Realização de Consultas Textuais Utilizando Oracle Text e Associação com Tesauro

Luiz B. Pimentel[1], Christian F. Peres[1], Luis Alexandre da Silva[1]

[1]Curso de Tecnologia em Banco de Dados - Faculdade de Tecnologia de Bauru (FATEC)
Rua Manoel Bento da Cruz, nº 30 Quadra 3 - Centro - 17.015-171 - Bauru, SP - Brasil

{luizgibber@gmail.com, christ.peres@gmail.com, luis.silva51@fatec.sp.gov.br}

Abstract. *The data query in database tables that contain many text fields, need alternatives to return all related results and correlated the searched words, generating a greater quantity and quality in the search results. The objective of this study is to verify the behavior and the number of responses returned when used selection methods using Like, Oracle Text and association of related words using the use of a thesaurus. Through experiments, it was possible to conclude that the sequence Oracle Text by the use of a thesaurus makes searches that are more efficient.*

Resumo. *A consulta de dados em tabelas de banco de dados que contém grande quantidade de campos textuais necessita de alternativas que retornem todos os resultados relacionados e correlacionados as palavras pesquisadas, gerando uma maior quantidade e qualidade nos resultados das pesquisas. O objetivo deste estudo é verificar o comportamento e a quantidade de respostas retornadas quando utilizados métodos de seleção usando Like, Oracle Text e associação de palavras correlacionadas utilizando o recurso de um tesauro. Através de uma sequência de experimentos foi possível concluir que o Oracle Text junto ao uso de um tesauro torna as pesquisas mais eficientes.*

1.Introdução

O registro de informações sobre diversos assuntos existe desde a antiguidade, onde as informações sobre diversos assuntos eram registradas nas pedras ou paredes de cavernas, posteriormente surgiram o papel, os livros, as bibliotecas, que acompanharam a humanidade por muitos séculos, até a criação dos computadores. Com o surgimento da informática e com o aumento da quantidade de dados, surgem os Banco de Dados (BD), recurso que compreende o container dos dados que posteriormente podem ser acessados e transformados em informações, sendo essenciais para a recuperação de informações.

O armazenamento de dados pode trazer diferentes tipos de interpretação, pois a construção de um vocabulário sofre constante modificação. Significados ganham novos valores constantemente e o que hoje pode trazer um grande uso, no futuro pode ser desconhecido ou considerado ultrapassado. A dificuldade gerada é a localização das informações porque elas podem se manifestar de formas variadas.

Quando se explora algo sem ter certeza, não se sabe o que pode ser obtido como resposta. Então todas as possibilidades devem ser pensadas e analisadas. Ao se executar uma busca dentro de um BD Relacional é necessário pensar como os outros usuários pensariam. Uma alternativa para este fato seria para aplicar ao BD um tesauro.

Ao utilizar um método de busca normal, são procurados os termos em um dicionário que está registrado a expressão utilizada por palavras semelhantes que podem também não ter o mesmo sinônimo ou contexto. Também as buscas podem ser feitas em dicionários que contém terminologias que contém o assunto. Nas opiniões de Krieger (2003) e Finatto (2004), um difere do outro somente na abrangência. Segundo Bocorny et al (2010) e Lancaster (2004), ao utilizar o método tesauro serão localizados também sinônimos e antônimos podendo gerar dezenas a centenas de termos a partir de uma única palavra.

A proposta deste trabalho é investigar o uso do tesauro em pesquisas de Banco de Dados Relacionais (BDR) para verificar se esse recurso pode trazer melhores resultados de pesquisa. Dessa forma, será possível verificar se juntos funcionam de forma correta trazendo algum tipo de benefício ao usuário.

Neste artigo serão apresentados os principais conceitos para o entendimento do assunto, definição dos materiais utilizados com a preparação do ambiente para execução dos testes, realização dos experimentos e resultados obtidos com suas conclusões.

2. Thesaurus

Tesauro é uma palavra de origem grega e latina que pode ser definida como léxico ou tesouro de palavras. Sendo que ficou conhecida quando publicada em 1852, na Inglaterra, a palavra “Thesaurus” (Roget, 1925). Segundo UNESCO (1973), tesauro é definido dependendo do aspecto que está sendo analisado. Se for considerar pela estrutura é um vocabulário controlado e dinâmico que relaciona termos através de suas formas semânticas e genéricas sobre um tema especifico. Quando considerada a função é definido como um dispositivo de controle terminológico que é usado para tradução da linguagem natural dos documentos, dos indexadores ou dos usuários numa linguagem do sistema.

Para Campos (2006), tesauro é uma linguagem documentaria que se baseia na relação de termos que tem por objetivo indexar ou recuperar uma informação. Segundo Lu (1995) é um conhecimento que envolve semântica e conceito que pode revelar o relacionamento de semelhança em informações de origens diferentes.

Em seu trabalho sobre indexação, Vieira et al. (2010) define tesauro como uma ferramenta que permite controlar termos atribuídos e é um instrumento de apoio para organizar, representar e recuperar informações dentro de um determinado grupo. Como benefício evita associação de termos a outros que gerem respostas confusas quando feito uma busca.

Em seu artigo sobre tesauro, Sabbah (2014) afirma que uma grande qualidade do tesauro é a disponibilidade para domínios específicos e algumas associações podem ter alcance e tamanho dentro de um mesmo domínio. Também afirma que pode ser reciclável e substituível, desta forma as associações dentro dele podem ser substituídas e também podem ser utilizadas para criar um outro tesauro.

2.1. Tesauro Conceitual

O tesauro conceitual surgiu devido a necessidade de associar mais de uma palavra para obter uma informação completa, porque somente uma não é o suficiente Lancaster (1986) e Gomes e Campos (2004). A alternativa encontrada foi uma expressão verbal para associação.

Medeiros (2012) afirma que os conceitos são ligados através das relações e cada um verifica quais as semelhanças com outros para compartilhar informações. Quanto

mais um conceito tiver suas características bem construídas, será possível associação com outros termos e dentro do tesauro este fato se caracteriza como importante, pois quando solicitado informações as respostas trarão mais resultados.
O uso do tesauro conceitual demonstra ser importante porque não vai considera somente a semelhança escrita ou de significado do termo, mas também o contexto em que ele se encontra [Medeiros, 2012]. Campos et al. (2009) afirma que como existem vários vocabulários eles devem ser sobrepostos para criar uma maior associação de significado.

Este será o modelo explorado neste trabalho.

3. Indexação

A indexação é um recurso utilizado dentro do Oracle para que as consultas sejam mais eficientes. Para um index ser construído é necessário estudar e analisar as estruturas das tabelas para que se obtenham resultados satisfatórios. Se forem feitas muitas inserções de dados na tabela, o Index é recalculado e o tempo que isso demora não compensa seu uso porque a consulta direta sem esse recurso seria mais rápida.

Conforme a funcionalidade os index podem ser classificados em: simples, compostos, Clustered e Non-Clustered. Os Clustered determinam qual será a ordem física do armazenamento das linhas na tabela e os Non-Clustered apontam para a linha que s encontram o objeto sem precisar importar com a ordem de armazenamento. Em relação aos Clustered quando uma chave primaria na tabela, ela será o Clustered index, caso contrário será associado o primeiro atributo está associada ao item que esteja em uma coluna não nula; as secundárias estarão ligadas a chave primária.
Os simples e os compostos partem do princípio da criação de chaves simples em compostas em tabelas. O simples possui uma coluna e o composto possui mais que uma coluna ou view [Nogare 2007].

Niemiec (1999), concluiu em seu estudo que vinte por cento dos problemas de desempenho em um sistema está ligado ao projeto e indexação do banco de dados.
Gomes & Marcondes (2003) explicam que o motivo do índice acelerar a consulta é o fato de ser um subconjunto de dados e por ser de tamanho menor facilita a consulta. Quando a informação é solicitada, o acesso aos dados ocorre pela chave de pesquisa associada. Essas chaves possuem valores e na localização é feita a comparação com a chave de menor valor dentro do índice para informar onde o registro será encontrado dentro do Banco de Dados.

Deve ser levado em consideração o tipo do usuário e os objetivos do sistema para que seja feita a escolha de uma linguagem de indexação adequada que permitirá uma eficácia melhor na consulta de dados [Vanti et al. 2011].

Wives (2000) define indexação como identificar características do assunto e armazena-las no índice, dessa forma em uma consulta o acesso da resposta é mais rápido. Qualquer informação sobre o item indexado deve ser registrada para poder ser utilizadas em demais consultas.

Na forma de construção do processo de indexação existe a forma manual e a automática. A automática baseia–se na associação de palavras significativas ou relevantes e associar a um termo que permita sua localização [Robredo, 1982]. De acordo com Vieira (1988), é feita através de programas computacionais que selecionam as palavras significativas que se associam ao item que será registrado no index.

Já na forma manual as palavras são associadas pela ação humana. Os problemas que podem ocorrer são a criação de relações falsas entre os termos, insuficiência e

excesso de termos indexados, hierarquia das ligações de atributos da forma incorreta e a falta de interação entre sistema e usuário.

4. Oracle Text

Oracle Text é uma ferramenta presente nas versões EE, SE, e XE do sistema de banco de dados Oracle. Envolve a linguagem natural de processamento, a procura livre de texto, a clusterização e classificação das palavras [Brüning, 2008].

Segundo Coulam (2009), o Oracle Text é uma ferramenta que é ignorada por ser grande e complexa e tem sua importância baseada na capacidade de pesquisar com eficiência bibliotecas com carga textual considerável permitindo o uso do BD por qualquer pessoa.

A principal funcionalidade desta ferramenta é que ela permite trabalhar com dados no formato VARCHAR2, CLOB ou BLOB, BFILE binários. Desta forma podem ser trabalhados colunas e arquivos com palavras pequenas ou com grandes textos pois, consegue fazer a indexação deles e armazenar em colunas específicas, fato que não ocorreria com o uso de SQLs na criação do índice.

A linguagem suportada, também chamada de Lexer no Oracle Text basicamente são as de origem europeia, as bases oferecidas pela Oracle são nos idiomas Inglês e Francês, mas podem ser aplicas as asiáticas, arábicas e outras fornecidas por um próprio dicionário de sinônimos criado em outro idioma. Assim é possível uma associação de palavras de origens idiomáticas diferentes.

Dentro de um sistema de BD com vários usuários, são incluídas e solicitadas informações de maneiras diferentes e Oracle Text apresenta uma solução para essa situação: associação com um tesauro.

O uso tesauro ele implicará na hierarquização do conteúdo e a definição dos sinônimos. A vantagem do uso do tesauro é que mesmo depois de haver a indexação novos termos podem ser associados. Outra opção é a aplicação dele sobre o Index.

Para Brüning (2008), o conteúdo indexado não é transacional, mas pode ser atualizado e quando entram novas informações dentro do sistema de dados do Oracle Text, pode ser criada uma fila de indexação que servirão para posterior atualização e sincronização do Index.

Os dados são estáticos e outra opção que pode ser adotada é que assim que forem feitas inserções o índice ser refeito através da sincronia com o sistema apresentando resultados precisos. Para essa opção ocorrer o Index precisa ser fragmentado na região que entrará a nova informação e depois remontado sem apagar dados antigos. Também pode ser feita a construção desde o início, mas o tempo gasto não é conveniente [Coulam, 2009].

A otimização do Index do Oracle Text permite um armazenamento de dados mais rápido e eficiente dentro do banco de dados. A localização de dados traz mais respostas e mesmo que ocorra uma exclusão as informações continuam registradas no caso de consulta, sairão somente na próxima atualização.

Os usos de letras maiúscula e minúsculas irão determinar a hierarquia das palavras e sentenças dentro do Oracle Text, mas no momento de consultas não ocorre essa diferenciação, nesse caso o uso do dicionário de sinônimos é importante. O tesauro pode ser criado a partir de uma enciclopédia que pode ser fornecida carregada no sistema gerando uma ligação ao dicionário de sinônimos podendo ser alterados por seus usuários.

Quando há mistura de tipos de letras elas são registradas como entraram e o dicionário de sinônimos e o index necessitam ter essa diferenciação. No caso do uso

tesauro essa situação, também chamada de case–sensitive, ao se realizarem as consultas, trazem mais respostas do que o dicionário de sinônimos que armazena todas as informações em maiúsculas. Uma consequência que ocorre no tesauro dentro do Oracle Text é se ele não diferenciar as letras e também converter todas para maiúsculas haverá comprometimento de resultados.

O Oracle Text, mesmo que seja extenso, tem suas funcionalidades e sabendo utiliza-las da forma correta trará benefícios na localização de arquivos e textos que estão guardados em algum local dentro do banco mesmo desconhecendo a forma que foram escritos ou armazenados. Coulam (2009) compara essa ferramenta à caixa de Pandora que muitos não sabem o que tem dentro e na curiosidade de saber o conteúdo descobrem suas características aprendendo a lidar com elas.

5. Materiais e Métodos

5.1. Configuração do Ambiente de Trabalho

Para o desenvolvimento desse trabalho a foi utilizado o programa Oracle Virtual Box para criar uma máquina virtual com as seguintes configurações e programas instalados:

- Memória principal: 2048MB.
- Capacidade do disco: 30GB.
- Sistema Operacional: Microsoft Windows 2008 Server 64 bits.
- Microsoft Excel 2010.
- Banco de Dados Oracle 11g Standard Edition.
- SQL Developer.

A Figura 1 a seguir apresenta um esquema com a forma de configuração do ambiente de trabalho já descrito:

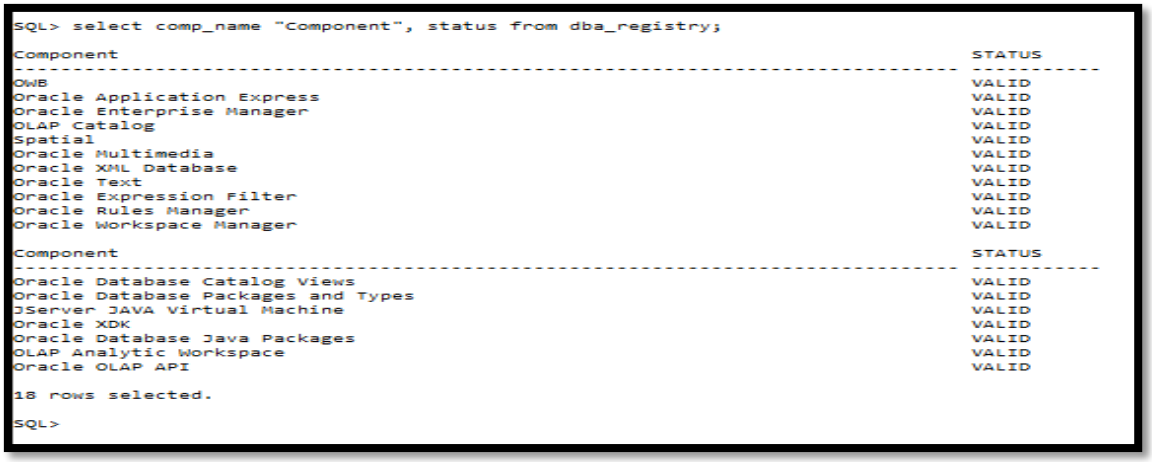
Figura 1. Ambiente de trabalho. Fonte: Elaborado pelos autores

5.2. Preparação do Oracle para a execução dos Experimentos

Após a instalação do Oracle estiver completa, com o usuário SYSDBA será consultado no sistema se o Oracle Text está instalado através do seguinte comando:

```
SELECT COMP_NAME "COMPONENT", STATUS FROM DBA_REGISTRY;
```

A Figura 2 demonstra a resposta do comando acima, deve ser observado que sobre o Oracle Text deve conter VALID em sua linha.



```
SQL> select comp_name "Component", status from dba_registry;

Component                                                                       STATUS
------------------------------------------------------------------------------- -----------
OWB                                                                             VALID
Oracle Application Express                                                      VALID
Oracle Enterprise Manager                                                       VALID
OLAP Catalog                                                                    VALID
Spatial                                                                         VALID
Oracle Multimedia                                                               VALID
Oracle XML Database                                                             VALID
Oracle Text                                                                     VALID
Oracle Expression Filter                                                        VALID
Oracle Rules Manager                                                            VALID
Oracle Workspace Manager                                                        VALID

Component                                                                       STATUS
------------------------------------------------------------------------------- -----------
Oracle Database Catalog Views                                                   VALID
Oracle Database Packages and Types                                              VALID
JServer JAVA Virtual Machine                                                    VALID
Oracle XDK                                                                      VALID
Oracle Database Java Packages                                                   VALID
OLAP Analytic Workspace                                                         VALID
Oracle OLAP API                                                                 VALID

18 rows selected.

SQL>
```

Figura 2. Exibição de componentes do Oracle. Fonte: Elaborado pelos autores

Após verificar que o Oracle Text está instalado, é feita a criação do usuário “tcc” com a senha “tcc” utilizando o comando para este procedimento, conforme as configurações apresentadas na da Tabela 1:

Tabela 1. Configurações e Comando para Criação de Usuário

Nome do Usuário	**Senha**
tcc	tcc
Função	**Comando**
Criar usuário	CREATE USER tcc IDENTIFIED BY tcc;

Para o usuário tcc serão atribuídas as seguintes permissões/atribuições que permitam o trabalho no Oracle via SQL PLUS ou SQL DEVELOPER de acordo com a Tabela2:

Tabela 2. Atribuições Permissões ao usuário “tcc”

Função/Permitir	**Comando**
Criar trigger	GRANT CREATE TRIGGER TO tcc ;
Criar tablespace	GRANT CREATE TABLESPACE TO tcc ;
Criar sequência	GRANT CREATE SEQUENCE TO tcc ;
Criar tabela	GRANT CREATE TABLE TO tcc ;
Criar procedure	GRANT CREATE PROCEDURE TO tcc ;
Criar sinônimo	GRANT CREATE SYNONYM TO tcc ;
Criar view	GRANT CREATE VIEW TO tcc ;
Criar type	GRANT CREATE TYPE TO tcc ;
Criar sessão	GRANT CREATE SESSION TO tcc ;
Tablespace ilimitado	GRANT UNLIMITED TABLESPACE TO tcc ;
Criar conexao	GRANT CREATE DATABASE LINK TO tcc ;
Alterar sessão	GRANT ALTER SESSION TO tcc ;
Permitir recurso	GRANT RESOURCE TO tcc;

Em relação ao Oracle Text, permissões específicas apresentadas na Tabela 3, devem ser concedidas ao usuário via SQLPLUS ou SQL DEVELOPER:

Tabela 3. Atribuições Permissões ao usuário "tcc" para uso do Oracle Text

Função/permitir	Comando
Uso do Oracle Text	GRANT "CTXAPP" TO tcc;
Ler biblioteca CTX_CLS	GRANT EXECUTE ON CTXSYS.CTX_CLS TO tcc;
Ler biblioteca CTX_DDL	GRANT EXECUTE ON CTXSYS.CTX_DDL TO tcc;
Ler biblioteca CTX_DOC	GRANT EXECUTE ON CTXSYS.CTX_DOC TO tcc;
Ler biblioteca CTX_OUTPUT	GRANT EXECUTE ON CTXSYS.CTX_OUTPUT TO tcc;
Ler biblioteca CTX_QUERY	GRANT EXECUTE ON CTXSYS.CTX_QUERY TO tcc;
Ler biblioteca CTX_REPORT	GRANT EXECUTE ON CTXSYS.CTX_REPORT TO tcc;

Especificamente via SQL DEVELOPER deve ser criada uma conexão TCC como usuário tcc, pelo motivo que a tabela que será utilizada no trabalho necessita de inserções e verificações atentas.

5.3. Criação da Tabela de Trabalho

A tabela criada para os experimentos tem o nome de "*MOVIES"*, pois os dados a serem explorados nos experimentos são somente sobre filmes e registrados na língua inglesa. Foram escolhidos 400 filmes diferentes que tem informações dispostas nas seguintes colunas, como demonstrado na Tabela 4.

Tabela 4. Estrutura da tabela MOVIES. Fonte: Elaborado pelo autor

Coluna	Tipo	Característica
ORIGINAL_TITLE	VARCHAR2(100 BYTE)	NOT NULL
PORTUGUESE_NAME	VARCHAR2(100 BYTE)	NULL
SYNOPSIS	VARCHAR2(500 BYTE)	NOT NULL
YEAR	NUMBER (4,0)	NULL
PLOT	CLOB	NOT NULL
GENRE	VARCHAR2(100 BYTE)	NULL
ID	NUMBER (3,0)	PRIMARY KEY

As colunas que serão trabalhadas para os testes são:

- ORIGINAL_TITLE: do tipo VARCHAR2, contém o nome original do filme, preferencialmente na língua inglesa.
- SYNOPSIS: do tipo VARCHAR2, no mecanismo de busca dentro do site www.imdb.com é feita a localização do filme e a sinopse escrita em Inglês é copiada e inserida no campo respectivo da tabela.
- PLOT: do tipo CLOB por conter texto extenso, contém o resumo do filme encontrado no Wikipédia em Inglês, da mesma forma que o campo SYNOPSIS, o conteúdo respectivo é copiado do site e inserido nos campos respectivos da tabela.

5.4. Indexação das Colunas

Como este trabalho visa verificar a eficiência e velocidade de consultas e serão efetuadas várias seleções buscou–se a indexação das colunas para facilitar as pesquisas. Durante o processo de indexação, para ser usada a consulta tipo *Like*, foi observado que não é possível indexar colunas tipo CLOB. Os índices referentes as outras colunas foram excluídos com a intenção de manter uma igualdade na consulta dentro das colunas.

Em relação ao uso do Oracle Text, para pesquisas usando essa ferramenta a indexação é obrigatória em cada coluna nas quais que serão efetuadas pesquisas e desta

vez é possível indexar colunas tipo CLOB. Foram realizados os seguintes comandos para indexação pelo Oracle Text de acordo Tabela 5:

Tabela 5. Criação de Índices pelo Oracle Text

Coluna Indexada	Nome do Índice	Comando
ORIGINAL_TITLE	IDX_TITLE	CREATE INDEX IDX_TITLE ON MOVIES (ORIGINAL TITLE) INDEXTYPE IS CTXSYS.CONTEXT;
SYNOPIS	IDX_SYNOPSIS	CREATE INDEX IDX_SYNOPSIS ON MOVIES (SYNOPIS) INDEXTYPE IS CTXSYS.CONTEXT;
PLOT	IDX_PLOT	CREATE INDEX IDX_PLOT ON MOVIES (PLOT) INDEXTYPE IS CTXSYS.CONTEXT;

É importante conhecer o funcionamento do Oracle Text pois ele não permite criar um índice que envolva mais de uma coluna, desta forma as consultas realizadas são específicas para cada coluna.

5.5. Criação do Tesauro

É possível inserir um tesauro já pronto dento do Oracle para a realização de consultas ou então criar um próprio. A opção escolhida foi criar um tesauro manualmente devido a tabela de trabalho ser temática e ter 400 inserções como já citado. Primeiramente através de um procedimento cria-se o arquivo do tesauro que no caso o nome escolhido foi "*films*":

```
BEGIN
CTX_THES.CREATE_THESAURUS ('FILMS');
END;
/
```

Foram escolhidas dez palavras para serem realizadas consultas nos experimentos, então foram escolhidos vários sinônimos para cada uma delas. Como exemplo a palavra "*Pirate*" que foi associada a outras que também remetessem a ela como: *sea*, *privateer*, *Peter Pan*, *treasure*, *island*, *hook*. O mesmo foi feito com a palavra "*Dcheroes*" que é uma palavra inexistente, mas que teve associações com nomes de heróis para verificar se o tesauro funcionará corretamente.

A Tabela 6 apresenta todas as palavras escolhidas para serem utilizadas nas consultas e sinônimos atribuídos que durante os testes deverão também ser apresentados nas respostas:

Tabela 6. Palavras-Teste e Sinônimos. Fonte: Elaborado pelos autores

Palavra-teste	Sinônimos
Witch	Witch, Harry Potter, fairy, magician, wicked, wand, wish, Oz, Merlin, Morgaine
King	Arthur, crown, sword, kingdom, emperor, queen, prince, princes, empire
Pirate	Pirates, sea, privateer, Peter Pan, treasure, island, hook
Vampire	Vamps, vamp, blood, wolfman, twilight, underworld, Helsing, Breakdown
Marvel	X-men', x2, Iron Man, Captain America, Hulk, Thor, Fantastic Four, Guardians of Galaxy, Avengers
Religion	Bible, Exodus, Jesus, Moses, Noah, cross, pope, god, gods, temptation, mythology
Disney	Aladdin, Snow White, Tarzan, Peter Pan, Maleficent, dalmatians, 'beauty and the beast, Pixar, Lion King
Dcheroes	Justice league, Batman, Wonderwoman, Green Lantern, man of steel, Superman, Flash
Terror	Fear, mummy, death, wolfman, monster, devil, screen, ghost
Space	Earth, sun, star, gravity, alien, planet, Starship, galaxy, moon, universe, comet

Para associar uma palavra a outra dentro de um tesauro no Oracle Text, também é necessário o uso de um procedimento para inserção que pode conter uma ou várias como no exemplo a seguir:

```
BEGIN
CTX_THES.CREATE_RELATION ('FILMS', 'SPACE', 'SYN', 'STARSHIP');
CTX_THES.CREATE_RELATION ('FILMS', 'KING', 'SYN', 'EMPEROR');
END;
/
```

A ordem do conteúdo dos parênteses é a seguinte escrevendo cada palavra por aspas: nome do tesauro, palavra-chave, *syn* para sinônimo e palavra associada.
Caso seja necessário remover uma associação é necessário um procedimento como no exemplo:

```
BEGIN
CTX_THES.DROP_RELATION ('FILMS', 'DISNEY', 'SYN', 'PETER PAN');
END;
/
```

5.6. Execução dos Experimentos

Nos experimentos foram realizadas seleções tipo *Like*, Oracle Text e Oracle Text nas colunas de trabalho, utilizando cada palavra-teste citada na Tabela 6, apresentada anteriormente. Junto com cada seleção também foi solicitado o plano de execução. Esse que esse procedimento se repetiu por dez vezes em instancias diferentes. Foram contabilizadas cem consultas para cada método de seleção. Juntamente com cada consulta foi solicitado o tempo de reposta e plano de execução (para se obter o Custo). No intervalo de cada consulta ocorreu o encerramento da instancia e depois uma nova instancia.

- **Testes tipo *Like*:** Utilizando cada uma das palavra-teste pelo comando *Like* é necessário evitar a diferenciação de letras maiúscula e minúsculas na primeira letra da palavra, pois este tipo de consulta diferencia os tipos de letras. Sendo assim foram executados comandos como no exemplo:

  ```
  SELECT ID, ORIGINAL_TITLE, PORTUGUESE_NAME FROM MOVIES WHERE
  (ORIGINAL_TITLE LIKE '%King%' OR ORIGINAL_TITLE LIKE '%king%')

  EXPLAIN PLAN FOR SELECT ID, ORIGINAL_TITLE, PORTUGUESE_NAME FROM MOVIES
  WHERE (ORIGINAL_TITLE LIKE '%KING%' OR ORIGINAL_TITLE LIKE '%KING%')

  SELECT * FROM TABLE(DBMS_XPLAN.DISPLAY);
  ```

  Dentro dos parênteses da primeira seleção contém a coluna consultada depois palavra-teste com a primeira letra em minúscula e depois palavra-teste com primeira letra em maiúsculo.

- **Teste com Oracle Text**: A consulta que envolve o Oracle Text envolve uma relação com a palavra "*contains*". Segue a seguinte estrutura como no exemplo:

  ```
  SELECT ID, ORIGINAL_TITLE, PORTUGUESE_NAME FROM MOVIES WHERE CONTAINS
  (SYNOPSIS,'WITCH', 1) >0;

  EXPLAIN PLAN FOR SELECT ID, ORIGINAL_TITLE, PORTUGUESE_NAME FROM MOVIES
  WHERE CONTAINS (SYNOPSIS,'WITCH', 1) >0;

  SELECT * FROM TABLE(DBMS_XPLAN.DISPLAY);
  ```

Dentro dos parênteses da primeira seleção deve ser preenchido o nome da coluna e depois a palavra-teste entre aspas simples.

- **Teste com Oracle Text associado ao tesauro criado:** Esse tipo de consulta é semelhante ao anterior e se diferencia no conteúdo dos parênteses onde primeiro aparece a coluna de consulta, depois uma aspa simples a palavra "*syn*" que abre novos parênteses contendo palavra-teste e nome do tesauro, depois deste parêntese uma aspa simples para encerramento. Tem a seguinte estrutura exemplificada:

```
SELECT ID, ORIGINAL_TITLE, PORTUGUESE_NAME FROM MOVIES WHERE
CONTAINS (SYNOPSIS, 'SYN (VAMPIRE, FILMS)') > 0;

EXPLAIN PLAN FOR SELECT ID, ORIGINAL_TITLE, PORTUGUESE_NAME FROM
MOVIES WHERE CONTAINS (SYNOPSIS, 'SYN (VAMPIRE, FILMS)') > 0;

SELECT * FROM TABLE(DBMS_XPLAN.DISPLAY);
```

6. Resultados

Em cada consulta executada foram anotados, o número de respostas obtidas, o tempo que foi necessário para exibir a resposta e o custo da CPU para o procedimento. Após todos esses dados serem registrados, foram feitas médias aritméticas para cada método de consulta e foram obtidos e representados pelos seguintes gráficos.

Na Figura 3, é apresentado o gráfico com as médias de custo do CPU em porcentagem. É possível verificar que o testes com Oracle Text gera um custo muito menor, aplicado sozinho apresenta médias 4,47% e com o tesauro 6,6%, que é praticamente dez vezes menor em relação ao teste com o comando Like que apresentou 68,7%. Este fato pode ser explicado pela diferença de processamento e uso de índices que são necessários para a execução do Oracle Text.

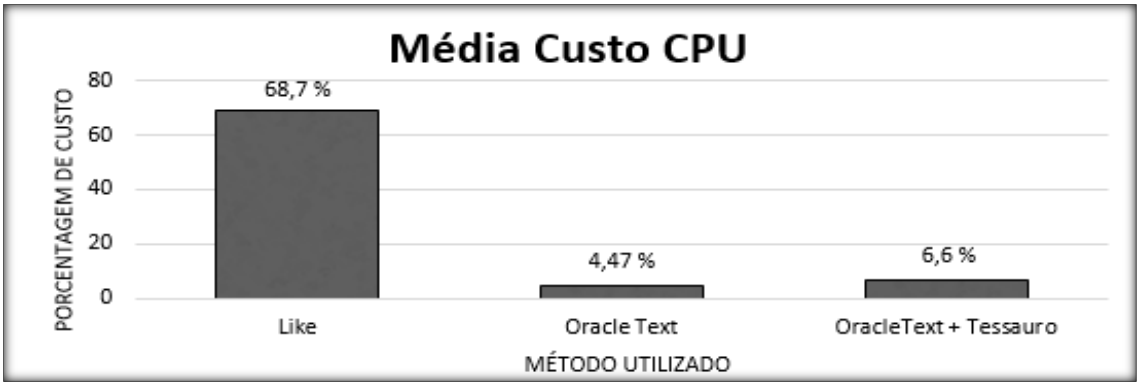


Figura 3. Média do Custo CPU. Fonte: Elaborado pelos autores

A Figura 4 apresenta um gráfico contendo duas informações obtidas por cada método de seleção. A primeira é a média de quantidade de respostas trazidas (barras claras) onde se observado que o uso do Oracle Text associado ao tesauro traz a média de 50,2 respostas, o Oracle Text sozinho traz a média de 37,4 respostas e *Like* traz a média de 19,3 respostas. Comparando os testes o uso do tesauro traz mais retorno nas consultas. A segunda é o tempo de resposta (barras escuras) observa-se que o tempo do Oracle Text junto ao tesauro é de 104 milésimos de segundo (ms) o que é maior em relação ao uso do Oracle Text sem associações que apresenta 43 ms. Isso pode ser

explicado pela quantidade de linhas lidas da tabela durante a consulta. Já no caso do uso do *Like* o tempo de resposta é 237 ms, o que é muito maior, onde também pode ser explicado pela ausência do uso de índices. É possível verificar que os testes usando *Like* tiveram os piores resultados quanto em tempo e quantidade de resposta.

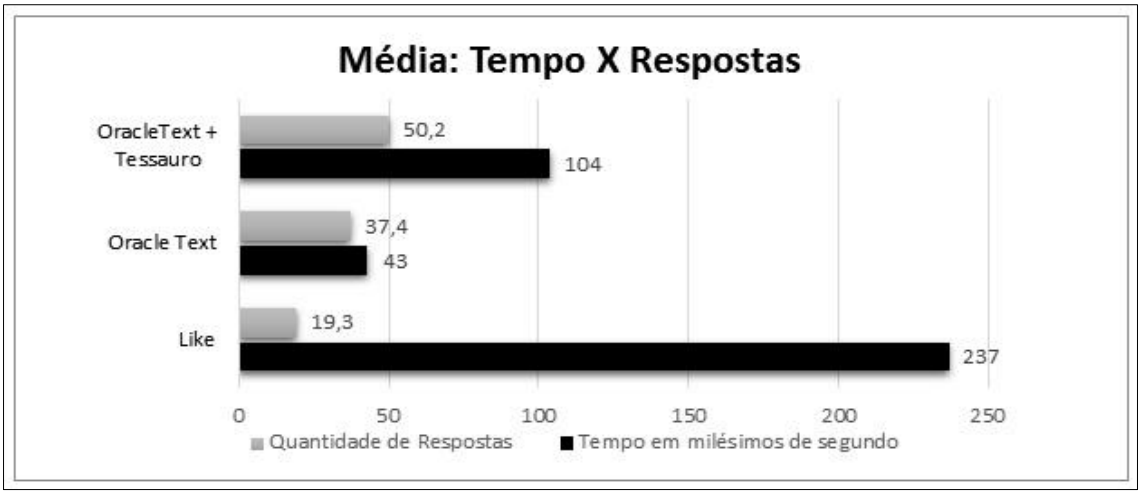


Figura 4. Média do tempo de processamento e quantidade de respostas obtidas em cada teste. Fonte: Elaborado pelos autores

A Tabela 7 apresenta exemplos, usando as palavras-teste "pirate" e "witch", para demonstrar que em alguns casos o uso do tesauro aumentou a quantidade de respostas. Para a palavra "*Pirate*", a consulta *Like* trouxe 15 resultados que diminuíram no uso do Oracle Text sozinho para 7 resultados e ao se usar o tesauro aumentou para 96 resultados. O mesmo tipo de comportamento ocorreu com a palavra "*Witch*".

Tabela 7. Exemplo de Respostas em Coluna CLOB

Palavra - Teste	*Like*	Oracle Text	O.T. + Tesauro
Pirate	15	7	96
Witch	47	32	73

7. Conclusões

Sobre o uso do método *Like* pode ser observado que além de não trazer tantas respostas demora muito mais que os outros métodos e gera um custo alto de processamento do banco de dados em relação aos outros métodos. O uso do Oracle Text associado ou não com tesauro gera um custo muito menor, fato que pode ser explicado devido a indexação realizada.

No caso, do uso do tesauro há um pequeno aumento de custo que decorrente da quantidade de novos termos que são pesquisados e quantidade de respostas trazidas. Pela análise das respostas das consultas durante os testes foi possível verificar que se for utilizado uma associação de um mesmo termo para palavras diferentes será exibido resultado para ambas.

Com uma correta associação de palavras com o uso do tesauro aliado ao recurso do Oracle Text é possível obter melhores resultados de uma consulta com menor custo de processamento do que os métodos tradicionais. E em muitos casos obtendo uma maior qualidade no retorno das consultas, que pelos métodos tradicionais omitem resultados, pois procuram apenas pelos termos exatos e não pelas suas correlações.

8. Referências

Bocorny, A. E. P.; Villavicencio A.; Kilian, C. K.; Wilkens, R. (2010) "A construção de um glossário bilíngue (inglês/português) multimeios online colaborativo para aprendizes baseado em corpus especializado da área de relações internacionais". Trama, v. 6, n. 12, p. 09 – 25.

Brüning, A. (2008) "Oracle Text 11g". https://www.informatik.hu-berlin.de/de/forschung/gebiete/wbi/teaching/archive/sose08/hk_text/oracle_text.pdf, May.

Campos, M. L. A.; Campos, M. L. M.; Gomes, H. E.; Campos, L. M.; Martins, A.E.; Sales, L. F. (2006) "Estudo comparativo de softwares de construção de tesauros". Perspectivas em Ciência da Informação, Belo Horizonte, v.11 n.1, p. 68-81, jan./abr..

Campos, M. L. A.; Gomes, H. E. (2006) "Metodologia de elaboração de tesauro conceitual: a categorização como princípio norteador". Perspectivas em Ciência da Informação, Belo Horizonte, v. 11, n. 3, p.348-359, set./dez.

Campos, M. L. A.; Dávila, A. M. R.; Gomes, H. E.; Campos, L. M.; Lira, L. M. (2009) "Aspectos Metodológicos no Reuso de Ontologias: um estudo a partir das anotações genômicas no domínio dos tripanosomatídeos". RECIIS - Revista Eletrônica de Comunicação, Informação e Inovação em Saúde, v. 3, p. 64-75.

Coulan, B. (2009) "Unlocking Hidden gems in Oracle Text". RMOUG: Training Days.

Finatto, M. J. B. (2004) "Introdução à terminologia: teoria e prática". São Paulo: Contexto.

Gomes, R. R., & Marcondes, M. R. (2003) "Estudo Comparativo de Técnicas de Inde-Indexação no Banco de Dados Oracle". http://www.esfcex.ensino.eb.br/revista/producaocientifica/arquivo/204_Artigo.pdf, Mai.

Gomes, H. E.; Campos, M. L. A.; Motta, D. F. (2004) "Elaboração do tesauro documentário: tutorial". http://conexaorio.com/biti/tesauro,Mar.

Krieger, M. G. (2003) "Dicionário de língua: um instrumento didático pouco explorado". In: Toldo, C. S. (Org.). Questões de Linguística. Passo Fundo: UPF Editora, p.70-87.

Lancaster, F. W. (1986) "Vocabulary control for information retrieval". 2. ed. Arlington: IRP.

Lancaster, F. W. (2004) "Indexação e resumos: teoria e prática". 2.ed. Brasília: Briquet de Lemos.

Lu, C.; Lee, K. H. ; Chen, H. Y.(1995) "TheSys-a comprehensive thesaurus system for intelligent document analysis and text retrieval Document Analysis and Recognition",p. 1169 – 1173, vol.2.

Medeiros, J. S. (2012) "Tesauros conceituais e ontologias de fundamentação: abordagem comparativa entre modelos conceituais". São Paulo: Ixtlan.

Niemiec, R. J. (1999) "Oracle Performance Tuning Tips & Techniques". Berkeley. McGraw-Hill

Nogare, D. (2007) “Melhorando desempenho de consultas utilizando Views Indexadas”. http://www.linhadecodigo.com.br/artigo/1308/melhorando-desempenho-de-consultas -utilizando-views-indexadas.aspx, May.

Robredo, J. (1982) “A indexação automática de textos: o presente já entrou no futuro”. Machado, U. O. , ed. Estudos Avançados em Biblioteconomia e Ciência da Informação. Brasília, ABDF, v. 1, p. 236-74.

Roget, P. M. (1925) “Thesaurus of English words and phrases”. New York: Longmans.

Sabbah, T.; Ashraf, M.; Herawan, T. (2014) “Effect of thesaurus size on schema matching quality”. Knowledge-Based Systems, p. 211–226.

UNESCO. (1973) “Guidelines for the establishment and development of monolingual thesauri”.

Vanti, N. et al. (2011) “Linguagens de indexação: uso das linguagens presentes na pratica da indexação”.In: Encontro regional de estudantes de biblioteconomia, documentação, ciência da informação e gestão da informação, Maranhão. Maranhão: EREBD. http://repositorio.ufrn.br:8080/jspui/handle/1/6176, May.

Vieira, S. B. (1988) “Indexação automática e manual: revisão de literatura”. Ciência da In - formação, Brasília, v.17. n.1, p.47-57.

Vieira, J.M.L. et al. (2010) “Estudo da construção e aplicação do tesauro na recuperação da informação de teses e dissertações do programa de pós-graduação em desenvolvimento urbano”. Biblionline, João Pessoa, n. esp., p. 71-80.

Wives, L. (2002) “Tecnologias de Descoberta de Conhecimento em Textos aplicadas à Inteligência Competitiva”. Porto Alegre, 2002. 100 f. Pós-Graduação em Computação. Universidade Federal do Rio Grande do Sul, Porto Alegre.