuma relação ou atributo com esse nome deve ser definida como um termo de dado.

Na sequência, o sistema deve montar os mapeamentos conforme cada palavra-chave. Desta forma, um mapeamento correto para essa consulta seria mapear a palavra-chave “*movie*” com a relação “*movies*”, a palavra-chave “*genres*” com o atributo “*movies_genres.genre*”. Devido a proximidade da palavra-chave “*drama*” com “*genres*”, e sabendo que ela não possui nenhuma relação com nomes de esquema e atributo, “*drama*” deve ser relacionado ao atributo *movies_genres.genre* conforme é mostrado na Tabela 4.3.

Tabela 4.3: Mapeamentos considerados relevantes para consulta “*movies genres drama*”

Relevância	Mapeamentos
Relevante	movie → movies
	genres → movies_genres.genre
	drama → movies_genres.genre
Irrelevante	movie → movies
	genres → directors_genres.genre
	drama → directors_genres.genre
Irrelevante	movie → movies
	genres → movies_genres.genre
	drama → movies.name

4.4 Métricas

Medidas para avaliar sistemas que recuperam informações devem quantificar se o conjunto de informações recuperadas é similar ao conjunto de informações consideradas relevantes [5, 34]. Sendo a relevância algo subjetivo, esta avaliação torna-se difícil. Dessa maneira, as métricas utilizadas são uma forma de quantificar o processo.

Duas métricas muito utilizadas em avaliação de sistemas que recuperam informação são Precisão e Revocação [34]. Elas avaliam um conjunto de resultados baseadas na suposição de relevância binária, ou seja, se cada resultado é relevante ou não para a consulta.

Os sistemas que o presente trabalho se propõe a avaliar utilizam critérios de ordenação por relevância devido ao fato de que durante o processo de tradução da consulta, diversas interpretações poderão surgir, juntamente com aquelas que podem não condizer com a intenção do usuário. Desta forma, para medir a relevância dos n primeiros resultados de uma lista ordenada é sugerido a métrica de Precisão em $n(P@n)$.

Desta forma, buscando avaliar o quão relevante o resultado é para um grupo de consultas, as métricas descritas a seguir foram selecionadas para serem utilizadas por este *benchmark*:

- **Número de top-1 resultados relevantes**

Avalia o número de consultas para as quais o primeiro resultado é relevante.

- **Precisão em n ($P@n$)**

Esta métrica mede a relevância dos n primeiros resultados retornados de uma lista ordenada. Utiliza-se da seguinte formula para cálculo da relevância.

$$P@n = \frac{r}{n} \quad (4-1)$$

onde, n é o número de resultados retornados e r é o número de resultados considerados relevantes e retornados até a posição n da lista ordenada.

- **Mean Reciprocal Rank (MRR)**

Mean Reciprocal Rank é uma medida estatística que avalia listas de respostas produzidas por um processo de consulta onde tais resultados estão ordenados por uma probabilidade de relevância, ou seja, esta métrica calcula a média dos melhores resultados relevantes classificados para cada consulta. O *reciprocal rank* tem por objetivo avaliar a posição do primeiro resultado relevante em relação ao topo do conjunto de resultados. O valor de MRR é determinado da seguinte maneira.

$$MRR = \frac{\sum_{i=1}^N \frac{1}{p_i}}{N} \quad (4-2)$$

em que, N é o número de consultas e p_i é a posição onde o resultado relevante, da consulta i , é encontrado na lista.

- **Average Precision (AP)**

Esta métrica é calculada com base na precisão em cada resultado relevante na posição, desta forma, é a média dos valores de precisão $n(P@n)$. Essa medida é útil para calcular um único valor de precisão ao comparar diferentes algoritmos de recuperação em uma consulta c .

$$AP = \frac{\sum_{n=1}^N P@n \times Rel(n)}{R_q} \quad (4-3)$$

em que, R_q é o número total de resultados considerados relevantes para a consulta, N é o número de resultados recuperados, e $rel(n)$ é uma função binária sobre a

relevância do n^{th} resultado.

$$Rel(n) = \begin{cases} 1 & , \text{se o } n^{th} \text{ é relevante,} \\ 0 & , \text{caso contrário.} \end{cases} \quad (4-4)$$

- **Mean Average Precision (MAP)**

Esta é a média das pontuações de precisão média para cada consulta.

$$MAP = \frac{\sum_{q=1}^Q AP(q)}{Q} \quad (4-5)$$

4.5 Comparação com Outra Abordagem

Nas seções anteriores foram apresentados os componentes da estrutura do *benchmark* proposto por este trabalho, detalhando os bancos de dados utilizados, conjuntos de consultas definidos, forma de julgamento de relevância e métricas utilizadas. Por outro lado, visando verificar possíveis contribuições ao estado da arte, esta seção tem por objetivo comparar o *benchmark* proposto por Coffman e Weaver [12, 11] com a proposta por este trabalho.

Tabela 4.4: Comparação do método proposto com outra abordagem

	<i>Benchmark</i> Coffman e Weaver [12]	<i>Benchmark</i> Proposto
Banco de Dados	Mondial IMDB Wikipedia	Mondial IMDB DBLP Northwind
Consultas	50 Consultas	50 Consultas (Distribuição proporcional dos termos de valor e termos de dado)
Avaliação Relevância	Relevância Binária	Relevância Binária Verificação das configurações
Métricas	Top-1 resultados relevantes <i>Reciprocal Rank</i> <i>Average Precision</i> <i>Mean Average Precision</i> Tempo de Execução Tempo de Resposta	Top-1 resultados relevantes <i>Precisão em $n(P@n)$</i> <i>Reciprocal Rank</i> <i>Mean Reciprocal Rank</i> <i>Average Precision</i> <i>Mean Average Precision</i>

A Tabela 4.4 apresenta uma comparação entre as duas propostas. Cada uma das linhas destaca a diferença entre as duas propostas em relação aos bancos de dados utilizados, consultas definidas, avaliação de relevância e métricas utilizadas.

Com relação aos bancos de dados, é possível observar que Coffman e Weaver [12] propõem a utilização de três: Mondial, IMDB e Wikipedia. O banco de dados do Wikipedia foi cogitado para utilização no presente trabalho, porém foi descartado visto que seus metadados não expressam assunto específico com relação ao domínio do banco de dados, dificultando assim a obtenção dos dados nos sistemas a serem avaliados. De antemão, verifica-se que os atuais sistemas não conseguem retornar resultados relevantes para banco de dados com essa estrutura. No banco de dados Wikipedia, os nomes das relações com seus respectivos atributos não representam o sentido da informação ao qual serão adicionados, ou seja, utiliza relações com nomes genéricos e essas armazenam tuplas de vários assuntos distintos. O esquema do Wikipedia pode ser visto na Figura 4.5.

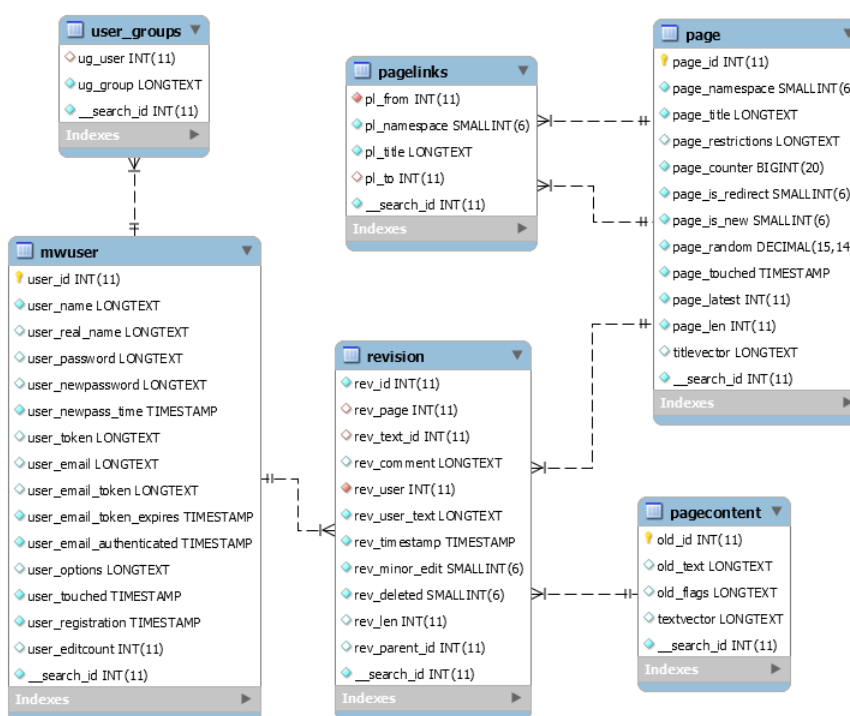


Figura 4.5: Esquema do banco de dados Wikipedia proposto por Coffman e Weaver[12]

Como visto no esquema do Wikipedia apresentado na Figura 4.5, as relações “page”, “revision” e “pagecontent” armazenam textos dos mais diversos assuntos tornando assim difícil de um sistema de busca por palavras-chave, que possui acesso somente a metadados do banco de dados, identificar que um dado termo da consulta pode ser relacionado a estes tipos de relações.

Com relação às consultas, Coffman e Weaver [12] utilizou um conjunto de 50 consultas, para cada banco de dados, em sua proposta de avaliação. O presente trabalho manteve a mesma quantidade, contudo, o conjunto de 50 consultas foi refeito buscando melhor atender características presentes nos sistemas de consultas por palavras-chave. A principal diferença entre as consulta esta na quantidade média termos da consulta, neste trabalho ouve uma preocupação em definir uma quantidade equivalente entre termos de valor e termos de esquema para que assim fosse possível destacar as diferenças de resultado em cada combinação.

O presente trabalho ficou definido que a avaliação de relevância seria binária onde seriam avaliados os mapeamentos de configuração gerados por cada técnica. Neste sentido, conforme descrito na Seção 4.3, deve-se verificar se cada palavra-chave da consulta condiz com o metadado correspondente no esquema do banco de dados. Nos trabalho de Coffman e Weaver [12, 11] são apresentados que a forma de avaliação de relevância é binária porem, os autores não detalham como a avaliação é realizada.

Coffman e Weaver [12] definem quatro métricas para avaliar eficácia de técnicas de consultas por palavras-chave a bancos de dados relacionais: Número de top-1 resultados relevantes, *Reciprocal Rank*, *Average Precision*, *Mean Average Precision (MAP)*. Para avaliar a eficiência dos sistemas em retornar resultados definiram: tempo de execução e tempo de resposta [11]. Visto que as métricas propostas nesse trabalho atende ao seu objetivo, decidiu-se mantê-las no entanto as métricas para avaliar a eficiência (tempo de execução e tempo de resposta) não foram utilizadas neste projeto por falta de tempo para análise das mesmas, porém esta será adicionada como trabalho futuro.

4.6 Considerações Finais

Este capítulo expôs, de forma detalhada, a estrutura proposta do *benchmark* para avaliação de sistemas que fazem consultas por palavras-chave a bancos de dados relacionais. Este método define uma forma padronizada de avaliação com cargas de trabalhos condizentes com o mundo real.

Foram definidas 50 consultas para cada banco de dados proposto para o método de forma a abranger vários cenários distintos de avaliação. As consultas foram criadas de forma a possuírem uma distribuição proporcional de termos de valor e termos de dados. Buscando quantificar o processo de avaliação dos sistemas, o capítulo descreve as métricas que foram selecionadas para este *benchmark*.

Ao final, este capítulo apresenta um comparativo entre o *benchmark* proposto com a abordagem apresentada por outros trabalhos.

O *benchmark* proposto por este trabalho atende aos quatro critérios definidos por Gray [20]. É relevante pois ao utiliza-lo será possível medir ao máximo os sistemas, visto

que, o conjunto dos quatro banco de dados possui características distintas, e juntamente com o conjunto de consultas, que abrangem vários cenários, proporcionam uma avaliação que extraia pontos distintos do sistema.

Neste trabalho, o *benchmark* foi utilizado para avaliar dois sistemas de consultas por palavras-chave a bancos de dados relacionais para técnicas que não possui acesso aos dados durante a interpretação da consulta, no entanto, esse *benchmark* pode ser utilizado também para outras técnicas de consultas por palavras-chave, sendo assim portátil.

Com base nos bancos de dados utilizados, é possível observar que o *benchmark* proposto é escalável pois a medida que os sistemas forem evoluindo é possível aumentar o tamanho dos bancos de dados. O IMDB, por exemplo, utilizado possui apenas uma pequena porção das tuplas da base de dados total.

Por fim, o *benchmark* é simples visto que apresenta de forma detalhada todas as consultas com suas respectivas sentidos bem definidos e uma forma clara de avaliação de relevância tornando assim o método compreensível aos usuários que o utilizar.

Resultados da Avaliação

Este capítulo detalha o processo de avaliação e os resultados obtidos a partir da aplicação do *benchmark* para análise dos sistemas propostos por Bergamaschi et al. [7] e Ramada et al. [39]. A Seção 5.1 apresenta as etapas que foram seguidas para execução da avaliação. A Seção 5.2 descreve a implementação das técnicas e configurações do ambiente de teste. A Seção 5.3 detalha os resultados obtidos da avaliação. Por fim, a Seção 5.5 faz uma discussão sobre os resultados da avaliação, sintetizando o capítulo dando destaque pontos importantes sobre cada técnica avaliada

5.1 Etapas da Avaliação

Assim como visto no capítulo 2, um *benchmark* é um método padronizado e bem definido para comparar produtos e objetos. Sendo assim, visando a obtenção de dados corretos sobre as diferentes técnicas, afim de compará-las de maneira equivalente, foi definido um processo sistemático de execução da avaliação conforme ilustrado na Figura 5.1.

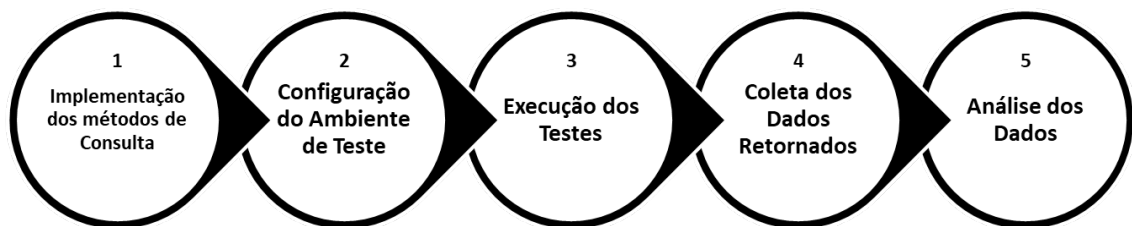


Figura 5.1: *Etapas de Avaliação do Benchmark*

A primeira etapa, implementação dos métodos de consulta, consiste na seleção e implementação das técnicas que serão avaliadas. Na segunda etapa, é realizada a configuração do ambiente de testes. Nesse momento é definida a máquina onde serão executados os sistemas, instalação do Sistema Gerenciador de Banco de Dados (SGBD) e importação os dados dos bancos de dados propostos para a avaliação.

A terceira e quarta etapas são executadas simultaneamente. Nelas, cada uma das cinquenta consultas, de cada banco de dados definido, são executadas nos sistemas selecionados e o retorno de cada sistema é devidamente anotado em planilhas para posterior análise.

Tendo em vista que uma única consulta por palavra-chave pode ter várias interpretações, a quarta etapa desempenha um papel importante no processo de avaliação dos métodos. Desta forma, nessa etapa é definida a relevância de cada resultado retornado por meio das configurações geradas por cada sistema conforme o contexto de cada banco de dados.

A quinta etapa consiste em quantificar os resultados conforme as métricas definidas e demonstrar graficamente as diferenças encontradas em cada sistema avaliado, para assim aferir a viabilidade de utilização das técnicas em um ambiente real.

5.2 Implementação e Configuração do Experimento

A implementação da técnica de consulta por palavras-chave proposta por Bergamaschi et al. [7] foi disponibilizada pelos autores em [15]. Esta ferramenta foi desenvolvida na linguagem de programação *Java*. Essa aplicação apresenta uma *interface* onde é possível informar uma consulta por palavras-chave e como retorno o sistema apresenta as configurações geradas e resultados obtidos por cada configuração. Conforme o sistema, vale destacar que o resultado da execução de um comando SQL gerado para alguma configuração específica só é executado assim que o usuário informa que deseja visualizar os resultados desta configuração.

A implementação da técnica de consultas por palavras-chaves proposta por Ramada et al. [39] foi obtida com seus respectivos autores. O código fonte disponibilizado foi implementado na linguagem de programação *Java*. Em sua versão original, após a execução, o sistema traz um arquivo contendo todas as configurações seguido da execução do comando SQL para cada configuração.

Visando uma avaliação uniforme, o código fonte da técnica proposta por Ramada et al. [39] foi ajustado para que este trouxesse todas as configurações antes da execução dos respectivos comandos SQL, e assim, ter um tempo exato de retorno das configurações para cada consulta conforme é feito na implementação de Bergamaschi et al. [7].

Para realização do experimento, os programas das técnicas avaliadas foram executados em uma máquina virtual *Java* versão 1.8.0 sobre o sistema operacional Microsoft Windows 10 de 64 bits em um computador com processador *Intel Core i7* de 1,8 GHz e 8 GB de memória RAM.

O Sistema Gerenciador de Banco de Dados(SGBD) utilizado para a avaliação foi o MySQL versão 10.1.13. Este foi escolhido após verificar que os códigos fontes dos

dois sistemas, desenvolvido e disponibilizado pelos seus respectivos autores, realizavam conexão com esse SGBD.

5.3 Resultados

Ao final da execução do *benchmark*, foi possível, por meio de uma análise sistemática, obter informações do desempenho das técnicas avaliadas. A Tabela 5.1 apresenta uma síntese das consultas que foram executadas com sucesso para cada banco de dados, por cada técnica de consulta. Vale destacar o número de consultas que não foram concluídas com sucesso. Foi definido 60 minutos como tempo máximo de espera para execução da consulta, sendo assim, após esse tempo caso a consulta não tenha sido finalizada a execução do sistema é encerrada e a consulta é marcada como exceção do tempo limite.

Tabela 5.1: *Síntese da quantidade de consultas finalizadas de um total de 50 consultas executadas*

Sistema	DBLP			IMDB			Mondial			Northwind		
	✓	✗	Rel	✓	✗	Rel	✓	✗	Rel	✓	✗	Rel
Keymantic	50	0	40	50	0	20	50	0	27	50	0	20
Ramada et al.	41	9	29	46	4	22	20	30	11	37	13	19

Legenda:

✓ - Total de consultas finalizadas com sucesso

✗ - Total de consultas não finalizadas

Rel - Total de consultas que obtiveram resultados relevantes

Na avaliação do sistema de Ramada et al., houve consultas em todos os bancos de dados que não finalizaram a execução. Para o banco de dados Mondial, o sistema não finaliza a execução de 30 das 50 consultas. Isso se dá devido à quantidade de combinações possíveis de acordo com o número de relações no esquema do banco de dados Mondial. Durante a execução dos testes foi observado que o código fonte entrava em um laço de repetição infinito para algumas consultas, o que levava a não finalizar a execução. Desta forma, problemas na implementação afetou os resultados da avaliação para essa técnica.

Em contrapartida, o sistema Keymantic finaliza todas as 50 consultas para cada banco de dados, porém um número expressivo de consultas não trazem resultados relevantes. Vale destacar que, conforme valores de resultados relevantes da Tabela 5.1, Keymantic obtém apenas 40%, 54% e 40% de resultados relevantes para os bancos de dados IMDB, Mondial e Northwind respectivamente.

Com base na avaliação, é possível observar que, apesar da grande quantidade de consultas não finalizadas por Ramada et al., esta técnica possui um percentual alto de resultados relevantes retornados para resultados finalizados. Em contraponto, Keymantic

finaliza a execução todas as consultas porém grande parte dos resultados não são relevantes.

Os resultado inicial indica que as técnicas avaliadas possuem um desempenho razoável para bancos de dados com esquemas menores, ou seja, uma quantidade menor de relações. A medida que a quantidade de relações aumenta impacta diretamente a eficácia das consultas. Nas sub seções seguintes serão examinados outros resultados obtidos conforme as métricas definidas pelo *benchmark* para avaliar a eficácia das técnicas.

5.3.1 Quantidade de Configurações

A Figura 5.2, apresenta a média de configurações geradas para cada conjunto de consultas nos bancos de dados utilizados na avaliação.

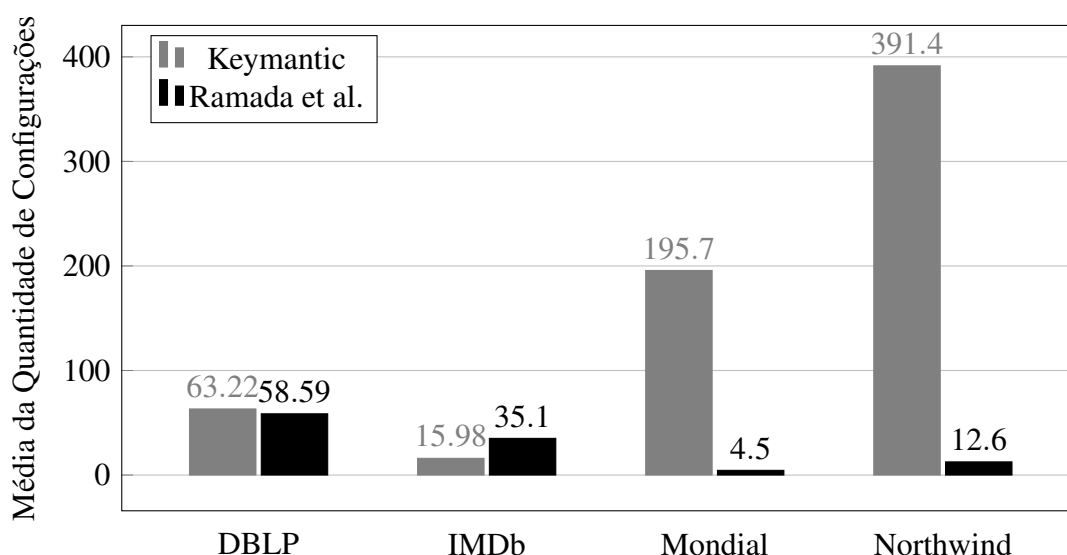


Figura 5.2: Quantidade média das configurações geradas

Com relação à quantidade de configurações geradas para os bancos de dados DBLP e IMDB, os dois sistemas obtiveram uma quantidade média próxima e baixa em relação as outras duas bases de dados. Isso foi influenciado pela pequena quantidade de relações do esquema dos dois bancos de dados.

Na base de dados do Mondial, o sistema da Ramada et al obteve uma média de configurações baixa pois, assim como visto na tabela 5.1, grande parte das consultas não foram finalizadas pelo sistema durante um tempo máximo de 60 minutos. Conforme definido nesta avaliação, às consultas que não retornaram resultados foi atribuído o valor zero(0) para cálculo desta métrica.

Com relação ao banco de dados Northwind, o sistema de Ramada et al. obteve uma média de 12.6 configurações devido à grande quantidade de consultas que não foram finalizadas no tempo máximo. Em contrapartida, o sistema Keymantic obteve a média

391.4 configurações. Essa discrepância ocorreu devido às consultas 44 e 48 da Tabela A.4 do Apêndice A que obtiveram, respectivamente, 7373 e 8999 configurações cada uma.

5.3.2 TOP-1 Resultados

Por meio desta métrica é possível destacar a capacidade das técnicas de retornarem um resultado relevante na primeira posição do conjunto de resultados. A Figura 5.3, apresenta o percentual de consultas que tiveram resultados relevantes classificados na primeira posição da lista de resultados retornados pelo sistema. Esse percentual possui variação de 0 a 100, e desta forma, percentuais maiores significam que dentre o conjunto de 50 consultas um maior número delas retornou o valor relevante na primeira posição.

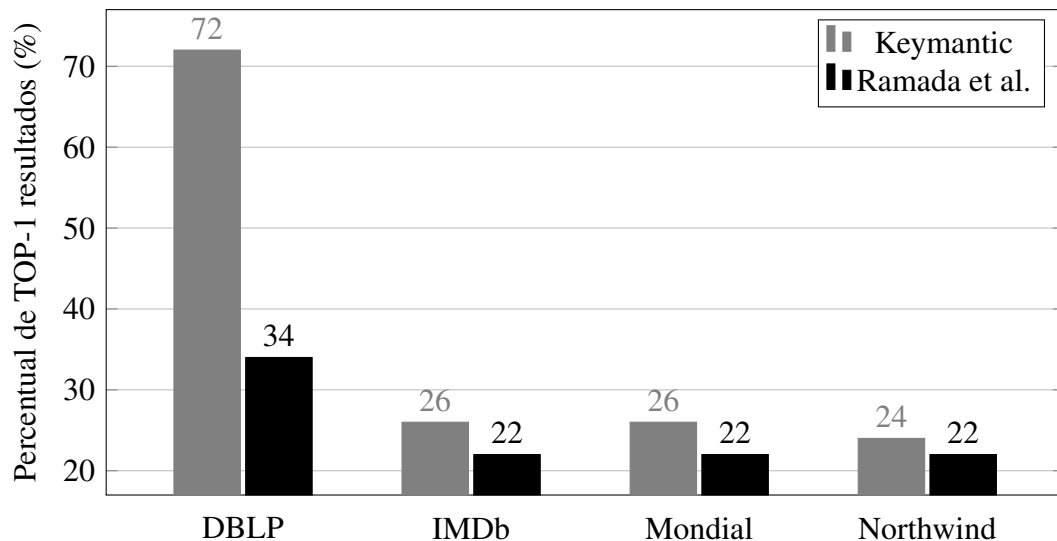


Figura 5.3: *Percentual de Top-1 Resultados relevantes (%)*

Dentre as bases de dados testadas, IMDb, Mondial e Northwind obtiveram resultados bem próximos para as duas técnicas avaliadas. No DBLP, o sistema Keymantic obteve melhor resultado, retornando resultados relevantes na primeira posição em 72% do total das 50 consultas. Uma possível justificativa desse resultado está no esquema do banco de dados DBLP que possui uma quantidade menor de relações onde, em sua estrutura, a tabela “*proceedings*” possui relacionamentos de chave primária-estrangeira com quase todas as outras relações.

5.3.3 Mean Reciprocal Rank (MRR)

Para a avaliação da métrica *Mean Reciprocal Rank (MRR)* é necessário primeiramente medir o Reciprocal Rank do resultado de cada uma das 50 consultas para cada

banco de dados. Desta forma, nesta etapa da avaliação foi verificada a posição do primeiro resultado relevante e assim realizado o cálculo da métrica.

As Tabelas 5.2 e 5.3 correspondem aos valores de *Reciprocal Rank* obtidos para cada consulta em cada banco de dados respectivamente, das técnicas de Ramada et al e Keymantic. Nestas tabelas, para cada consulta numerada na coluna com título “Nº”, são apresentados os valores correspondentes de MRR para os quatro bancos de dados, sendo que, esse valor representa a posição do primeiro resultado relevante em relação ao topo do conjunto de resultados.

Os resultados do *Reciprocal Rank* variam de 0 a 1, sendo que, quanto mais próximo de 1, significa que o primeiro resultado relevante para a consulta estaria mais próximo da primeira colocação da lista de respostas. Caso contrário, se o valor estiver mais próximo de 0, o resultado relevante estará mais distante do topo da lista ou para aquela consulta a técnica não conseguiu retornar um resultado relevante.

Tabela 5.2: *Reciprocal Rank de cada consulta da Técnica Ramada et al [39] (Nº = número da consulta, D = DBLP, I = IMDB, M = Mondial e N = Northwind)*

Nº	D	I	M	N	Nº	D	I	M	N
1	0	0	0	0	26	1	0,25	0	0,111
2	0	0	0	0	27	1	0	0	0
3	0	0	0	0	28	0,5	0,5	0	0,111
4	1	0	0	0	29	0	0	0	0
5	1	0	1	0	30	1	0	0	0
6	1	1	1	0	31	0	1	0	0
7	0	1	1	0	32	0	1	0	0
8	1	1	0	0	33	0,25	0	0	0
9	1	0,5	1	0	34	0	0	0	0
10	1	1	1	0	35	1	0	0	0,142
11	1	0	1	0	36	1	0	0	0
12	0,5	1	0	0,142	37	1	0,058	0	1
13	1	0	1	0	38	1	0,025	0	0
14	1	0	0	1	39	0	0	0	0
15	0,5	0,5	1	1	40	0,021	0,333	0	0
16	1	0,333	x	1	41	0	0	0	0
17	0	0	0	0	42	0,125	0,333	0	0
18	0,333	0	1	0	43	0	0	0	1
19	0	0	1	0	44	0,333	0	0	0
20	0,076	0	0	1	45	0,083	0	0	0
21	0,066	1	0	0,2	46	0	0,066	0	0
22	0	1	0	0	47	0	0	0	0
23	0,25	0	0	0,25	48	0	0,1	0	0
24	0	1	0	0,111	49	0	0	0	0
25	0	0,5	0	1	50	0	0	0	0

Tabela 5.3: *Reciprocal Rank de cada consulta da Técnica Keyman-
tic [7] (Nº = número da consulta, D = DBLP, I =
IMDB, M = Mondial e N = Northwind)*

Nº	D	I	M	N	Nº	D	I	M	N
1	0,25	0	0,125	0,032	26	1	0	0	0,25
2	0	0	0,166	0,032	27	1	0	1	0
3	0	0,125	0,041	0,018	28	1	0	1	0
4	1	0,125	0,041	1	29	1	0	1	0
5	1	0,5	1	1	30	1	1	0	0
6	1	1	1	1	31	0	1	0	0
7	0	1	1	1	32	1	1	0	0
8	1	1	1	1	33	1	0	0	0
9	1	0,5	1	0	34	1	0	1	0
10	1	1	1	0	35	0	0	0,5	0
11	1	0	1	0	36	1	0	0	0,5
12	1	0	1	0	37	0,1	0	0,5	1
13	1	1	1	0	38	1	0	0,5	0
14	0,5	0	0	1	39	1	1	0	0
15	1	0	0	1	40	1	0,5	0,2	0
16	1	0	0	1	41	0	0	0,5	0
17	0	0	0	1	42	1	0,5	0,5	0
18	1	0	0,25	0,1	43	1	0	0,5	1
19	1	0	0	0	44	1	0	0,5	0
20	0	0	0	0	45	1	0	0	0,076
21	0	1	0	1	46	1	1	0	0
22	1	0,5	0	0	47	1	0	0	0
23	1	0	0	0,5	48	1	1	1	0
24	0	0	0	0	49	1	1	0	0
25	1	0	0	0	50	1	0	0	0

Tendo por base os valores de *Reciprocal Rank*, é possível obter as médias dos melhores resultados relevantes para cada banco de dados em cada técnica avaliada. Esta média é apresentada na Figura 5.4.

Keymantic retornou resultados mais relevantes no banco de dados DBLP com MRR igual a 0.76 em comparação com Ramada et al. com MRR de 0.4. Sendo assim, para o banco de dados DBLP, Keymantic obteve uma número maior de resultados relevantes retornados na primeira posição ou próximo da primeira posição da lista de resultados da consulta. Para os outros bancos de dados, Keymantic também obteve um MRR maior que a técnica proposta por Ramada et al., todavia a diferença entre eles foi pequena.

5.3.4 Precisão em $n(P@n)$

Uma das métricas mais utilizadas para avaliar sistemas de recuperação de informações, a Precisão (*Precision*) verifica a fração de resultados considerados relevantes,

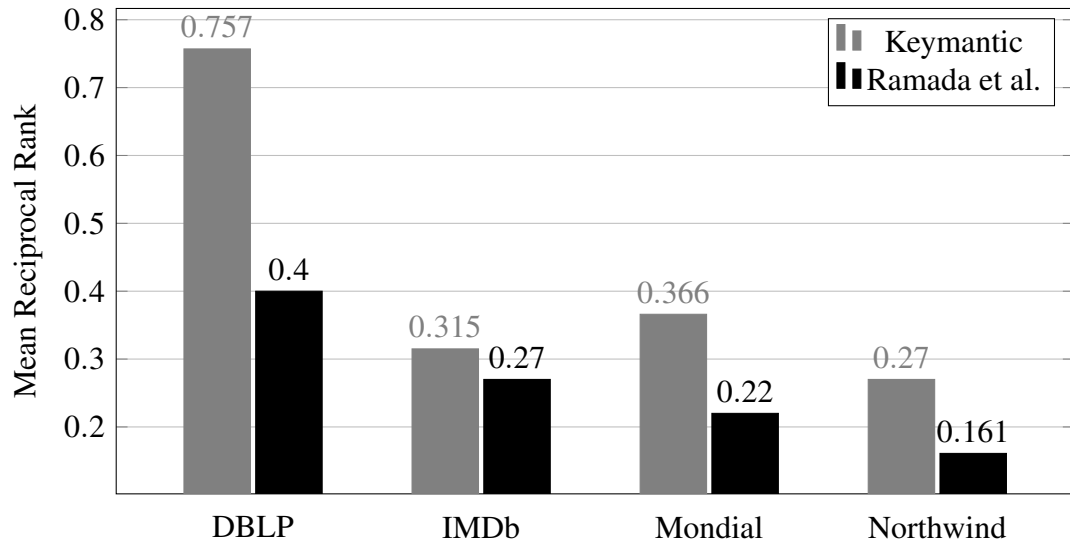


Figura 5.4: Mean Reciprocal Rank

dentro de um conjunto de resultados retornados. Os valores desta métrica podem variar de 0 a 1, sendo que valores mais próximos de 1 expressam maior precisão dos resultados retornados. De forma similar, quanto mais próximo de zero for o valor, menos resultados relevantes foram retornados.

Conforme explicado na Seção 4.4, a Precisão em n ($P@n$) verifica a precisão para os n primeiros resultados retornados. Neste trabalho, foram verificadas as precisões: $P@1$, $P@2$, $P@3$, $P@4$, $P@5$, $P@6$, $P@7$, $P@8$, $P@9$ e $P@10$. Os resultados de cada uma delas pode ser conferido no Apêndice B. No corpo deste trabalho são apresentados os gráficos para as métricas $P@5$ e $P@10$, respectivamente as figuras 5.5 e 5.6.

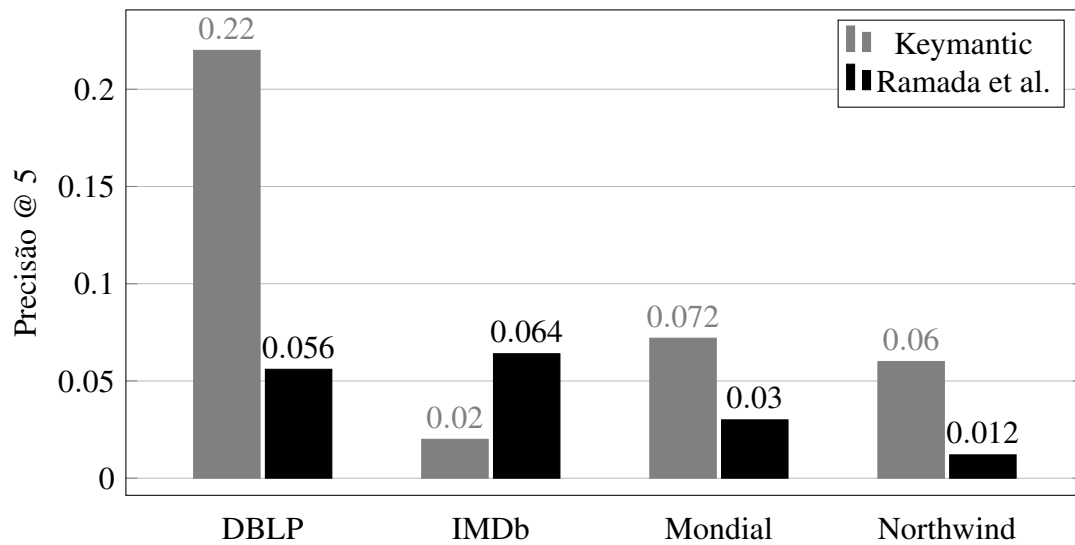


Figura 5.5: Precisão @ 5

A Figura 5.5 apresenta os valores de precisão para os 5 (cinco) primeiros resultados retornados na avaliação dos dois sistemas. Ao utilizar o banco de dados

DBLP, o valor retornado para o sistema Keymantic foi de 0.22, sendo este melhor que o sistema de Ramada et. al, cujo valor retornado foi de 0.056. Essa distorção ocorre porque Keymantic finaliza a execução todas as consultas para o DBLP e a maioria dos resultados relevantes estão entre as cinco primeiras da lista de resultados. Outro ponto a ser destacado neste resultado refere-se a estrutura do banco DBLP, que das 6 relações do banco de dados, a tabela “*Proceeding*” possui relacionamento de chave-estrangeira com 4 das 5 demais, o que facilita a definição dos caminhos de junção da consulta.

Para o banco de dados IMDb, foram retornados os valores 0.02 e 0.064, respectivamente Keymantic e Ramada et. al, tendo este último melhor resultado. No entanto, o sistema Keymantic apresentou valores mais altos, ou seja, melhores resultados quando testados com os bancos de dados Mondial (0.072) e NorthWind (0.06), em contraponto com o sistema de Ramada et. al, que obteve os valores 0.03 (Mondial) e 0.012 (IMDb). Nessas duas últimas bases de dados a diferença entre os resultados deve-se a quantidade de consultas não finalizadas com sucesso pelo sistema de Ramada et al.

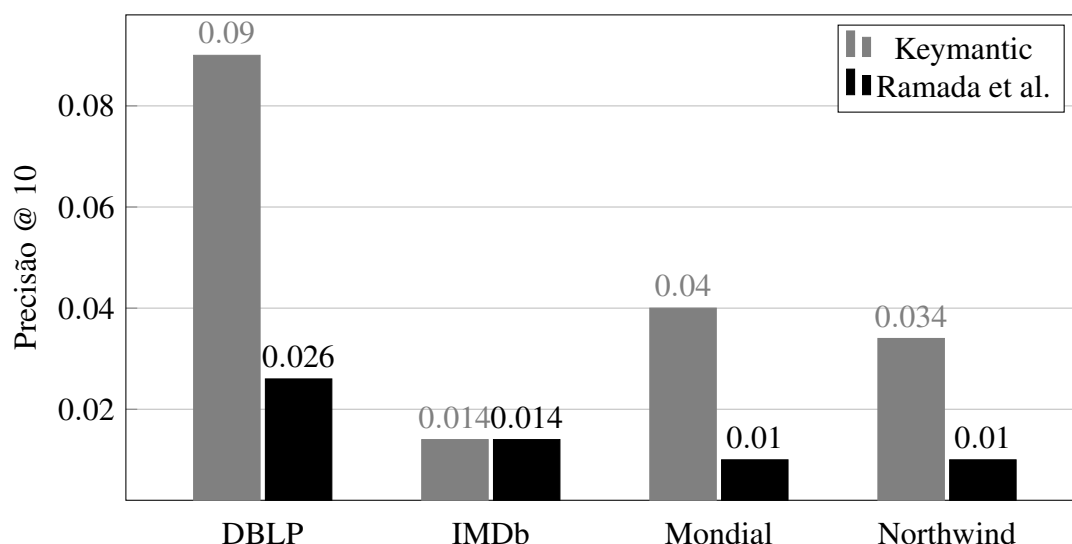


Figura 5.6: *Precisão @ 10*

O gráfico da Figura 5.6 exibe os valores de precisão para os 10 (dez) primeiros resultados retornados na avaliação ($P@10$). No banco de dados DBLP, o sistema Keymantic obteve valores mais altos em relação do sistema desenvolvido por Ramada et. al., sendo esses valores, respectivamente, 0.09 e 0.026. Ao realizar os testes com o banco de dados IMDb, foram retornados os mesmos valores de precisão para os dois sistemas: 0.014. Os valores retornados utilizando o bancos de dados Mondial foi de 0.04 e 0.01, respectivamente, para os sistemas Keymantic e Ramada et al. Os testes efetuados com o banco de dados NorthWind apresentaram valores próximos aos do banco de dados anterior, sendo 0.034 para o sistema Keymantic e 0.01 para o sistema de Ramada et al.

5.3.5 Average Precision (AP)

Uma forma de calcular a média das precisões em cada resultado na sua respectiva posição é através da métrica *Average Precision (AP)*. Desta forma, para cada consulta, após a obtenção das precisões $P@n$ dos dez primeiros resultados retornados, foi realizado o cálculo dessa métrica conforme a formula apresentada na Seção 4.4.

As Tabelas B.1, B.2, B.3, B.4, B.5, B.6, B.7 e B.8 do Apêndice B possuem a última coluna com os valores de *average precision* de cada uma das 50 consultas. Conforme os resultados obtidos pela métrica, foi possível gerar o gráfico das Figuras 5.7 e 5.8, em que são apresentados as variações de dispersão dos resultados conforme as 50 consultas para cada banco de dados.

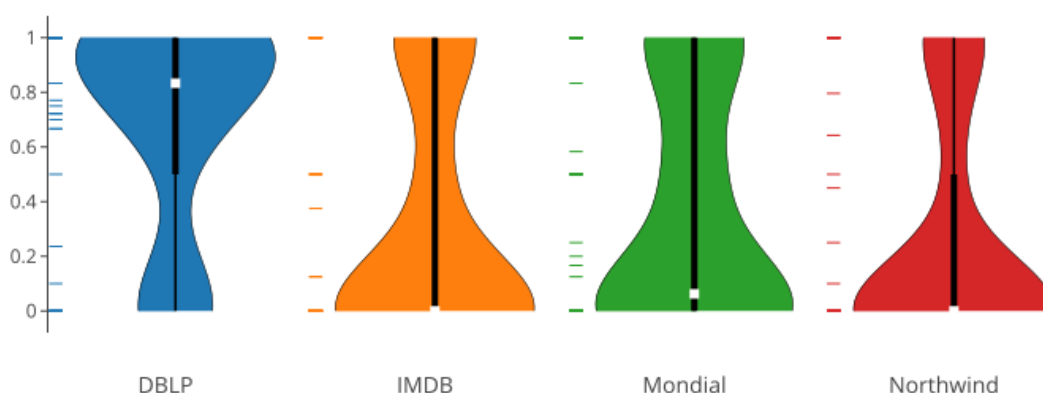


Figura 5.7: Variação dos resultados de *Average Precision* aplicadas Técnica Keymantic[7]

A Figura 5.7 destaca os resultados para o banco de dados DBLP que apresenta uma mediana dos resultados de *average precision (AP)* em torno de 0,8. Nesse mesmo banco de dados, 75% dos resultados encontram-se próximos a 1. Para os demais bancos de dados, os resultados da métrica apresentam medianas em zero ou próximo de zero. IMDB e Mondial apresentam 50% dos valores obtidos distribuídos entre zero e 1, onde, conforme a largura do gráfico, é possível observar que existe um percentual maior dos valores próximos a zero. Northwind apresenta 50% dos valores em zero.

Na Figura 5.8 os melhores resultados de *average precision* são obtidos com o banco de dados DBLP, em que aproximadamente 50% dos resultados estão distribuídos entre 0 e 1 com uma quantidade maior próximo de 1 em relação as demais bases de dados. No banco de dados Mondial mais de 75% dos resultados estão na faixa zero. Isso acontece devido à grande maioria das consultas não terem sido finalizadas em um tempo inferior a 1 hora definido para esta avaliação.

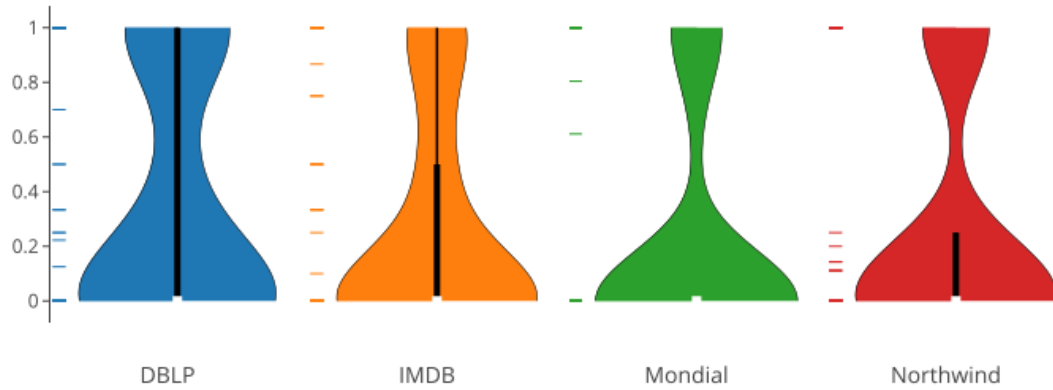


Figura 5.8: Variação dos resultados de Average Precision aplicadas a Técnica de Ramada et al.[39]

5.3.6 Mean Average Precision (MAP)

O MAP tem por objetivo calcular uma média das pontuações de precisão média. Nessa avaliação, por meio dos valores obtidos de *Average Precision (AP)* de cada consulta, apresentados nas Tabelas B.1, B.2, B.3, B.4, B.5, B.6, B.7 e B.8 do Apêndice B, foi calculado o MAP das técnicas para cada banco de dados e seus respectivos valores apresentados na Figura 5.9.

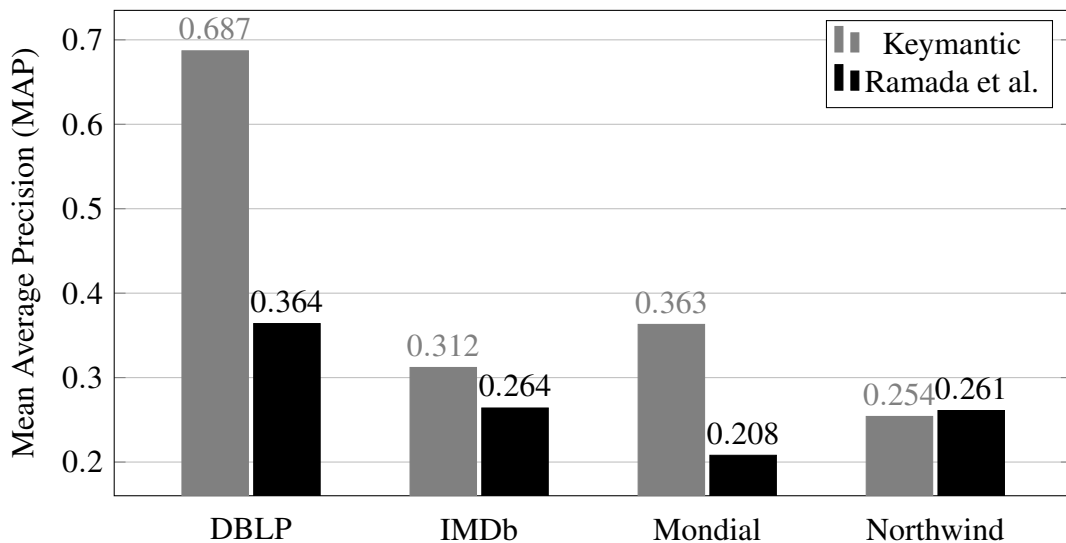


Figura 5.9: Mean Average Precision (MAP)

Os valores para essa métrica variam entre 0 e 1, sendo que os valores mais próximos de 1 expressam melhores resultados. Com o resultado desta avaliação, é possível destacar que a eficácia dos sistemas não são afetadas devido ao tamanho do banco de dados, visto que, IMDB que possui uma quantidade maior de tuplas possui resultados próximos dos obtidos pelas demais bases de dados. No entanto, existe uma diferença expressiva entre os valores das duas técnicas para o banco de dados Mundial. Isso ocorre

pois, durante a avaliação, grande parte dos resultados da técnica de Ramada et al não foram finalizados, não retornando assim resultados relevantes.

Para o banco de dado DBLP, a técnica Keymantic obtém melhor pontuação, isso é reflexo da quantidade de resultados relevantes retornados na primeira posição assim como apresentados na Figura 5.3. Com Keymantic, para o banco de dados DBLP, 72% das 50 consultas realizadas são retornadas na primeira posição isso reflete nessa métrica.

5.4 Discussão dos Resultados

Ao utilizar o *benchmark* proposto por este trabalho, foi possível identificar pontos importantes sobre cada técnica avaliada. As duas técnicas não tiveram bons resultados, visto que, assim como observado na avaliação, uma maior quantidade de consultas não obtém resultados relevantes. A avaliação inclusive destaca que um número alto de consultas não foi sequer finalizada com sucesso em um período menor que 1 hora.

Durante a avaliação foi diagnosticado problemas na codificação da técnica realizada por Ramada et al [39]. Durante a execução do código fonte para algumas consultas o sistema entrava em um laço de repetição infinito. Esse problema influenciou de forma negativa na avaliação realizada para a data técnica. No entanto, foi observado que o problema era de codificação e não um problema da técnica utilizada para realizar consultas por palavras-chave, ou seja, esse problema não pode ser levado em consideração para classificar mal o sistemas quanto a sua eficácia. Esse problema aconteceu principalmente no banco de dados Mundial devido a grande quantidade de relações e relacionamentos entre elas. Desta forma, sugere-se em trabalhos futuros a re-implementação de todas as técnicas para nova avaliação.

Uma ponto observado durante a avaliação, e que merece destaque, foi com relação ao desempenho das técnicas em realizar consultas em bancos de dados que possuem uma quantidade maior de relações, assim como é o caso dos bancos de dados Mondial e Northwind. As duas técnicas avaliadas produzem muitas configurações para esse bancos de dados assim como visto na Figura 5.2. Ramada et al. [39] apresenta valores baixos nessa figura pois no processo de geração de todas as combinações de configurações o código implementado extrapola o tempo definido por este *benchmark* de 1 hora para finalização.

Outro ponto a ser destacado nesta avaliação é com relação aos resultados retornados próximos as primeiras posições. Conforme a Figura 5.3 foi observado que, com exceção do resultado de Keymantic [7] para o banco de dados DBLP, as técnicas obtiveram baixos resultados retornados na primeira posição dos resultados. Esse mesmo comportamento se reflete na precisão nos 10 primeiros resultados retornados.

5.5 Considerações Finais

Para demonstrar a viabilidade da utilização do *benchmark* proposto, este capítulo abordou e categorizou as etapas da avaliação realizada, descreve como as técnicas a serem avaliadas foram implementadas, define o ambiente do experimento e apresenta a análise dos resultados retornados conforme as métricas.

A utilização do *benchmark* para avaliação das técnicas contribui para a identificação de pontos positivos e negativos de cada técnica avaliada. Isto pode ser observado por meio das métricas utilizadas. Por meio dos resultados desta avaliação é possível demonstrar o quanto cada técnica obtém sucesso ao retornar resultados relevantes nas primeiras posições.

Conclusão

Visando avaliar técnicas de consultas por palavras-chaves a bancos de dados relacionais que não possuem acessos aos dados durante a fase de interpretação da consulta, este trabalho propôs um *benchmark* com bancos de dados, consultas, avaliação de relevância e métricas padronizadas para melhor extrair pontos fortes e fracos de cada sistema.

As avaliações realizadas pelos pesquisadores não possuem um conjunto de dados e consultas padronizadas que melhor representassem dados do mundo real. Com base nos trabalhos relacionados, foi possível identificar as características importantes a serem avaliadas em cada técnica e com isso propor o conjunto de dados que melhor se adequasse a cada uma delas. Desta forma, espera-se que seja possível uma melhor identificação dos problemas de cada técnica e assim contribuir para melhoria do estado da arte.

Esta proposta define uma estrutura comum para avaliações de eficácia dos sistemas atuais e futuros. Ela foi projetada para contemplar todos os aspectos dos sistemas a serem avaliados. Visando contribuir para a avaliação de outros trabalhos, o conjunto de dados e consultas estão disponíveis em <http://inf.ufg.br/~jcs/benchmark/>.

A avaliação apresentada neste trabalho, utilizando o *benchmark* proposto, foi realizada com os sistemas Keymantic[7] e Ramada et al.[39]. Foram utilizados os códigos fontes desenvolvidos e disponibilizados pelos autores. Os resultados indicaram que nenhuma das duas técnicas obteve bons resultados para a eficácia da pesquisa. Ambas as técnicas possuem dificuldades em retornar resultados relevantes em bancos de dados com uma quantidade maior de relações. Mesmo em bancos de dados com poucas relações, para algumas consultas, os sistemas não conseguem retornar resultados relevantes entre os 10 primeiros valores do conjunto de resultados retornados para a consulta, o que contradiz as avaliações anteriores que aparecem na literatura.

As próximas seções apresentam uma síntese das principais contribuições obtidas a partir deste trabalho e as propostas de trabalhos futuros definidos a partir desta pesquisa.

6.1 Contribuições

As contribuições deste trabalho são apresentadas a seguir:

- **Formas de avaliação das técnicas de consultas por palavras-chave.** Foi realizada uma revisão de bibliografias buscando-se verificar as formas de avaliação utilizadas por cada técnica de consultas por palavras-chave a bancos de dados relacionais. Essa avaliação serviu como base para definir as melhores práticas a ser utilizado para a proposta deste trabalho.
- **Definição de bases de dados padronizadas.** Foram definidos 3 bancos de dados (DBLP, IMDB, Mondial) com cargas que representam dados reais, e o banco Northwind que é um banco sintético para serem utilizados como base para avaliações futuras. A escolha dos bancos de dados foi conforme o tamanho, quantidade de relações, quantidade de tuplas e estrutura do esquema do banco de dados.
- **Definição de um conjunto de consultas para avaliação.** Definição de um conjunto de 50 consultas com as necessidades de informações para cada banco de dados. Para definição desse conjunto de consultas foi levado em consideração além da qualidade de palavras-chave a quantidade de termos de esquema e de valor.
- **Formalização do *benchmark*.** Foi definido um *benchmark* onde é formalizado toda a metodologia de avaliação de sistema de consultas por palavras-chave com definições de métricas e avaliação de relevância.
- **Avaliação das técnicas Keymantic[7] e Ramada et al.[39].** Realização da avaliação de duas técnicas onde além de validar a proposta do *benchmark* pode-se constatar a eficácia dos sistemas.

6.2 Trabalhos Futuros

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

- O *benchmark* proposto por este trabalho se ateve a avaliar a eficácia dos resultados retornados pelas técnicas de consultas por palavras-chave a bancos de dados relacionais que não possuíam acesso aos dados durante a interpretação da consultas, no entanto, não foi adicionado a ele nenhuma forma de avaliar a eficiência. Como trabalho futuro devem ser consideradas métricas que afirmam o tempo gasto por cada técnica para concluir cada consulta para as bases de dados propostas por esta avaliação.
- Foram avaliadas por este trabalho as técnicas Keymantic [7] e Ramada et al. [39]. Como proposta de trabalho futuro pretende-se avaliar a técnica proposta por Sousa

et al. [41], assim como avaliar outras técnicas de consultas por palavras-chave a bancos de dados relacionais que possuem acesso aos dados durante a interpretação da consulta. Para essa avaliação, planeja-se realizar a implementação de todas as técnicas a serem avaliadas, desta forma, evita discrepâncias na comparação dos sistemas devido a codificações diferentes.

- Criação de um método de extração de consultas de usuários reais em um mecanismo de busca. É interessante definir um grupo de pessoas para definir um conjunto melhor das consultas com as necessidades reais de usuários. Essa forma de definição de consultas é utilizado por fóruns de avaliação padrão como o TREC[23] e INEX[19].
- A avaliação realizada neste trabalho foi realizada de forma manual, ou seja, foram executadas as consultas em cada técnica para cada banco de dados, após o retorno dos resultados da consulta foi verificado se as configurações geradas 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.

Referências Bibliográficas

- [1] **The mondial database.** <http://www.dbis.informatik.uni-goettingen.de/Mondial>. Accessed: 2017-02-30.
- [2] **Relational dataset repository.** <https://relational.fit.cvut.cz/>. Accessed: 2017-02-30.
- [3] AGRAWAL, S.; CHAUDHURI, S.; DAS, G. **Dbxplorer: a system for keyword-based search over relational databases.** In: *Proceedings 18th International Conference on Data Engineering*, p. 5–16, 2002.
- [4] BAEZA-YATES, R.; RIBEIRO-NETO, B. **Recuperação de Informação: Conceitos e Tecnologia das Máquinas de Busca.** Bookman Editora, 2013.
- [5] BAEZA-YATES, R.; RIBEIRO-NETO, B.; OTHERS. **Modern information retrieval**, volume 463. ACM press New York, 1999.
- [6] BERGAMASCHI, S.; DOMNORI, E.; GUERRA, F.; ORSINI, M.; LADO, R. T.; VELEGRAKIS, Y. **Keymantic: Semantic keyword-based searching in data integration systems.** *Proc. VLDB Endow.*, 3(1-2):1637–1640, Sept. 2010.
- [7] BERGAMASCHI, S.; DOMNORI, E.; GUERRA, F.; TRILLO LADO, R.; VELEGRAKIS, Y. **Keyword search over relational databases: A metadata approach.** In: *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, SIGMOD '11, p. 565–576, New York, NY, USA, 2011. ACM.
- [8] 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, 2002.
- [9] BORAL, H.; DEWITT, D. J. **A methodology for database system performance evaluation.** *SIGMOD Rec.*, 14(2):176–185, June 1984.
- [10] CLEVERDON, C. **The cranfield tests on index language devices.** *Links*, 942:42, 1997.

- [11] 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, Jan 2014.
- [12] COFFMAN, J.; WEAVER, A. C. **A framework for evaluating database keyword search strategies.** In: *Proceedings of the 19th ACM International Conference on Information and Knowledge Management, CIKM '10*, p. 729–738, New York, NY, USA, 2010. ACM.
- [13] 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.
- [14] DING, B.; YU, J. X.; WANG, S.; QIN, L.; ZHANG, X.; LIN, X. **Finding top-k min-cost connected trees in databases.** In: *2007 IEEE 23rd International Conference on Data Engineering*, p. 836–845, April 2007.
- [15] DOMNORI, E. **Keyword based search engine over relational database.** <https://sourceforge.net/p/keymantic/wiki/Home/>, jan 2017.
- [16] ELMASRI, R.; NAVATHE, S. **Fundamentals of database systems.** Addison-Wesley Publishing Company, 2010.
- [17] FAKHRAEE, S.; FOTOUHI, F. **Dbsemsplorer: Semantic-based keyword search system over relational databases for knowledge discovery.** In: *Proceedings of the Third International Workshop on Keyword Search on Structured Data, KEYS '12*, p. 54–62, New York, NY, USA, 2012. ACM.
- [18] FREITAS, R. A. P.; RAMALHO, J. C. **Significant properties in the preservation of relational databases.** In: *Research and Advanced Technology for Digital Libraries, 14th European Conference, ECDL2010*. Springer, 2010.
- [19] GÖVERT, N. F. N.; LALMAS, G. K. M. **Inex: Initiative for the evaluation of xml retrieval**, 2003.
- [20] GRAY, J. **The Benchmark Handbook for Database and Transaction System.** Morgan Kaufmann Publishers Inc., 1993.
- [21] GU, J.; KITAGAWA, H. **Extending keyword search to metadata on relational databases.** In: *2008 International Workshop on Information-Explosion and Next Generation Search*, p. 97–103, April 2008.

- [22] HAAM, D.; LEE, K. Y.; KIM, M. H. **Keyword search on relational databases using keyword query interpretation.** In: *5th International Conference on Computer Sciences and Convergence Information Technology*, p. 957–961, Nov 2010.
- [23] HARMAN, D. **Overview of the second text retrieval conference (trec-2).** In: *Proceedings of the workshop on Human Language Technology*, p. 351–357. Association for Computational Linguistics, 1994.
- [24] HE, H.; WANG, H.; YANG, J.; YU, P. S. **Blinks: Ranked keyword searches on graphs.** In: *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*, SIGMOD '07, p. 305–316, New York, NY, USA, 2007. ACM.
- [25] HRISTIDIS, V.; PAPAKONSTANTINOY, Y. **Chapter 58 - discover: Keyword search in relational databases.** In: Bernstein, P. A.; ; Ioannidis, Y. E.; ; Ramakrishnan, R.; ; Papadias, D., editors, *{VLDB} '02: Proceedings of the 28th International Conference on Very Large Databases*, p. 670 – 681. Morgan Kaufmann, San Francisco, 2002.
- [26] HRISTIDIS, V.; PAPAKONSTANTINOY, Y.; GRAVANO, L. **Efficient ir-style keyword search over relational databases*.** In: Freytag, J.-C.; ; Lockemann, P.; ; Abiteboul, S.; ; Carey, M.; ; Selinger, P.; ; Heuer, A., editors, *Proceedings 2003 {VLDB} Conference*, p. 850 – 861. Morgan Kaufmann, San Francisco, 2003.
- [27] KACHOLIA, V.; PANDIT, S.; CHAKRABARTI, S.; SUDARSHAN, S.; DESAI, R.; KARAMBELKAR, H. **Bidirectional expansion for keyword search on graph databases.** In: *Proceedings of the 31st International Conference on Very Large Data Bases*, VLDB '05, p. 505–516. VLDB Endowment, 2005.
- [28] KARGAR, M.; AN, A.; CERCONE, N.; GODFREY, P.; SZLICHTA, J.; YU, X. **Meanks: Meaningful keyword search in relational databases with complex schema.** In: *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, SIGMOD '14, p. 905–908, New York, NY, USA, 2014. ACM.
- [29] KASNECI, G.; RAMANATH, M.; SOZIO, M.; SUCHANEK, F. M.; WEIKUM, G. **Star: Steiner-tree approximation in relationship graphs.** In: *2009 IEEE 25th International Conference on Data Engineering*, p. 868–879, March 2009.
- [30] LI, G.; FENG, J.; ZHOU, X.; WANG, J. **Providing built-in keyword search capabilities in rdbms.** *The VLDB Journal*, 20(1):1–19, Feb 2011.
- [31] LI, G.; JI, S.; LI, C.; FENG, J. **Efficient type-ahead search on relational data: A tastier approach.** In: *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data*, SIGMOD '09, p. 695–706, New York, NY, USA, 2009. ACM.

- [32] LIU, F.; YU, C.; MENG, W.; CHOWDHURY, A. **Effective keyword search in relational databases**. In: *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data*, SIGMOD '06, p. 563–574, New York, NY, USA, 2006. ACM.
- [33] LUO, Y.; LIN, X.; WANG, W.; ZHOU, X. **Spark: Top-k keyword query in relational databases**. In: *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*, SIGMOD '07, p. 115–126, New York, NY, USA, 2007. ACM.
- [34] MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H.; OTHERS. **Introduction to information retrieval**, volume 1. Cambridge university press Cambridge, 2008.
- [35] MOTL, J.; SCHULTE, O. **The ctu prague relational learning repository**. *arXiv preprint arXiv:1511.03086*, 2015.
- [36] NANDI, A.; JAGADISH, H. V. **Assisted querying using instant-response interfaces**. In: *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*, SIGMOD '07, p. 1156–1158, New York, NY, USA, 2007. ACM.
- [37] PU, K. Q.; YU, X. **Frisk: Keyword query cleaning and processing in action**. In: *2009 IEEE 25th International Conference on Data Engineering*, p. 1531–1534, March 2009.
- [38] RAMADA, M.; DA SILVA, J. C.; JÚNIOR, P. D. S. L. **Data extraction from structured databases using keyword-based queries**. In: *SBBD*, p. 57–66, 2014.
- [39] RAMADA, M. S.; DA SILVA, J. C.; DE SÁ LEITAO-JÚNIOR, P. **A method for semantic analysis of keyword-based queries to access information in web databases**. In: *Proceedings of the IADIS International Conference WWW/Internet*, p. 35–42. IADIS, 2014.
- [40] SOLID IT. **Db-engines ranking of relational dbms**. <https://db-engines.com/en/ranking>, Fev 2018.
- [41] SOUSA, W. P.; OTHERS. **Uma técnica para ranqueamento de interpretações sql oriundas de consultas com palavras-chave**. Master's thesis, Universidade Federal de Goiás, 2017.
- [42] STAPENHURST, T. **The benchmarking book**. Routledge, 2009.
- [43] TATA, S.; LOHMAN, G. M. **Sqak: Doing more with keywords**. In: *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, p. 889–902, New York, NY, USA, 2008. ACM.
- [44] TPC. **Tpc benchmarks**. <http://www.tpc.org>, Fev 2018.

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

Consultas e Sentidos Pretendidos

Tabela A.1: Consultas para o Banco de Dados DBLP

Nº	Consulta	Sentido
1	springer	Informações sobre o periódico springer
2	dblp	Informações sobre a base dblp
3	ISBN	Informações sobre os ISBNs
4	author	Informações sobre autores de artigos
5	publisher	Informações sobre editores
6	series	Informação sobre as series das conferências/periódicos
7	title springer	Informações sobre o periódico springer
8	author morgan	Informações sobre o autor morgan
9	author inproceeding	Informações sobre autores e os artigos por eles escritos
10	author name	Nome dos autores cadastrados
11	inproceeding title	Título dos artigos
12	proceeding title	Nome das conferências/periódicos
13	publisher name	Nome dos editores
14	publisher springer	Informações sobre o editor elsevier
15	series title	Título das series das conferências/periódicos
16	author name jacques	Informações sobre o autor de nome jacques
17	author id 2	Informações sobre autor com id = 2
18	inproceeding pages 10	Informações sobre artigos com na página 10
19	proceeding year 2002	Conferências do ano de 2012
20	author proceeding 2000	Informações sobre autores em periódicos no ano de 2010
21	author inproceeding 2001	Autores com artigos publicados em 2010
22	proceeding title year	Nome e ano dos periódicos/conferências
23	proceeding title database	Informações sobre periódicos/conferências de banco de dados
24	proceeding url dblp	Informações sobre periódicos/conferências com url dblp
25	publisher name elsevier	Informações sobre o editor de nome elsevier
26	author name bruno	Informações sobre autor de nome bruno
27	series title computing	Informação sobre series com título computing
28	author name proceeding	Nome dos autores e conferências/periodicos que os mesmos participaram
29	proceeding ISBN title	Isbn e título das conferências/periódicos
30	author inproceeding title graph	Informações sobre autores e os artigos por eles escritos com título contendo "graph"
31	inproceeding title year 1995	Título dos artigos publicados no ano 1995
32	proceeding year 1992 publisher	Informações sobre editores das conferências/periódicos do ano 2000
33	author name proceeding title	Nome dos autores e títulos das conferências/periodicos que os mesmos participaram
34	proceeding title year 1993	Título das conferências/periódicos que ocorreram em 1993
35	author name series "computer science"	Nome dos atores que participarem das series de ciência da computação
36	author name inproceeding title	Nome dos autores e títulos dos artigos que esses autores publicaram
37	author inproceeding publisher name "IEEE Computer Society"	Informações sobre autores que publicaram artigos pela editora "IEEE Computer Society"
38	inproceeding title publisher name "AAAI Press"	Informações sobre artigos publicados pela editora "AAAI press"
39	author name proceeding year 2002	Nome dos autores que publicaram em periódicos / conferências em 2010
40	author name pierre proceeding year	Informações sobre as conferências/periódicos que o autor pierre publicou
41	proceeding year 2000 publisher "Oxford University Press"	Informações sobre as conferências/periódicos que publicaram artigos pela editora "Oxford University Press"
42	author name proceeding title "artificial evolution"	Nome dos autores que participaram das conferências/periódicos com título "artificial evolution"
43	publisher name proceeding year 1991	Nome dos editores que publicaram nas conferências / periódicos em 1991
44	author name inproceeding pages 8	Nome dos autores que possuem artigos com 8 páginas
45	author Dominique inproceeding series title	Título das séries dos artigos publicados pelo autor Dominique
46	author name pierre proceeding year 1993	Informações sobre as conferências/periódicos que o autor pierre publicou no ano de 1998
47	series title information proceeding year 2001	Informações sobre series com título information de conferências/periódicos do ano 2005
48	proceeding year 2002 publisher name acm	Periodicos/conferencias publicadas pela ACM em 2002
49	author name inproceeding title proceeding title publisher	Nome dos autores, títulos dos artigos publicados por eles nas conferências / periódicos de 2011
50	author name patrick inproceeding title proceeding year 2001	Títulos dos artigos publicados pelo autor Patrick nas conferências / periódicos de 2011

Tabela A.2: Consultas para o Banco de Dados 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
11	actor "tom hanks"	Informações sobre o ator Tom Hanks
12	actors name	Lista de nomes de atores
13	movies genre	Lista de generos de filmes
14	movie "Wizard of oz"	Informações sobre o filme Wizard of Oz
15	movie 2012	Informações sobre o filme 2012
16	movies 1984	Lista de Filmes lançados em 1984
17	movies Comedy	Lista de filmes de comédia
18	"Indiana Jones and the Last Crusade" roles	Personagens do filme "Indiana Jones and The Last Crusade"
19	director "James Cameron"	Informações sobre diretor James cameron
20	name "Jacques Clouseau"	Informações sobre "Jacques Clouseau"
21	movie name higher	Informações sbre o filme Higher
22	movies genres drama	Informações sobre os filmes do Gênero Drama
23	documentary year 2012	Documentários que foram lançados no ano de 2012
24	movie titanic director	Diretor do filme Titanic
25	movie "spider man"genre	Gênero cinematográfico que o filme "Spider-Man" se enquadra
26	movie titanic actors	Atores que atuaram no filme Titanic
27	actor "Leonardo DiCaprio"movies	Filmes que o ator Leonardo DiCaprio atuou
28	movie "mission impossible"genre	Gênero do filme "Mission Impossible"
29	movie name "independece day"type	Qual o tipo do filme "Independece Day"
30	movie name higher rank	O nome do filme com o maior rank O rank do filme de nome Higher
31	movie name "independece day"genre	Qual o gênero do filme Independece Day
32	movie name actor Tom	Nomes dos filmes que o ator Tom Cruise atuou
33	comedy movies actor "Jim Carrey"	Filmmes de comédia em que o ator Jim Carey participou
34	"Quentin Tarantino"action movies 1998	Filmes de ação lançados em 1998 e que foram dirigidos por Quentin Tarantino
35	name james cameron year	Filmes com nomes James Cameron dirigiu por ano
36	russell crowe gladiator role	Papel de Russell Crowe no filme Gladiador
37	Actor leonardo movies year 2000	Filmes lançados no ano 2000 qnos quais o ator Leonardo participou
38	movie director spilberg actor leonardo	Filmes dirigidos por Spielberg nos quais o ator Leonardo atua
39	actor name movie year 2000	Nome dos atores que atuaram nos filmes lançados no ano 2000
40	movies year 2000 higher rank	Filmes lançados no ano 2000 com maior pontuação
41	actor "jackie chan"movies year 2000	Filmes lançados no ano 2000 que o ator Jackie Chan participou
42	genre movies name year 1950	Gênero e nome dos filmes lançados no ano de 1950
43	film name "the godfather"player gender F	Atores do gênero Feminino do filme com nome "The Godfather"
44	actors name action movies director quentin	Nome dos atores que atuaram nos filmes dirigidos pelo diretor Quentin
45	Char name movie "Back to the Future"gender F	Nome dos personagens do filme "Voltar ao Futuro"do gênero feminino
46	movies name genre comedy year 1989	Nome dos filmes de gênero comédia que foram lançados no ano de 1989
47	drama movies name actors m year 1995	Atores do sexo masculino que atuaram em filmes de drama no ano de 1995
48	actor gender m movie year 1994	Atores do gênero masculino que atuaram em filmes no ano de 1994
49	movies year 1986 actors gender F roles	Papeis dos filmes do ano de 1986 com atores do gênero Feminino
50	actors name gender f drama movies year 1990	Atores do sexo feminino que atuaram em filmes de drama no ano de 1990

Tabela A.3: Consultas para o Banco de Dados Mundial

Nº	Consulta	Sentido
1	Brazil	Informações acerca do Brasil
2	Africa	Informações acerca do Continente Africano
3	Mediterranean	Informações sobre o mar mediterraneo
4	Java	Informações sobre o Ilha de Java
5	Country	Lista de países
6	Lakes	Lista de lagos
7	Sea	Lista de mares
8	Island	Lista de ilhas
9	Continent	Lista de continentes
10	country name	Nome de todos os países
11	country capital	Nome de todas as capitais de todos os países
12	Continent Island	Informações sobre ilhas de cada continente
13	Country Province	Informações de estados de cada país
14	Caribbean Economic	Informações sobre economia caribenha
15	Desert Atacama	Informações sobre o deserto do Atacama
16	country china	Informações sobre o país de nome China
17	Sea Brazil	Informações sobre mares do Brasil
18	River Amazonas	Informações sobre o rio Amazonas
19	country china province	Informações sobre os estados da china
20	Country Brazil river	Informações sobre os rios do Brasil
21	continent america country	Informações sobre os países do continente americano
22	country "south africa"religion	Religião(ões) predominante(s) na África do Sul
23	canada province lake	Informações sobre rios por estados do canada
24	country brazil amazonas	Informações sobre o Estado do Amazonas do País Brasil Informações sobre o rio amazonas do Brasil
25	Population Country Italy	Informações sobre a População do país Itália
26	Mountains Everest Himalaya	Informações sobre a Montanha Everest no Himalaya
27	Language Country name	Línguas faladas por cada país
28	Mountains name Elevation	Nome de montanhas com suas altitudes
29	country province capital	Informações sobre as capitais dos estados dos países
30	Island name coordinates	Nome de ilhas e suas coordenadas
31	country china province gansu	Informações sobre a província chinesa de Gansu
32	Catholic religion European continent	Informações sobre religião católica no continente europeu
33	country brazil province capital	Informações sobre a capital dos estados brasileiros
34	sea name "indian ocean"slands	Informações sobre as ilhas do oceano índico
35	country name canada lake	Informações sobre os lagos canadenses
36	country japan economy agriculture	Informações econômicas sobre a agricultura japonesa
37	province name "minas gerais"mountains	Informações sobre as montanhas do estado de Minas Gerais
38	Country name Africa continent	Informações sobre nome dos países do continente Africano
39	continent africa country sea	Nome dos mares que banham o continente africano
40	country name brazil religion percentage	Informações sobre os percentuais das religiões do Brasil
41	country name japan economy agriculture	Informações econômicas sobre a agricultura japonesa
42	sea name "pacific ocean"river name	Nome dos rios que desaguam no oceano Pacífico
43	continent name africa country sea	Nome dos mares que banham o continente africano
44	country name italy island name	Nome das ilhas italianas
45	country brasil province pernambuco city recife	Informações sobre a cidade brasileira do estado do Pernambuco Recife
46	rivers country name brazil province sergipe	Informações sobre os rios do estado de Sergipe
47	city name delhi river sea "Indian Ocean"	Informações sobre os rios que banham a cidade de Delhi e que desaguam no oceano Índico
48	country name egypt province desert name	Informações sobre estados do Egito que possuem deserto
49	language portuguese percent continent name europe country	Informações sobre o percentual dos habitantes dos países do continente europeu que falam português
50	country name chile sea name "pacific ocean" river	Nomes dos rios chilenos que deságuam no oceano pacífico

Tabela A.4: Consultas para o Banco de Dados Northwind

Nº	Consulta	Sentido
1	"Robert King"	Informações sobre o colaborador "Robert king"
2	Boston	Informações sobre o território "Boston"
3	Mayumi	Informações sobre a fornecedora "Mayumi"
4	customers	Informações sobre clientes
5	employees	Informações sobre colaboradores
6	orders	Informações de compras efetuadas
7	products	Informações sobre produtos
8	suppliers	Informações sobre revendedores
9	employee "Andrew Fuller"	Informações sobre o colaborador "Andrew Fuller"
10	supplier "tokyo Traders"	Informações sobre o revendedor "tokyo Traders"
11	product Tourtire	Informações sobre o produto "Tourtire"
12	Orders Brazil	Informações sobre os pedidos efetuados com destino ao Brasil
13	Orders 2016	Informações sobre pedidos realizados no ano de 2016
14	Products Categories	Informações sobre produtos e suas categorias
15	Employees Region	Informações sobre colaboradores e suas regiões de atuação
16	Products Suppliers	Informações sobre produtos e seus revendedores
17	Orders Territories	Informações sobre os pedidos e seus territórios
18	employee Janet Leverling	Informações sobre a colaboradora Janet Leverling
19	Orders Products Ikura	Informações sobre as encomendas dos produtos "Ikura"
20	Employees Region Northern	Informações sobre os coladoradores da região norte
21	Suppliers City London	Informações sobre revendedores da cidade de Londres
22	Customer Berlin Germany	Informações sobre clientes da cidade alemã Berlim
23	Products Category Cheeses	Informações sobre os produtos que pertencem à categoria queijos
24	Count customers Canada	Quantidade de clientes do Canada
25	Average Orders customers	Informações sobre a média das encomendas realizadas pelos clientes
26	employee Janet Orders	Informações sobre as encomendas vendidas pela colaboradora Janet
27	Customer region "New York"	Informações sobre clientes da região "New York"
28	Shippers "Federal Shipping"Canada	Informações sobre o remetente canadense "Federal Shipping"
29	employee Orders date 1996	Informações sobre as encomendas de 1996
30	Products orders date 1995	Informações sobre os produtos das encomendas realizadas em 1995
31	Products Orders Employee Michael	Informações sobre encomendas e produtos vendidos pelo colaborador Michael
32	Customer region Boston employees	Informações sobre clientes da região de Boston que são empregados
33	Customer Company Name "Familia Arquibaldo"	Informações sobre clientes da empresa com nome "Familia Arquibaldo"
34	Customer city "São Paulo"Pedro	Informações sobre o cliente Pedro da cidade de São Paulo
35	suppliers Melbourne products tofu	Informações sobre distribuidores de tofu do território de Melbourne
36	Orders orderID 36 territories	Informações sobre a encomenda com ID 36 e seu território
37	Employee Orders territories region	Informações sobre as encomendas, seus territórios e regiões
38	Costumers Orders territories region	Informações sobre as clientes, suas encomendas, territórios e regiões
39	Employee firt_name Costumers Country Canada	Informações sobre o primeiro nome dos colaboradores que atenderem clientes com informações do país
40	Products Orders Employee Firstname Laura	Informações sobre os produtos e encomendas atendidas pelo colaboradora Laura
41	Products Categories beers suppliers Frankfurt	Informações sobre produtos da categoria cervejas e seus revendedores
42	Quantity productname chai order RJ	Informações sobre a quantidade de chai vendidos nas encomendas com destino ao Rio de Janeiro
43	employees lastname Peacock region postcode	Informações sobre os colaboradores que possuem sobrenome Peacock, as regiões que atendem e código postal
44	Region Southern Employees orders orderid 150	Informações sobre colaboradores que atendem a região sul e a encomenda de id=150
45	Territory Atlanta Employee orders "order ID"5	Informações sobre os colaboradores do Território Atlanta e a encomenda de id = 5
46	Margaret orders product categories categoryName cereal	Informações sobre as encomenda atendidas pela colaboradora Margaret que possuem produtos da categoria cereal
47	Costumers France orders product supliers country	Informações sobre as encomendas com destino à França, produtos, seus revendedores e países
48	Region Southern Employees orders orderid 150 products	Informações sobre colaboradores que atendem a região sul da encomenda com id = 50
49	Employee Andrew orders product categories categoryName bread	Informações sobre informações sobre as encomendas atendidas pelo colaborator Andrew que contenham produtos da categoria pão
50	Costumers country France orders product categories fish	Informações sobre clientes franceses e suas encomendas que possuem produtos da categoria peixe

Precisão $n(P@n)$ das Técnicas Avaliadas

Tabela B.1: *Precisão $n(P@n)$ da técnica Keymantic[7] no banco de dados DBLP*

Nº	P@1	P@2	P@3	P@4	P@5	P@6	P@7	P@8	P@9	P@10	Average Precision
1	0	0	0	0,25	0,2	0,1667	0,1429	0,125	0,2222	0,2	0,2361
2	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0
4	1	1	0	0	0	0	0	0	0	0	1
5	1	1	0	0	0	0	0	0	0	0	1
6	1	0	0	0	0	0	0	0	0	0	1
7	0	0	0	0	0	0	0	0	0	0	0
8	1	1	0,6667	0,5	0	0	0	0	0	0	1
9	1	0,5	0,6667	0,75	0,6	0,6667	0	0	0	0	0,7708
10	1	1	0,6667	0,5	0	0	0	0	0	0	1
11	1	0,5	0,3333	0,25	0,4	0,3333	0,2857	0,25	0	0	0,7
12	1	1	0,6667	0,5	0,4	0,3333	0,2857	0,25	0	0	1
13	1	1	0,6667	0,5	0	0	0	0	0	0	1
14	0	0,5	0,3333	0,25	0	0	0	0	0	0	0,5
15	1	0,5	0,3333	0	0	0	0	0	0	0	1
16	1	1	0,6667	0,5	0,4	0,3333	0,2857	0,25	0	0	1
17	0	0	0	0	0	0	0	0	0	0	0
18	1	0,5	0,3333	0,25	0	0	0	0	0	0	1
19	1	0,5	0,3333	0,25	0	0	0	0	0	0	1
20	0	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0
22	1	1	0,6667	0,5	0,4	0,3333	0,2857	0,25	0,2222	0	1
23	1	0,5	0,3333	0,5	0,4	0,3333	0,2857	0,25	0,2222	0,2	0,75
24	0	0	0	0	0	0	0	0	0	0	0
25	1	0,5	0,6667	0,5	0,4	0,3333	0,2857	0,25	0	0	0,8333
26	1	1	0,6667	0,5	0,4	0,3333	0,2857	0,25	0	0	1
27	1	0,5	0,3333	0,25	0,2	0,1667	0	0	0	0	1
28	1	1	0,6667	0,5	0,4	0,3333	0,2857	0,25	0,2222	0,2	1
29	1	1	0,6667	0,5	0,4	0,3333	0,2857	0,25	0,2222	0	1
30	1	0,5	0,6667	0,5	0,4	0,3333	0,2857	0,25	0,2222	0,2	0,8333
31	0	0	0	0	0	0	0	0	0	0	0
32	1	0,5	0,6667	0,5	0,4	0,5	0,4286	0,375	0,3333	0,3	0,7222
33	1	1	0,6667	0,5	0,4	0,3333	0,2857	0,25	0,2222	0,2	1
34	1	0,5	0,6667	0,5	0,4	0,3333	0,2857	0,25	0,2222	0,2	0,8333
35	0	0	0	0	0	0	0	0	0	0	0
36	1	1	0,6667	0,5	0,4	0,5	0,4286	0,375	0,3333	0,3	0,8333
37	0	0	0	0	0	0	0	0	0	0,1	0,1
38	1	0,5	0,6667	0,5	0,4	0,3333	0,2857	0,25	0,2222	0,2	0,8333
39	1	0,5	0,6667	0,5	0,4	0,3333	0,2857	0,25	0,2222	0,2	0,8333
40	1	1	0,6667	0,5	0,4	0,3333	0,2857	0,25	0,2222	0,2	1
41	0	0	0	0	0	0	0	0	0	0	0
42	1	0,5	0,6667	0,5	0,4	0,3333	0,2857	0,25	0,2222	0,2	0,8333
43	1	0,5	0,6667	0,5	0,4	0,5	0,4286	0,375	0,3333	0,3	0,7222
44	1	0,5	0,6667	0,5	0,4	0,5	0,4286	0,375	0,3333	0,3	0,7222
45	1	1	0,6667	0,5	0,4	0,3333	0,2857	0,25	0,2222	0,2	1
46	1	0,5	0,3333	0,25	0,2	0,3333	0,2857	0,25	0,2222	0,2	0,6666
47	1	0,5	0,3333	0,25	0,2	0,1667	0,1429	0,125	0,1111	0,1	1
48	1	0,5	0,3333	0,25	0,2	0,3333	0,2857	0,25	0,2222	0,2	0,6666
49	1	1	1	1	0,8	0,6667	0,5714	0,5	0,4444	0,4	1
50	1	0,5	0,3333	0,25	0,2	0,1667	0,1429	0,125	0,1111	0,1	1

Tabela B.3: Precisão $n(P@n)$ da técnica Keymantic[7] no banco de dados Mondial

[illegible]

Tabela B.4: Precisão $n(P@n)$ da técnica Keymantic[7] no banco de dados Northwind

[illegible]

Tabela B.6: *Precisão $n(P@n)$ da técnica Ramada[39] no banco de dados IMDB*

Nº	P@1	P@2	P@3	P@4	P@5	P@6	P@7	P@8	P@9	P@10	Average Precision
1	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0
6	1	0	0	0	0	0	0	0	0	0	1
7	1	0	0	0	0	0	0	0	0	0	1
8	1	0	0	0	0	0	0	0	0	0	1
9	1	1	0	0	0	0	0	0	0	0	1
10	1	0	0	0	0	0	0	0	0	0	1
11	0	0	0	0	0	0	0	0	0	0	0
12	1	0	0	0	0	0	0	0	0	0	1
13	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0
15	0	0,5	0,33	0,25	0	0	0	0	0	0	0,5
16	0	0	0,33	0,25	0	0	0	0	0	0	0,3333
17	0	0	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0
21	1	1	1	1	0,8	0,67	0	0	0	0	1
22	1	0,5	0,33	0,25	0,2	0,17	0,14	0,13	0	0	1
23	0	0	0	0	0	0	0	0	0	0	0
24	1	1	0,67	0,5	0,6	0,5	0,43	0	0	0	0,8666
25	0	0,5	0,33	0,25	0,2	0,17	0,14	0	0	0	0,5
26	0	0	0	0,25	0,2	0,17	0,14	0,13	0,11	0,1	0,25
27	0	0	0	0	0	0	0	0	0	0	0
28	0	0,5	0,33	0,25	0,2	0,17	0,14	0	0	0	0,5
29	0	0	0	0	0	0	0	0	0	0	0
30	0	0	0	0	0	0	0	0	0	0	0
31	1	0,5	0,33	0,5	0,4	0,33	0,29	0,25	0,22	0,2	0,75
32	1	0,5	0,33	0,5	0,4	0,33	0,29	0,25	0,22	0,2	0,75
33	0	0	0	0	0	0	0	0	0	0	0
34	0	0	0	0	0	0	0	0	0	0	0
35	0	0	0	0	0	0	0	0	0	0	0
36	0	0	0	0	0	0	0	0	0	0	0
37	0	0	0	0	0	0	0	0	0	0	0
38	0	0	0	0	0	0	0	0	0	0	0
39	0	0	0	0	0	0	0	0	0	0	0
40	0	0	0,33	0,25	0	0	0	0	0	0	0,3333
41	0	0	0	0	0	0	0	0	0	0	0
42	0	0	0,33	0,25	0,2	0,17	0,14	0,13	0,11	0,1	0,3333
43	0	0	0	0	0	0	0	0	0	0	0
44	0	0	0	0	0	0	0	0	0	0	0
45	0	0	0	0	0	0	0	0	0	0	0
46	0	0	0	0	0	0	0	0	0	0	0
47	0	0	0	0	0	0	0	0	0	0	0
48	0	0	0	0	0	0	0	0	0	0,1	0,1
49	0	0	0	0	0	0	0	0	0	0	0
50	0	0	0	0	0	0	0	0	0	0	0

