Talita dos Reis Lopes Berbel

Recomendação Semântica de Documentos de Texto Mediante a Personalização de Agregações OLAP

Sorocaba, SP

23 de Março de 2015

Talita dos Reis Lopes Berbel

Recomendação Semântica de Documentos de Texto Mediante a Personalização de Agregações OLAP

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação (PPGCCS) da Universidade Federal de São Carlos como parte dos requisitos exigidos para a obtenção do título de Mestre em Ciência da Computação. Área de concentração: Engenharia de Software e Gestão do Conhecimento: Banco de Dados.

Universidade Federal de São Carlos – UFSCar

Centro de Ciências em Gestão e Tecnologia – CCGT

Programa de Pós-Graduação em Ciência da Computação – PPGCCS

Orientador: Sahudy Montenegro González

Sorocaba, SP

23 de Março de 2015

Berbel, Talita dos Reis Lopes.
B484r Recomendação semântica de documentos de texto mediante a personalização de agregações OLAP. / Talita dos Reis Lopes Berbel. – – 2015.
114 f. : 28 cm.

Dissertação (mestrado)-Universidade Federal de São Carlos, *Campus* Sorocaba, Sorocaba, 2015
Orientador: Sahudy Montenegro González
Banca examinadora: Marcio Katsumi Oikawa, Tiemi Christine Sakata
Bibliografia

1. Tecnologia OLAP. 2. Ontologia. 3. Semântica. I. Título. II. Sorocaba-Universidade Federal de São Carlos.

CDD 006.301

Ficha catalográfica elaborada pela Biblioteca *Campus* Sorocaba.

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências em Gestão e Tecnologia
Programa de Pós-Graduação em Ciência da Computação

Folha de Aprovação

Assinaturas dos membros da comissão examinadora que avaliou e aprovou a Defesa de Dissertação de Mestrado da candidata Talita dos Reis Lopes Berbel, realizada em 23/03/2015:

Profa. Dra. Sahudy Montenegro González
UFSCar

Prof. Dr. Marcio Katsumi Oikawa
UFABC

Profa. Dra. Tiemi Christine Sakata
UFSCar

Ao meu querido marido Bruno e aos meus pais que contribuíram com seu carinho, afeto e paciência ao longo dos anos em que tenho dedicado aos meus estudos.

Agradecimentos

A Deus, por me ter proporcionar vivência, amadurecimento e educação durante toda a minha vida.

Ao meu marido, Bruno Berbel, aos meus pais, minha irmã e minha sobrinha pelo amor, carinho e apoio a cada novo desafio.

Às amizades cultivadas durante o período do curso, em especial aos colegas de trabalho: Angelina Melaré, Jane Piantoni, Ricardo Leme e Joaquim Machado por sua compreensão, cooperação e pelos grandes momentos de superação e alegria que vivemos juntos.

Aos professores da UFSCar, por seus ensinamentos, dedicação e constante instrução. Aos professores especialistas nas áreas biomédicas que também contribuíram na avaliação dos resultados do trabalho.

À minha orientadora, em especial, pelos seus ensinamentos, seu apoio e seu talento aplicados neste trabalho.

À UFSCar pela sua estrutura e pelo apoio no desenvolvimento deste projeto.

Por fim agradeço a todos que, de alguma forma, contribuíram para que esse trabalho fosse realizado. A todos vocês, a minha sincera gratidão.

"A mente que se abre a uma nova ideia
jamais volta ao seu tamanho original".
(Albert Einstein)

Resumo

Com o crescimento do volume dos dados não estruturados, como os documentos de texto, torna-se cada vez mais interessante e necessário extrair informações deste tipo de dado para dar suporte à tomada de decisão em sistemas de *Business Intelligence*. Recomendações podem ser utilizadas no processo OLAP, pois permitem que os usuários tenham uma experiência diferenciada na exploração dos dados. O processo de recomendação, aliado à possibilidade da personalização das consultas dos usuários, tomadores de decisão, permite que as recomendações possam ser cada vez mais relevantes. A principal contribuição deste trabalho é a proposta de uma solução eficaz para a recomendação semântica de documentos mediante a personalização de consultas de agregação OLAP em um ambiente de *Data Warehousing*. Com o intuito de agregar e recomendar documentos propõe-se a utilização da similaridade semântica. A ontologia de domínio e a medida estatística de frequência são utilizadas com o objetivo de verificar a similaridade entre os documentos. O limiar de similaridade entre os documentos no processo de recomendação pode ser parametrizado e é esta a personalização que oferece ao usuário uma maneira interativa de melhorar a relevância dos resultados obtidos. O estudo de caso proposto se baseia em artigos da PubMed e em sua ontologia de domínio com o propósito de criar um protótipo utilizando dados reais. Os resultados dos experimentos realizados são expostos e analisados, mostrando que boas recomendações e agregações são possíveis utilizando a abordagem sugerida. Os resultados são discutidos com base nas métricas de avaliação: *precision*, *recall* e *F1-measure*.

Palavras-chave: *Data Warehouse*. OLAP. Dados textuais. Agregação. Recomendação. Semântica. Ontologia. LCA. Personalização de consultas. MeSH.

Abstract

With the rapid growth of unstructured data, such as text documents, it becomes more and more interesting and necessary to extract such information to support decision making in business intelligence systems. Recommendations can be used in the OLAP process, because they allow users to have a particular experience in exploiting data. The process of recommendation, together with the possibility of query personalisation, allows recommendations to be increasingly relevant. The main contribution of this work is to propose an effective solution for semantic recommendation of documents through personalisation of OLAP aggregation queries in a data warehousing environment. In order to aggregate and recommend documents, we propose the use of semantic similarity. Domain ontology and the statistical measure of frequency are used in order to verify the similarity between documents. The threshold of similarity between documents in the recommendation process is adjustable and this is the personalisation that provides to the user an interactive way to improve the relevance of the results. The proposed case study is based on articles from PubMed and its domain ontology in order to create a prototype using real data. The results of the experiments are presented and discussed, showing that good recommendations and aggregations are possible with the suggested approach. The results are discussed on the basis of evaluation measures: precision, recall and F1-measure.

Key-words: Data warehouse. OLAP. Textual data. Aggregation. Recommendation. Semantic. Ontology. LCA. Query Personalization. MeSH.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

BI	Business Intelligence (Inteligência nos Negócios)
BPM	Business Performance Management (Gestão do Desempenho Empresarial)
CRM	Customer Relationship Management (Gestão de Relacionamento com o Cliente)
DSS	Decision Support Systems (Sistemas de Apoio à Decisão)
DW	Data Warehouse (Armazém de Dados)
EIS	Executive Information Systems (Sistemas de Informações Executivas)
ERP	Enterprise Resource Planning (Planejamento de Recursos Empresariais)
ETL	Extract Transform Load (Extração, Transformação e Carregamento)
LCA	Lowest Common Ancestor (Ancestral Comum de Maior Profundidade)
MeSH	Medical Subject Headings
NLM	U. S. National Library of Medicine
OLAP	On-line Analytical Processing (Processamento Analítico *On-line*)
OLTP	On-line Transaction Processing (Processamento de Transações *On-line*)
PAMDES	PubMed Article Multidimensional Recommendation System
POQT	Personalized OLAP Queries for Texts
RI	Recuperação de Informações
TM	Text Mining (Mineração de Textos)
XML	Extensible Markup Language (Linguagem de Marcação Extensível)

Sumário

1 Introdução

Nos últimos anos, o Processamento Analítico *On-line* (OLAP) aliado ao armazenamento multidimensional geraram metodologias, ferramentas e sistemas de gestão de recursos para a análise de dados convencionais. Porém, mesmo com esses avanços, não foi estabelecido um consenso sobre como incorporar os dados não estruturados em *Data Warehouses*. O volume desse tipo de dados cresce constantemente e por esta razão torna-se necessário fornecer uma análise multidimensional diferenciada ([RAVAT; TESTE; TOURNIER, 2007]).

Dados não estruturados são dados heterogêneos e possuem diferentes formatos, como ocorre em textos, documentos, imagens, arquivo de áudio e vídeos. Dados textuais são difíceis de analisar e os sistemas geralmente usam medidas numéricas para contornar a análise de textos. Os *Data Warehouses* possuem operações OLAP ideais para manipular dados numéricos, porém a análise de texto não está em sua natureza.

A análise de dados não estruturados pode revelar inter-relações importantes que anteriormente eram difíceis ou impossíveis de determinar apenas com as técnicas existentes para a análise dos dados estruturados. Os trabalhos, no entanto, que tratam sobre modelos multidimensionais que privilegiavam a integração dos dados não estruturados, não chegam a um consenso sobre uma solução de como armazenar e manipular de forma adequada este tipo de dado em um *Data Warehousing* [GONZÁLEZ; BERBEL, 2014]. Portanto, as abordagens recentes oferecem alguma forma de apoio a esses dados, mas não há uma implementação robusta que trate esses dados como nativos nas operações OLAP.

Algumas soluções para armazenamento de textos e imagens criam cubos de consultas baseados nesses dados e outras propostas introduziram a semântica nos modelos multidimensionais. Para analisar as relações semânticas entre documentos, é possível, por exemplo, utilizar a agregação multidimensional para sumarizar os documentos.

Os sistemas OLAP tradicionais oferecem multidimensionalidade e um grande volume de dados, mas segundo [JERBI et al., 2009], não se pode confiar unicamente na exploração padrão de bases multidimensionais. Utilizar recomendações torna o processo de análise mais fácil, além de ajudar os usuários a encontrarem rapidamente os dados relevantes para a tomada de decisão.

Além da recomendação, outra possibilidade de facilitar a exploração dos dados é tornar o usuário o centro da decisão e considerar suas necessidades e preferências nas consultas. Incluir a personalização na análise multidimensional é uma tendência de pesquisa exposta em [BENTAYEB; FAVRE, 2009] e [KANG; CHOI, 2011] para melhorar o processo de decisão centrado no usuário.

A recomendação e a personalização, portanto, podem se tornar aliadas no processo de exploração de uma base de documentos de texto. As propostas, sobre modelos multidimensionais e operações OLAP para dados textuais não comtemplam esses tópicos, enquanto que trabalhos sobre recomendação/personalização não dão ênfase aos dados não estruturados, ou seja, propostas que trabalhem com personalização de agregações OLAP visando a recomendação semântica de documentos de texto. Dessa forma, esse assunto se tornou o principal objeto de estudo deste trabalho.

1.1 Objetivo

O principal objetivo deste trabalho consiste em propor um modelo conceitual multidimensional e um algoritmo de agregação, que permite a agregação e recomendação personalizada de documentos de texto para consultas OLAP.

Com o modelo multidimensional e o OLAP para documentos de texto pretende-se prover respostas de maneira mais precisa e eficiente para os tomadores de decisão.

1.2 Metodologia

O primeiro passo para o desenvolvimento do trabalho foi realizar um estudo sobre os fundamentos sobre *Data Warehouse* e as operações OLAP. Na sequência foi realizada uma revisão sistemática, metodologia usada para identificar e sintetizar estudos, para explorar o estado da arte sobre os *Data Warehouses* que armazenam e utilizam dados não estruturados nos processos de tomada de decisão.

A revisão sistemática [GONZÁLEZ; BERBEL, 2014] revelou os esforços que estão sendo realizados na área e alguns trabalhos analisados propuseram novas possibilidades de operações OLAP para textos, como a operação de agregação. No entanto, esses estudos não traziam aplicações práticas na área, ou seja, eles apenas sintetizavam os conceitos. A partir da análise desses estudos, foi encontrada uma vertente que ainda não havia sido explorada até o momento, que é aliar o OLAP textual e a recomendação personalizada de documentos.

A Figura 1 mostra não apenas a estrutura da dissertação, mas também os principais pontos da metodologia desenvolvida. Após o estudo dos conceitos utilizados amplamente na proposta e a verificação do estado da arte, passou-se a trabalhar na etapa seguinte: a construção de um modelo multidimensional adequado para dados textuais.

Foi desenvolvida a proposta de um modelo multidimensional, denominado *Personalized OLAP Queries for Texts* (POQT) e algoritmos para personalizar agregações textuais (OLAP textual), a partir de um determinado documento escolhido pelo usuário final. O conjunto de algoritmos idealizado para a recomendação é baseado em uma ontologia de

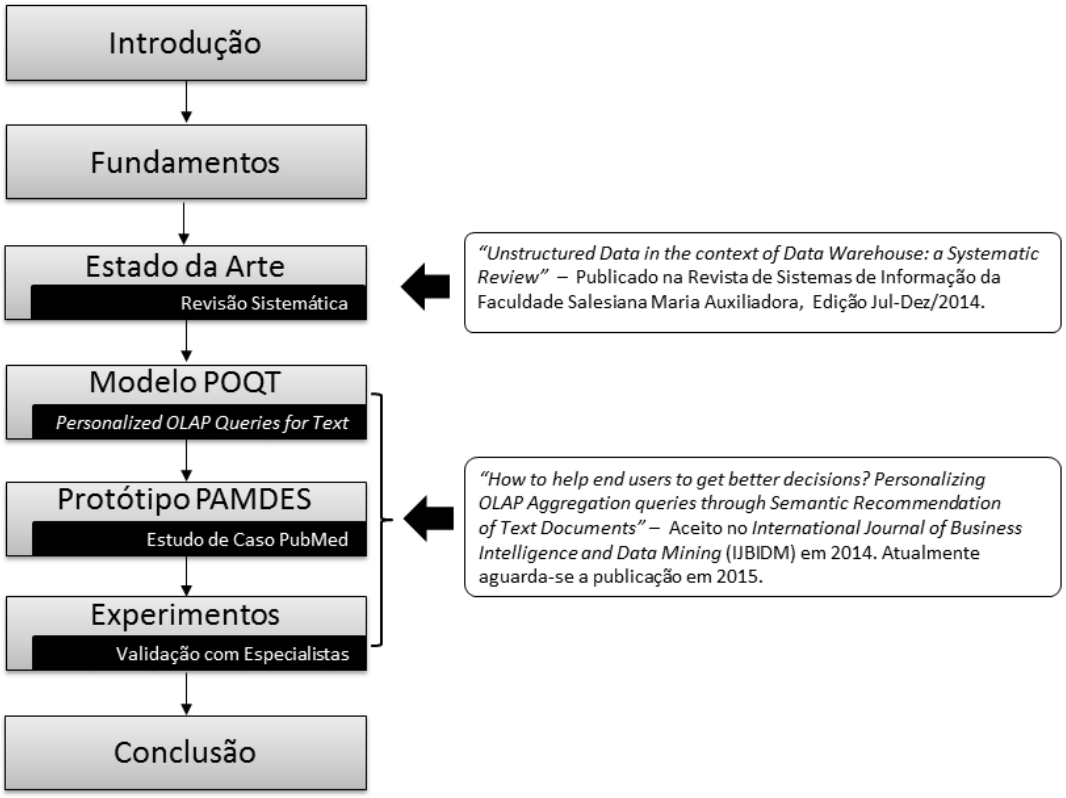


Figura 1 – Estrutura da dissertação e metodologia

domínio e conceitos de estatística para direcionar o processamento do OLAP textual. A fim de agregar e recomendar documentos, foi necessário utilizar medidas de similaridade semântica.

O uso de uma ontologia é fundamental, justamente porque ela visa fornecer a conceitualização de um domínio específico do problema, ou seja, um acordo comum de vocabulário para que os dados sejam referenciados. Desta maneira, uma ontologia descreve o significado e as relações dos termos de um domínio [SKOUTAS; SIMITSIS, 2007]. A ontologia tem a função de agregar documentos a partir de suas palavras-chave.

O significado e a maneira de abordar o conceito da semelhança entre documentos de texto foi construído a partir da agregação do algoritmo *AVG_KW* descrito em [RAVAT; TESTE; TOURNIER, 2007]. Para isso, foi proposta a utilização da métrica de frequência estatística para calcular como o número de combinações de palavras-chave irão satisfazer os parâmetros do usuário. O objetivo da personalização da consulta é oferecer ao usuário uma forma interativa para se obter uma agregação relevante de documentos baseados nesses parâmetros.

O protótipo *PubMed Article Multidimensional Recommendation System* (PAMDES) foi concebido a partir do modelo proposto (POQT) e a sua arquitetura permite o pré-processamento das recomendações personalizadas entre os documentos existentes

no *Data Warehouse*. A aplicação foi implementada com a finalidade de possibilitar a análise multidimensional de dados da PubMed [MEDICAL LITERATURE ANALYSIS RETRIEVEL SYSTEM ONLINE, 2014]. No estudo de caso para biomedicina, foi utilizado o MeSH (*Medical Subject Headings*) que é uma representação hierárquica fornecida pela Biblioteca de Medicina dos Estados Unidos [NATIONAL LIBRARY OF MEDICINE, 2014]

Experimentos foram conduzidos no sentido de validar os parâmetros de personalização, a avaliação das recomendações e dos resultados da expansão da recomendação. As validações são discutidas com base nas métricas de avaliação: *precision*, *recall* e *F1-measure*. Os resultados dos experimentos realizados com os especialistas mostram que boas recomendações são possíveis.

1.3 Organização do Trabalho

O trabalho está estruturado como segue (vide Figura 1). O Capítulo 2 descreve uma visão geral sobre *Data Warehouse* e OLAP, além de trazer uma revisão sobre os principais trabalhos correlatos, a fim de se conhecer os atuais modelos propostos nesta área, sejam eles modelo de OLAP textual ou de personalização e recomendação de consultas semântica. O Capítulo 3 descreve as definições do modelo conceitual, bem como a arquitetura do sistema e os algoritmos de recomendação. O Capítulo 4 tem o objetivo de detalhar como foi desenvolvido o protótipo PAMDES. O Capítulo 5 apresenta os resultados das avaliações experimentais. Por fim, no Capítulo 6 são apresentadas as conclusões e os trabalhos futuros.

2 Fundamentos e Estado da Arte

Este capítulo descreve aspectos relacionados à fundamentação teórica sobre os principais conceitos relacionados a *Data Warehouse*, sua estrutura, modelo, arquitetura e operações OLAP, assim como a integração de documentos de texto em ambientes de *Data Warehouse* por meio de ontologias. A seguir são detalhados os conceitos relacionados a OLAP para documentos de texto e o que a literatura científica atual oferece nesse sentido. Outro aspecto importante é melhorar a relevância dos resultados para os tomadores de decisão; por esta razão, é importante conhecer os conceitos sobre a personalização e recomendação semântica de consultas, assim como saber quais as propostas atuais tratam sobre esses assuntos.

2.1 Fundamentos sobre *Data Warehouse*

Um *Data Warehouse* (DW) é o componente central de uma estrutura de *Business Intelligence* (BI) que permite a análise multidimensional de grandes volumes de dados. Os sistemas transacionais, chamados de *On-line Transaction Processing* (OLTP), fornecem os dados para o *Data Warehouse* e é justamente neste novo ambiente que tais dados são consolidados para que seus usuários possam gerar informações estratégicas às empresas.

2.1.1 Definição

Segundo [INMON; GLASSEY; L., 1996], um DW é como uma coleção de dados orientada a assuntos, integrada, variável em relação ao tempo e não volátil, voltado à tomada de decisão.

Seguindo essa abordagem, as características de um *Data Warehouse* compreendem [HAN; KAMBER, 2006]:

- **Orientado a assuntos**: cada empresa possui uma necessidade diferente de informação, organizada por assunto ou tema. Pode-se citar como exemplos: as vendas de produtos a diferentes tipos de clientes, atendimento e diagnósticos de pacientes, rendimento de estudantes e conjunto de documentos para um determinado autor. *Data Warehouses* fornecem uma visão simples e concisa sobre uma questão particular, por meio da exclusão de dados que não são úteis ao processo de tomada de decisão.

- **Integrado**: diferentes nomenclaturas, formatos e estruturas das fontes de dados precisam ser inseridas em um único esquema para prover uma visão unificada e

consistente da informação, já que os *Data Warehouses* são construídos usualmente pela integração de múltiplas e heterogêneas fontes de dados.

- **Variável em relação ao tempo**: o histórico dos dados por um período de tempo superior ao usual em banco de dados transacionais permite analisar tendências e mudanças. Cada estrutura de chave no *Data Warehouse* contém, de forma explícita ou implícita, um elemento de tempo, pois os dados do *Data Warehouse* referem-se a algum momento específico, diferente portanto dos dados de produção que são alterados em decorrências das alterações frequentes nas operações das empresas. A cada ocorrência de mudança, uma nova entrada é criada no *Data Warehouse* para marcar a alteração.

- **Não volátil**: os dados de um DW não são modificados como em sistemas transacionais (exceto para correções), mas somente carregados e acessados para leituras, com apenas atualizações periódicas.

Os *Data Warehouses* fornecem dados para consulta e análise em diversos tipos de aplicações, tais como: *Decision Support Systems* (DSS), *Executive Information Systems* (EIS) ou mineração de dados. Os *Data Warehouses* também são elementos de sistemas relacionados a iniciativas estratégicas, como *Customer Relationship Management* (CRM), *Business Performance Management* (BPM) e à gestão da cadeia de suprimentos [ARIYACHANDRA; WATSON, 2010].

2.1.2 Arquitetura

A estrutura de um ambiente de *Data Warehousing* favorece a análise e obtenção de informações estratégicas, possibilitando um melhor suporte às decisões baseadas em ocorrências passadas e à predição de possíveis eventos futuros. Um DW consiste em uma plataforma de dados histórica, e por definição, os dados inseridos em um *Data Warehouse* não podem ser apagados, salvo em caso de erro, embora isto não seja uma regra.

A arquitetura mostrada na Figura 2, uma interpretação da apresentada em [KIMBALL; ROSS, 2002], representa a estrutura básica de interações realizadas em um ambiente de *Data Warehousing*. Nessa arquitetura podem ser identificados quatro elementos básicos:

- **Fontes provedoras**: os dados inseridos em um *Data Warehouse*, quase sempre, possuem origem em uma base de dados operacional (aplicativos comerciais, *Enterprise Resource Planning* (ERP), CRM, etc.), com isso, normalmente estão separados fisicamente do banco de origem que os armazena na forma primitiva do dado.

- **Área de Trabalho**: os dados introduzidos em um DW geralmente passam por uma etapa conhecida como área de *Data Stage*, que ocorre quando existem processos

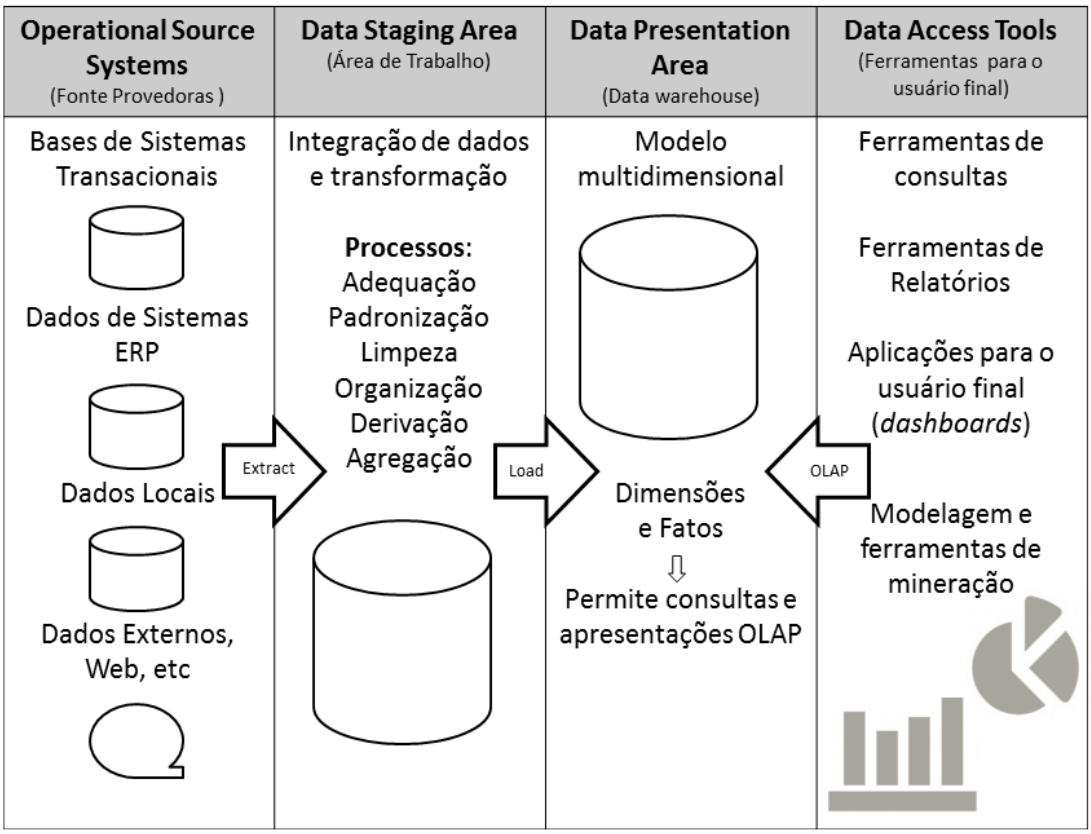


Figura 2 – Elementos básicos de um *Data Warehouse*

periódicos de leitura de dados e, em seguida, ocorre a integração e transformação de dados que consiste na adequação, padronização, limpeza, organização, derivação e agregação dos dados. Depois dessa etapa de transformação dos dados, ocorre a gravação dos dados no DW. Assim, pode-se observar que os dados são submetidos as etapas do processo *Extract Transform Load* (ETL).

A extração é a etapa inicial do processo ETL, ele consiste basicamente em ler e entender as fontes de dados e conduzí-las para a próxima etapa, onde serão realizadas as transformações sobre esses dados.

Após a extração, os dados passam por uma série de processos, de modo a convertê-los em formato adequado para serem armazenados no *Data Warehouse*. Essa conversão pode envolver um ou vários processos, dependendo da necessidade e situação. Seguem alguns dos processos mais comumente utilizados: limpeza, conversão, eliminação, combinação, cálculos, desnormalização e normalização.

A carga dos dados é a etapa final no processo ETL. Ela possui uma enorme complexidade, e assim os seguintes fatores devem ser levados em conta: integridade dos dados, tipo de carga a ser realizada (incremental ou total), otimização do processo de carga e suporte completo ao processo de carga.

- ***Data Warehouse***: área onde os dados estão organizados, armazenados e são

disponibilizados para consulta direta pelos usuários, geradores de relatórios e outras aplicações analíticas.

- **Ferramentas para o usuário final**: são as ferramentas de acesso à informação e, portanto, área que realiza a comunicação com o usuário final.

No *Data Warehouse*, os dados podem ser consultados por intermédio de ferramentas de consultas (*front-end*) pelos usuários finais, ou seja, são ferramentas de visualização para análise dos dados, como relatórios e gráficos. Já na área de *Data Stage* (*back-end*), não é acessível para os usuários e não fornece serviços de consulta ou apresentação. Já no

2.1.3 Modelagem Multidimensional

A modelagem de dados para *Data Warehouse* é diferente da modelagem de dados para sistemas OLTP, pois são modelagens para focos distintos. Sistemas OLTP são fundamentais para transações empresariais, pois são sistemas voltados para tarefas repetitivas. Esses sistemas transacionais possuem foco no nível operacional da organização, visando a execução operacional do negócio. No *Data Warehouse*, as operações OLAP permitem analisar um grande volume de informações em diferentes perspectivas, sendo assim, o seu foco está no nível estratégico da organização, visando a análise empresarial e tomada de decisão.

Um esquema é um projeto lógico ou físico de um conjunto de tabelas de banco de dados, indicando a relação entre as tabelas [KIMBALL; ROSS, 2002]. Assim como um banco de dados, um *Data Warehouse* também exige a definição de um esquema. Um banco de dados utiliza o modelo relacional e os *Data Warehouses* podem usar os seguintes esquemas: estrela *(star)*, *snowflake* e o *fact constellation*, também chamado de *galaxy*. Esses esquemas possuem variações, no entanto qualquer que seja o esquema na modelagem multidimensional, eles possuem basicamente dois tipos de tabelas: tabela fato e tabela dimensão.

O objeto de análise, ou seja, um fato, é um agrupamento conceitual de medidas. Essas medidas são valores (dados quantitativos) extraídos de sistemas operacionais e tradicionalmente são medições numéricas do negócio, porém a análise textual requer que essas medidas sejam textuais. Uma medida textual é uma medida que contém dados textuais, ou seja, representam palavras, parágrafos ou documentos completos [RAVAT; TESTE; TOURNIER, 2007].

As dimensões permitem a análise de um determinado modelo de diversas formas (visões). As dimensões são compostas por um conjunto de parâmetros (dados qualitativos) que são organizados em uma ou mais hierarquias.

O conceito de hierarquia é definido como uma sequência de mapeamentos de um conjunto de conceitos de baixo nível para conceitos de mais alto nível [HAN; KAMBER, 2006]. Um exemplo de mapeamento para uma hierarquia da dimensão *localização* seria ter cidades no mais baixo nível e no mais alto nível detalhe, ter a generalização dessas cidades em país.

O modelo estrela é caracterizado por uma tabela fato, parte central da estrela, que possui chaves compostas conectadas a um número determinado de tabelas de dimensão, sendo que cada uma possui uma única chave primária, como mostrado na Figura 3. As tabelas de dimensão possuem um conjunto de atributos. Os dados de vendas de uma determinada companhia constituem os dados de uma tabela fato, por exemplo. Essa tabela fato `Vendas` está cercada por quatro tabelas de dimensão, como `Filial`, `Tempo`, `Localização` e `Item`. Dessa maneira, a tabela fato contém chaves para cada uma das quatro dimensões. A tabela fato `Vendas` também pode conter atributos como o valor total de vendas e quantidade de unidades vendidas.

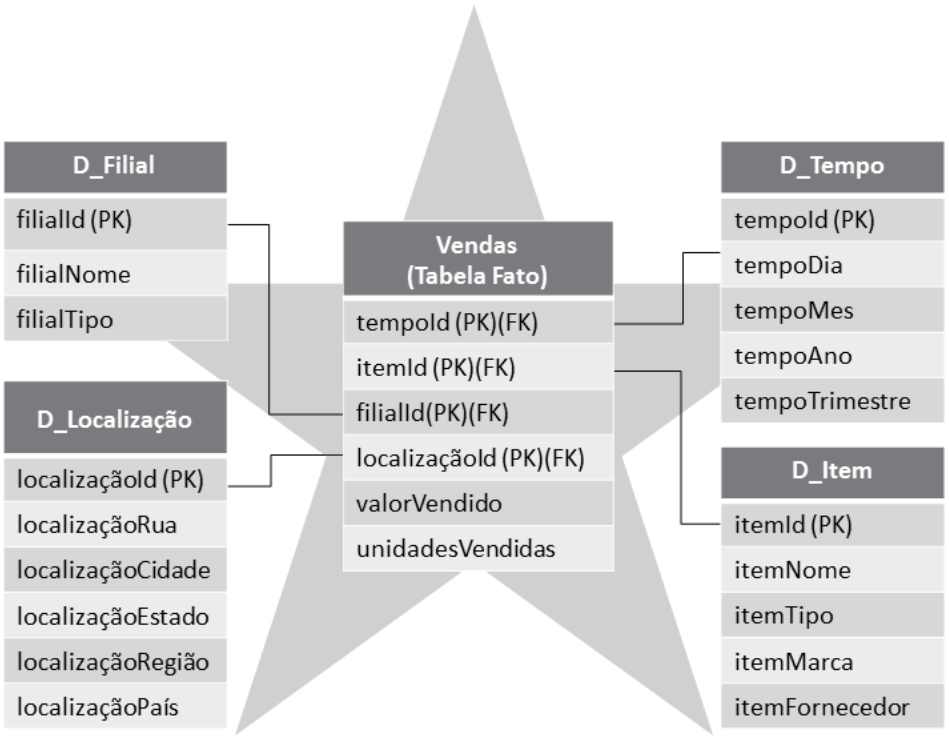


Figura 3 – Exemplo de modelo estrela

O esquema *snowflake* é uma variação do modelo estrela e o que difere um do outro é que algumas tabelas dimensão do esquema *snowflake* são normalizadas. Já o esquema *fact constellation* consiste em um modelo que possui várias tabelas fato.

Modelos multidimensionais implementados em bancos de dados multidimensionais são chamados de cubos OLAP [KIMBALL; ROSS, 2002]. Quando os dados são carregados em um cubo OLAP, eles são armazenados e indexados usando formatos e técnicas para dados

dimensionais, usando agregações, tabelas com dados pré-calculados e outras otimizações, o que permite que o seu desempenho seja melhor do que as consultas tradicionais. Os cubos consistem de um grande conjunto de fatos ou medidas.

Após a definição de um esquema lógico, torna-se necessário identificar quais dimensões serão usadas no cubo OLAP, principalmente para aplicar operações OLAP aos dados armazenados no *Data Warehouse*. O modelo multidimensional, portanto, permite que sejam dadas as respostas que atendam as questões complexas de análise de negócios, possibilitando essa análise sobre diversos pontos de vista. É com base nessas várias perspectivas, que muitas decisões acertadas podem ser tomadas.

2.1.4 Processamento Analítico On-line

Os dados armazenados em um *Data Warehouse* podem ser vistos em diferentes níveis de detalhe ou abstração, e a partir da combinação de diferentes atributos e dimensões. O OLAP consiste em um conjunto de técnicas para exploração de consultas multidimensionais de valores [TAN; STEINBACH; KUMAS, 2009]. Funções de análises relacionadas a OLAP enfocam diversas formas de analisar os dados usando visões multidimensionais complexas e elaboradas.

Dada uma determinada dimensão de interesse, os dados de um *Data Warehouse* podem ser organizados em uma hierarquia. As hierarquias permitem que os dados podem ser analisados em diferentes perspectivas. No modelo de dados multidimensional, os operadores OLAP básicos são:

- ***Roll-up***: também chamado de operação de agregação, analisa os dados em níveis de agregação progressivamente menos detalhados ou de maior granularidade, ou seja, retira-se detalhes na análise.

  Uma operação de agregação é caracterizada por permitir a combinação de células, dados, de uma ou mais dimensões do *Data Warehouse* para se obter informações mais gerais sobre um determinado assunto. Os atributos quantitativos, como preço e quantidade, são geralmente agregados por meio da soma ou média dos elementos. Já um atributo qualitativo, como um item, pode ser omitido ou resumido como uma lista de todos os itens que foram vendidos naquele local [TAN; STEINBACH; KUMAS, 2009].

- ***Drill-down***: é a operação inversa *roll-up*, cujo objetivo é obter uma informação mais detalhada, ou seja, descer na hierarquia de dimensão.

- ***Slice e Dice***: projeta valores específicos de uma dimensão. O *slice* é a operação que corta o cubo, mas mantêm a mesma perspectiva de visualização dos dados. A

operação *dice* é caracterizada por *slices* consecutivos, os quais mudam a perspectiva da visão.

- **Visualização Pivô (Rotate)**: muda a posição ou orientação das dimensões na visualização bidimensional dos dados.

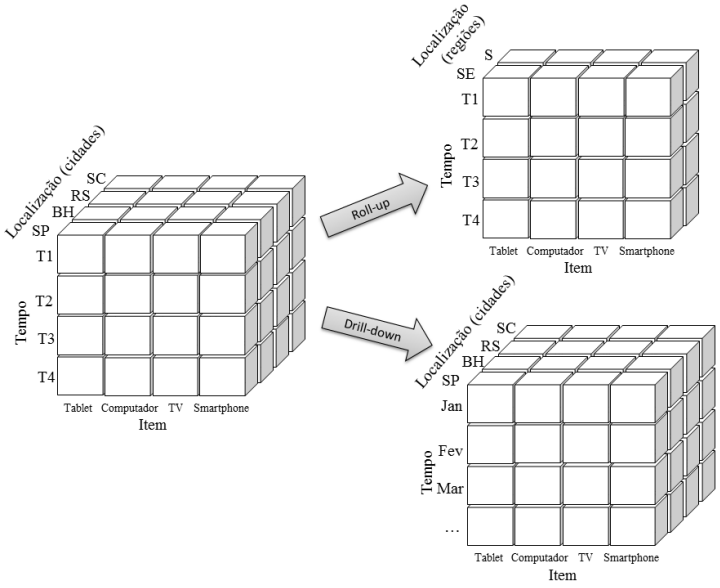


Figura 4 – Exemplos de operações OLAP *roll-up* e *drill-down* em banco de dados multidimensionais. Adaptado de [HAN; KAMBER, 2006]

Na análise multidimensional, busca-se descobrir um comportamento conforme a perspectiva de análise dos dados. Na Figura 4, baseada em [HAN; KAMBER, 2006], pode-se observar o relacionamento das seguintes dimensões: `Localização - cidades` (`SP`, `BH`, `RS` e `SC`), `Tempo - trimestres` (`T1`, `T2`, `T3` e `T4`) e `Item - tipos de produtos` (`Tablet`, `Computador`, `TV` e `Smartphone`), as quais caracterizam valores de vendas de produtos por trimestre e cidade.

O resultado da operação de *roll-up* permite visualizar a agregação dos dados, onde as quatro cidades são sumarizadas em duas regiões `Sul` (`S`) e `Sudeste` (`SE`), ou seja, a quantidade de produtos nas cidades é somada a fim de se obter o valor correspondente por região. A operação de *drill-down* mostra os dados em um nível de detalhe maior. No caso, a dimensão `Tempo` visualizada em trimestres é expandida para `Meses`, logo, serão relacionados os valores específicos de vendas de cada produto, a cada mês e cidade.

A operação de *dice* é representada por três operações *slice* consecutivas na Figura 5. A primeira define apenas duas cidade, `SP` e `BH`. A segunda restringe apenas os dois primeiros trimestres, e a terceira mostra dados de apenas dois produtos (`Tablet` e `Computador`).

Como resultante da operação de *slice* aplicada ao cubo (Figura 5), tem-se uma visão específica, que restringe a visualização de apenas duas dimensões, ou seja, há um

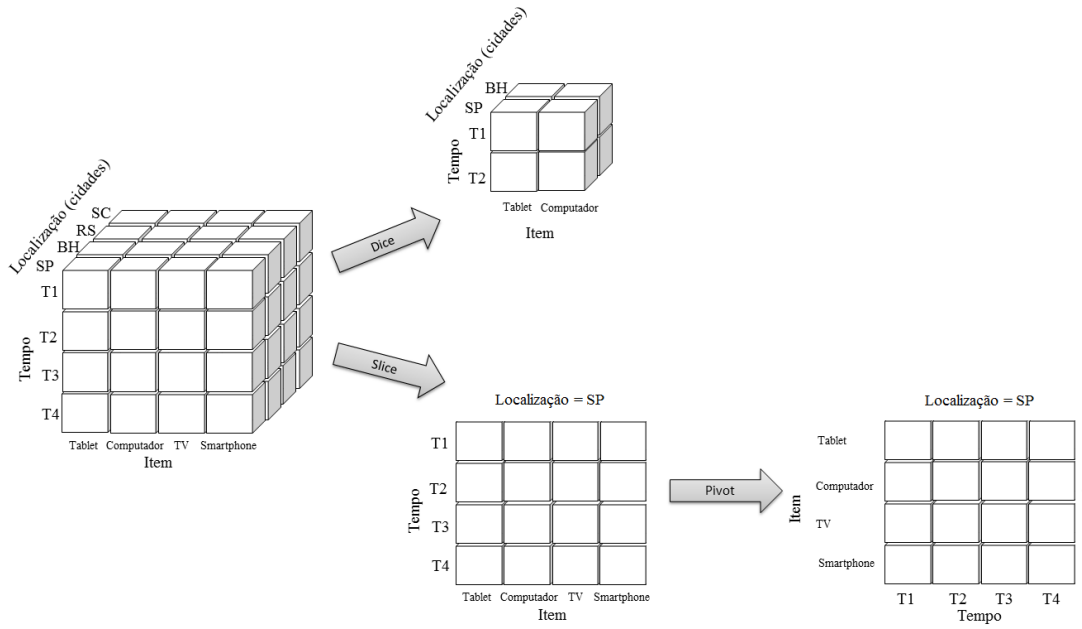


Figura 5 – Exemplos de operações OLAP *slice*, *dice* e *pivô* em banco de dados multidimensionais. Adaptado de [HAN; KAMBER, 2006]

corte em relação as informações completas de vendas, já que apenas se considera as vendas totais apenas para a cidade `SP`. A operação de *pivô* é mostrada na sequência ilustrando a troca de eixo da visualização bidimensional dos dados representados.

Além dos exemplos de operadores de consulta OLAP mostrados, as operações ainda podem ser combinadas entre si, ou seja, pode-se realizar um *slice* e um *dice* juntamente com operações de *drill-down* ou *roll-up*.

Os dados não estruturados, no entanto, não estão dentro do alcance das operações OLAP devido à falta de operações e gerenciamento de recursos para dados textuais. OLAP fornece ferramentas poderosas e métodos, mas com um rígido *framework* [RAVAT; TESTE; TOURNIER, 2007], portanto, dados não estruturados não se enquadram neste *framework*.

2.2 Conceitos Complementares

A integração da semântica de textos e a personalização de consultas multidimensionais tem o objetivo de melhorar a relevância dos resultados obtidos pelos tomadores de decisão. Os sistemas de recomendação oferecem esta possibilidade de personalização e uma das formas de obter a relação semântica entre documentos é por meio de uma ontologia. Por essa razão são descritas a seguir as tecnologias que também são utilizadas neste trabalho: ontologia e sistemas de recomendação.

No desenvolvido do protótipo PAMDES (Capítulo 5), uma ontologia foi utilizada para auxiliar no desenvolvimento de algoritmos para permitir a agregação (*roll-up*) de

documentos em bases multidimensionais. A ontologia utilizada no desenvolvimento reúne os conceitos e relacionamentos de domínio da aplicação, que nesse caso, trata-se do domínio da biomedicina.

2.2.1 Ontologia

Uma ontologia é uma especificação formal e explícita sobre um conceito compartilhado [SKOUTAS; SIMITSIS, 2007], onde os conceitos são ricamente definidos e organizados em hierarquias de subsunção [NEUMAYR; ANDERLIK; SCHREFL, 2012]. Em [ZHUOLUN; SUFEN, 2008], o termo ontologia é definido como uma maneira de descrever metadados em comum por meio de uma coleção semelhante de objetos.

Ontologia é um conceito isométrico emprestado da filosofia pelas comunidades da Inteligência Artificial e Tecnologia da Informação. Os elementos de sua composição como entidades, atributos, relações e axiomas podem ser expressos de uma maneira que não apenas os seres humanos consigam compreender, mas também máquinas podem interpretá-los, já que existe uma lógica na criação e na manutenção de ontologias.

No contexto de uma aplicação de *Data Warehouse*, as ontologias podem ser usadas como um modelo conceitual para descrever aspectos relacionados à semântica contida nas fontes de dados, permitindo assim que o uso de técnicas racionais possa ser aplicada para inferir correspondências e apontar conflitos entre as fontes [SKOUTAS; SIMITSIS, 2007]. O acesso à informação e o gerenciamento da informação são aspectos cada vez mais desafiadores quando se trata da alta taxa de volume de crescimento de dados. Por essa razão, as ontologias de domínio são uma forma fundamental de representação do conhecimento em um determinado domínio [MUSTAPHA et al., 2012].

Por sua vez, a ontologia deve ser bem construída para que possa realmente ajudar o desenvolvimento de pesquisas e gestão de sistemas de informação baseados em conhecimento, tais como ferramentas de busca, sistemas de recomendação, sistemas de classificação automática de textos, sistemas de gerenciamento de conteúdo, entre outros. No entanto, a efetividade desses sistemas depende da ontologia ser a mais completa possível e possuir uma representação que se adapte no processo de buscas em bases textuais.

Existem diferentes tipos ou níveis de ontologia. [BEPPLER, 2008] cita três tipos básicos de ontologia:

- *Ontologia de domínio*: representam o conhecimento de um domínio particular, como é o caso das ontologias específicas para as áreas de medicina, biomedicina, eletrônica, engenharia mecânica, comércio eletrônica, biologia dentre outras. Esse tipo de ontologia é a que foi utilizada neste trabalho.

- *Ontologias genéricas ou de senso comum*: ontologias utilizadas para armazenar o

conhecimento geral sobre o mundo, além de noções básicas e conceitos diversos, tais como: tempo, espaço, estado, evento, ação, etc. Essa categoria de ontologia é, portanto, independente de um problema ou domínio específico.

- *Ontologias representacionais*: armazenam a descrição de estruturas conceituais e metaestruturas que podem ser aplicadas em um âmbito geral, que são baseadas em visões lógicas e filosóficas não necessariamente focadas em aplicações [BEPPLER, 2008].

Em resumo, uma ontologia pode fornecer uma forma para descrever o significado e as relações dos termos em um domínio e por ser padronizada, a comunidade aceita a descrição do conteúdo da informação para resolver heterogeneidade, que incluem problemas semânticos [ZHUOLUN; SUFEN, 2008]. Por essas razões, as estruturas de conhecimento semântico podem fornecer uma valiosa base de conhecimento de domínio e de informações do usuário [MIDDLETON; ALANI; ROURE, 2002].

Dessa forma, a ontologia em si pode ser o coração de um sistema baseado em conhecimento, pois é por seu intermédio que o conhecimento de um domínio pode ser representado. O seu uso permite o desenvolvimento de sistemas inteligentes [BEPPLER, 2008]. Já o uso de ontologias na área da computação trouxe significativos avanços, pois mesmo com o avanço de *software*, os desenvolvedores e analistas de sistemas perceberam a importância de focar nos dados, sobre os quais os seus sistemas operam, em vez de focar em funcionalidades e aspectos procedurais dos próprios sistemas [BEPPLER, 2008].

2.2.2 Sistemas de Recomendação

A personalização da recuperação de informação e os sistemas de recomendação têm o propósito de encontrar a informação correta para os usuários com interesses diferentes [KANG; CHOI, 2011]. Os sistemas de recomendação surgiram como uma área de pesquisa independente em meados de 1990, quando os pesquisadores começaram a se concentrar em problemas de recomendação que explicitamente dependiam de uma estrutura de classificação. Em sua forma mais comum, o problema era reduzido a estimar a classificação para itens que não deveriam ser apresentados para os usuários [ADOMAVICIUS; TUZHILIN, 2005]. De forma intuitiva, essa estimativa baseia-se nas classificações dadas por esses usuários para outros itens, ou seja, assim seria possível estimar a classificação para itens ainda não classificados e que podem ser altamente recomendados para esses usuários.

Segundo [ADOMAVICIUS; TUZHILIN, 2005], são desenvolvidas novas abordagens sobre sistemas de recomendação tanto no meio acadêmico quanto na indústria. O interesse nessa área permanece elevado, pois é uma área de pesquisa que possui questões em aberto e também pela abundância de aplicações práticas. Essas aplicações e *softwares* que ajudam

os usuários a trabalhar com uma sobrecarga de informações e que fornecem recomendações personalizadas sobre conteúdos e serviços para eles.

Atualmente, a maioria dos usuários da *Web*, por exemplo, gastam uma grande quantidade de tempo para encontrar o conteúdo de que desejam, porque os sistemas de recuperação de informações possuem uma capacidade limitada para identificar os documentos apropriados, já que o volume de informações disponíveis é cada vez maior [KANG; CHOI, 2011].

Os sistemas de recomendação normalmente são utilizados para sugerir produtos baseados em compras ou pesquisas realizadas pelo cliente no passado, baseados no índice de produtos mais vendidos ou relacionados ao perfil do usuário. O uso das informações de contexto ajuda a melhorar os sistemas de recomendação, tornando-os mais refinados e precisos.

Ao se recomendar um item, é necessário definir quais parâmetros serão levados em consideração, seja, por exemplo, ao consumir produtos em um comércio eletrônico ou por número de acessos a uma determinada página de uma categoria, havendo ou não interação do usuário e suas preferências [JESUS; BRITO, 2011]. A identificação do usuário é opcional em sistemas de recomendação, já que seus perfis podem ou não ser considerados como parâmetros. Os parâmetros que serão usados pelo sistema de recomendação são definidos e os dados são coletados, e a partir desse ponto, é aplicada uma estratégia de recomendação que, por fim, é visualizada por meio dos itens recomendados ao usuário [BARCELLOS et al., 2007].

Os sistemas de recomendação podem ser classificados em três tipos [VIEIRA; NUNES, 2012]:

- *Sistema de recomendação baseado em conteúdo*: utiliza as informações de preferência do usuário para sugerir novos itens, ou seja, recupera informações já apontadas pelo usuário, assim recomendando itens similares àqueles já escolhidos.
- *Sistema de recomendação colaborativo*: considera as escolhas realizadas por usuários com características similares, considerando que usuários com características semelhantes terão as mesmas preferências.
- *Sistema de recomendação híbrido*: efetua a junção das técnicas de recomendação baseados em conteúdo e também focados nos usuários (colaborativa). Dessa forma, tanto a similaridade entre usuários e entre os itens utilizados são considerados, podendo assim sugerir mais itens e recomendando itens que não parecem estar relacionados [VIEIRA; NUNES, 2012].

A escolha do algoritmo mais apropriado para um sistema de recomendação depende de muitos problemas, incluindo o tipo de serviço específico, a natureza dos itens, em

conjunto com o tipo e a quantidade de informações disponíveis. Por exemplo, se os itens são documentos, um algoritmo baseado em conteúdo é mais apropriado, pois é capaz de lidar com os problemas relacionados com a análise automática de texto [DEHURI, 2012]. Baseado nesse princípio, a estratégia de recomendação utilizada no desenvolvimento do protótipo PAMDES é baseada em conteúdo, a qual permite a agregação (agrupamento) de dados textuais.

A avaliação do desempenho dos algoritmos de recomendação ocorre analisando essencialmente a satisfação do usuário, ou seja, o grau de aceitação das recomendações. Na maior parte dos casos existe um interesse maior em avaliar os itens que são de interesse do usuário e que deveriam ser recomendados, assim distinguindo as boas das más recomendações, uma vez que o objetivo dos sistemas de recomendação consiste em produzir boas recomendação [COSTA et al., 2014]. As três métricas mais comuns de avaliação são: *precision*, *recall* e *F1-measure*. Essas medidas serão exploradas na seção que descreve os experimentos e as validações relativas ao protótipo PAMDES no Capítulo 5.

2.3 Sumário

No decorrer deste capítulo foram abordados os principais conceitos relacionados a *Data Warehouses*. A arquitetura do *Data Warehouse* favorece a consulta e análise para diversas aplicações voltadas para a tomada de decisão. O processo de ETL é fundamental para tornar as informações integradas, padronizadas e disponíveis para ferramentas de acesso e análise de dados.

Os modelos multidimensionais, como o estrela, armazenam dados de maneira a responder questões complexas de análise de negócios. Os operadores OLAP são responsáveis pela análise dos dados em diversas perspectivas, em diferentes níveis de detalhe ou abstração, além de permitir a seleção de atributos e dimensões. A operação de agregação, *roll-up*, consolida os dados, ou seja, diminui o nível de detalhes. Essa operação é utilizada para permitir a recomendação semântica de documentos nesse trabalho.

Uma ontologia de domínio é utilizada nesse trabalho para se estabelecer a similaridade entre os descritores dos documentos. Já, os sistemas de recomendação e a personalização possuem o propósito de encontrar a informação correta para usuários com objetivos diferentes. O sistema de recomendação proposto é baseado em conteúdo, que nesse caso, são documentos de textos. Ambos os conceitos complementares, ontologia e sistema de recomendação, são utilizados para melhorar a relevância dos resultados das agregações OLAP.

3 Estado da Arte: Personalização, Recomendação e OLAP para Documentos de Texto

A seguir são apresentados propostas que a literatura científica atual oferece para trabalhar com documentos nos ambientes de *Data Warehouse*. Nesse primeiro grupo de trabalhos, algumas propostas apresentam operações OLAP adaptadas ou idealizadas para documentos de texto, o que é denominado como OLAP textual. Essas propostas são as principais abordagens que utilizam dados não estruturados para extrair informações úteis aos tomadores de decisão. Algumas abordagens utilizam medidas de RI e outras são baseadas em semântica e ontologia, porém nenhuma traz aspectos relacionados a formas de personalizar agregações e recomendar documentos.

A personalização e recomendação são os assuntos de um segundo grupo de trabalhos que são discutidos no decorrer do capítulo, justamente porque eles oferecem a possibilidade de trazer resultados mais interessantes aos tomadores de decisão. No entanto, essas propostas possuem como objeto de análise dados estruturados, ou seja, não há nenhuma abordagem que trabalhe a personalização e recomendação no contexto de dados não estruturados. Por esta razão, no decorrer da proposta do modelo multidimensional estendido desse trabalho, os temas OLAP textual, personalização e recomendação são explorados. Uma análise qualitativa e comparativa foi realizada com base no conjunto de trabalhos descritos no capítulo.

3.1 OLAP Textual

Os *Document Warehouses* surgiram a partir da necessidade de armazenar e obter informações mediante a exploração de dados textuais, pois são formados por uma estrutura de *software* para análise, compartilhamento e reutilização de dados não estruturados, não sendo exclusivo apenas para documentos de texto, mas também dados multimídia [CEMBALO; PISANO; ROMANO, 2012].

Para [TSENG; CHOU, 2006], a análise no *Document Warehouse* é realizada sobre o identificador de documento, ou seja, a análise é centrada no documento, além disso a estrutura provê um mecanismo de consulta com base em atributos de documentos e palavras-chave contidas em documentos. Já para [ISHIKAWA; OHTA; KATO, 2001], a navegação deve ser facilitada para a recuperação baseada em conteúdo.

Assim como em [RAVAT et al., 2007], [PUJOLLE et al., 2011] e [BOUKRAA; BOUSSAID; BENTAYEB, 2010], algumas abordagens apontam o *eXtensible Markup Lan-*

guage (XML) como uma solução ou parte da solução para analisar dados semiestruturados, sejam para modelos centrados em medidas numéricas ou não. O XML é uma linguagem de marcação recomendada pela W3C (*World Wide Web Consortium*) para a criação de documentos com dados organizados hierarquicamente. A linguagem XML é classificada como extensível, porque permite definir os marcadores, além adaptar-se aos mais diversos tipos de sistemas, ou seja, trabalhando com bases de textos ou de dados multimídias, como apresentado em [KIM; PARK, 2003].

Os modelos centrados em medidas numéricas comumente utilizam técnicas de Recuperação de Informações (RI) para extrair essas medidas e calcular seus pesos para documentos de texto, viabilizando a análise de textos. Ao utilizar técnicas de RI tem-se especial preocupação com a organização e recuperação de informações, a partir de um grande número de documentos baseados em texto [RAVAT et al., 2007].

No modelo espaço vetorial em sistemas de RI o método $tf.idf$, utilizado para cálculo de pesos, considera a frequência do termo no documento (*term frequency* - tf), onde quanto maior for o seu valor, mais relevante é o termo para o documento. O inverso da frequência do termo entre os documentos (*inverse document frequency* - idf) revela que o termo que aparece em muitos documentos não é útil para distinguir a relevância entre os documentos. Essas métricas, por sua vez, visam adicionar semântica para o OLAP de documentos de texto.

Alguns autores utilizam RI para documentos de textos aliados ao processamento OLAP, como [CHEN; GARCIA-ALVARADO; ORDONEZ, 2010] que utiliza além de RI, técnicas de *Text Mining* (TM). Já em [PÉREZ; BERLANGA; ARAMBURU, 2009], os autores utilizam conceitos de RI para criar cubos de documentos. A outra vertente de análise de documento de texto traz modelos de dados centrados em ontologia e Web semântica, as quais podem auxiliar a construção de um *Data Warehouse* específico para dados não estruturados [GONZÁLEZ; BERBEL, 2014].

Na literatura, existem propostas de modelagem de *Data Warehouse* que contemplam documentos de textos, porém existem abordagens distintas em relação a esse tipo de dados. Alguns sistemas trabalham com documentos, mas considerando suas medidas de RI baseadas no conteúdo do documento e outros adotam modelos centrados em semântica e ontologia. As propostas compreendem desde modelos multidimensionais tradicionais adaptados para dados não estruturados até aos que foram idealizados para trabalhar com dados de texto.

A seguir são descritos os seguintes modelos multidimensionais para texto que utilizam medidas de RI: modelo *Galaxy*, o *Text cube*, o *Topic cube*, *R-cube*, *Enhancing document exploration with OLAP* e o *Index cube*.

Modelo *Galaxy*

Os autores propõem, em [RAVAT et al., 2007], um modelo OLAP multidimensional conceitual, o *Galaxy*, sem a tabela fato. O modelo possui apenas eixos de análise, nomeados de dimensões. Estas dimensões são reunidas em grupos para indicar dimensões compatíveis para uma análise comum.

Para organizar os documentos é utilizada uma árvore de dados hierarquicamente estruturada. Os documentos podem referenciar a si mesmos ou outros documentos, e essas ligações precisam ser explícitas para garantir a navegação durante a análise dos dados.

Com o objetivo de exemplificar o modelo multidimensional, foi descrito o seguinte problema: *realizar uma análise sobre a atividade dos institutos de pesquisa, onde neste caso o tomador de decisão analisa publicações científicas, bem como os relatórios produzidos por esses institutos*. Para obter resultados a partir dessa análise, foi criado o modelo *Galaxy*. Neste exemplo, podem ser utilizadas duas ligações recursivas para navegar por intermédio: (1) das referências de artigos e (2) dos institutos dos autores. O documento foi analisado de forma contextual, por assunto. As operações de manipulações são um pouco diferentes das operações tradicionais, as quais são: *Focussing* e *Selection*, *Drilling operations*, *Analysis Reorganisation*, e o uso de links recursivos.

As agregações são realizadas utilizando a função de agregação *Top_Keyword* apresentada com detalhes em [RAVAT et al., 2008]. *Top_Keyword* é uma função que determina a compatibilidade entre dimensões usando o método $tf.idf$. A função tem o objetivo de selecionar as principais palavras-chave em um documento de texto.

Text Cube

O *Text Cube* [LIN et al., 2008] associa o OLAP tradicional e técnicas de RI para o tratamento dos textos. Dois tipos de hierarquias são utilizados: dimensão hierárquica, que é o mesmo conceito dos cubos de dados tradicionais, e hierarquia de termo (assunto). Este último conceito é introduzido no *Text Cube* como uma hierarquia para especificar os níveis de semântica e relacionamento entre os termos.

A hierarquia de termo (T) é construída sobre um conjunto de termos (W) para especificar os níveis semânticos dos termos e seus relacionamentos. Cada nó (n) em T é chamado de *generalized term*, representando um conjunto de termos. Cada nó folha da árvore possui apenas um termo e a raiz de T é um conjunto de todos os termos. Alguns outros termos também são definidos, tais como, o conjunto de nós filhos de um nó v, $chd(v)$, conjunto de nós descendentes de v, $des(v)$, e o nó pai de v ($par(v)$).

O *term level* (L) é um conjunto de nós em T, $L \subset T$. O conjunto de todos os nós folhas, L_0, é um termo de nível chamado de *top level*. Outros níveis de termos são generalizados por duas novas operações OLAP: *Pull-up* L em v e *Push-down* L em v.

Existem, também, medidas criadas para permitir de forma eficiente a recuperação de informações. Foram usadas duas medidas para se obter a agregação de dados de texto: tf e idf.

Topic Cube

O *Topic Cube* foi proposto em [ZHANG et al., 2009] e é construído baseado em um banco de dados de texto multidimensional. Um outro componente usado para a construção do cubo é uma árvore hierárquica de tópicos. A construção de um *Topic Cube*, inicialmente, traz dados gerais e em cada célula do cubo, ele armazena um conjunto de documentos agregados na dimensão texto. Assim, a partir do conjunto de documentos em cada célula, é realizada a mineração das distribuições das palavras nos tópicos definidos na árvore hierárquica nível por nível.

A abordagem envolve estender o cubo tradicional incorporando o modelo de análise semântica latente probabilística (*Probabilistic Latent Semantic Analysis* - PLSA) com um cubo de dados que possui parâmetros de um modelo probabilístico que pode indicar o conteúdo de um texto em uma célula.

O objetivo central do trabalho é usar a distribuição probabilística sobre as palavras do modelo para extrair um tópico no texto. Como exemplo, pode-se verificar a distribuição de probabilidades altas de algumas palavras e desta forma, capturar as distribuições de palavras que caracterizam cada tópico da hierarquia. Uma vez que se tem a representação da distribuição de cada tópico, pode-se mapear qualquer conjunto de documentos para a hierarquia de tópicos [ZHANG et al., 2009].

R-cube

O armazenamento contextualizado permite a análise em um cubo OLAP, chamado *R-cube* [PÉREZ; BERLANGA; ARAMBURU, 2009]. Os autores propõem que sejam utilizadas ferramentas de OLAP e de RI para realizar consultas e buscar informações em documentos de textos. O uso do XML permite que ocorra a união entre essas duas abordagens, tornando possível o desenvolvimento de um *Data Warehouse* para dados semiestruturados.

Foi criado uma nova técnica, a *Relevance Modeling*, para poder eleger os fatos descritos nos textos dos documentos de acordo com sua relevância calculada para a consulta IR-OLAP. Dessa forma, o *R-cube* mantém dimensões para recuperar documentos relevantes para o contexto selecionado. Essas dimensões são calculadas usando recuperação de informação e modelos probabilísticos. Operações algébricas para gerenciar os *R-cubes* também são descritas na proposta.

A arquitetura do armazenamento contextualizado proposto pelos autores combina

fontes de dados estruturados e documentos. Seus três componentes principais são: *Data Warehouse* corporativo, o *Document Warehouse* e módulo extrator de fatos.

Em linhas gerais, o *Data Warehouse* corporativo integra dados de diferentes áreas de uma empresa e permite a geração de cubos OLAP. O *Document Warehouse* armazena dados não estruturados provenientes de fontes internas e externas. Os dados desses documentos oferecem aos usuários informações adicionais relacionadas aos fatos. Já o módulo extrator tem o objetivo de relacionar os fatos do *Data Warehouse* corporativo com os documentos que descrevem o seu contexto. O módulo identifica os valores das dimensões em metadados e conteúdo textual dos documentos, a partir desse ponto faz um link entre os fatos caracterizados pelos mesmos valores de dimensão. A análise é realizada a partir do *R-cube*, o qual se materializa por meio da recuperação dos documentos e fatos relevantes no contexto selecionado pelo usuário.

Exploração de documentos aprimorada por meio do OLAP

A proposta em [CHEN; GARCIA-ALVARADO; ORDONEZ, 2010] é *keyword-centric* para consultas OLAP, ou seja, o plano é construir cubos de dados de possíveis palavras-chave. O estudo baseia-se em RI em documentos de textos e utiliza o processamento OLAP. Em um primeiro momento é feita uma classificação dos documentos usando técnicas de mineração de textos, tais como, *stemming* e *Vector Space Model* (VSM), para calcular a similaridade entre os documentos.

A partir dessa etapa, os cubos de dados são gerados e as operações OLAP podem ser executadas de forma otimizada utilizando as funções definidas pelo usuário, bem como usando o SQL padrão. O último passo seria prover a integração entre a pesquisa desejada e o OLAP, para que os documentos que possuam palavras-chave semelhantes a palavra-chave de pesquisa possam ser recuperados.

O cubo de dados gerado armazena tanto o número de ocorrências da combinação no espaço de documentos, bem como o número total de documentos em que aparece o conjunto de palavras-chave. Após a finalização da estrutura do cubo o usuário pode definir as suas consultas de interesse.

Index Cube

Em [JANET; REDDY, 2011], é apresentado um modelo de cubo, *Index Cube*, como um cubo tridimensional utilizando três estruturas de índice, e suas operações OLAP para documentos de texto.

No *Index Cube*, os usuários têm a flexibilidade para agregar medidas em um subconjunto de dimensões:

- *Hierarquia de documentos*: utilizada para especificar o relacionamento entre palavras

de um documento. Esta hierarquia é utilizada para analisar as palavras em vários níveis de abstração nos documentos.

- *Hierarquia de termos*: sua função é especificar os níveis semânticos e relacionamentos entre os termos dos textos. É utilizada para agrupar palavras para generalização em coleções de documentos.

- *Dimensão hierárquica*: é similar as dimensões comumente utilizadas em cubos de dados tradicionais.

A proposta do *Index Cube* mostra um modelo de índice de cubo com hierarquia de documentos e duas novas operações chamadas de *Scroll up* e *Scroll down*, que permitem analisar documentos em diferentes níveis de abstração.As três importantes estruturas de índice são: o índice direto, o índice invertido e o índice *Next-Word*, os quais são integrados em um único modelo de índice de cubo. A sua eficiência em comparação com a agregação e armazenamento foi estudada com a ajuda do índice *terrier*.

Os modelos baseados em semântica e ontologias são apresentados a seguir. Os estudos trabalham com dados estruturados ou semiestruturados, no caso documentos XML. Alguns modelos apresentam também possíveis operações OLAP baseadas no conteúdo do documento de texto.

Agregação a partir de uma ontologia

O trabalho, descrito em [RAVAT; TESTE; TOURNIER, 2007], possui uma abordagem de análise qualitativa e quantitativa de documentos de textos, isto é, analisam as palavras-chave de uma publicação a fim de obter uma descrição geral do conteúdo da publicação. São definidas medidas textuais adicionadas a um modelo multidimensional tradicional.

O esquema multidimensional dos autores ilustra um exemplo, onde se deseja obter uma visão sobre os assuntos de uma coleção de artigos científicos. Com esse objetivo, o tomador de decisão analisa as palavras-chave utilizadas pelos autores. A tabela fato é chamada de *Articles* e possui algumas medidas numéricas (*acceptance*, *text* e *keywords*). A partir da estrutura no modelo multidimensional, é definido o algoritmo *AVG_KW*, a fim de gerar agregações. Esse algoritmo é baseado em um modelo conceitual que fornece : (1) conceitos adaptados para apoiar medidas textuais não numéricas e (2) um novo conceito para conduzir o processamento da agregação do OLAP textual com o uso de uma ontologia domínio.

A função de agregação, descrita no algoritmo *AVG_KW*, é projetada para agregar conjuntos de palavras-chave. Dado um conjunto de palavras-chave (*KW_List*), uma

distância máxima (D_{MAX}) e uma ontologia (O) como entrada, a função gera um novo conjunto de palavras-chave agregadas.

Para cada par de palavras-chave, a função encontra o menor ancestral comum entre elas por meio da ontologia. Ao agregar palavras-chave muito distantes, não importa o quão profunda é a ontologia, existe uma alta probabilidade de retornar a palavra-chave raiz da ontologia. Para evitar que isso ocorra, deve ser especificado um limite dentro do processo agregação, ou seja, o algoritmo usa uma distância máxima autorizada ao agregar palavras-chave denominada de D_{MAX}.

Dessa forma, caso o valor obtido pelo cálculo do LCA (*Least Common Ancestor*) seja inferior ou igual ao D_{MAX} significa que as palavras-chave podem ser agregadas. A palavra-chave que corresponde a agregação é adicionada ao conjunto de retorno do algoritmo, o conjunto de palavras-chave agregadas (*output_List*).

Algorithm: *AVG_KW*
Input: *KW_list*, D_{MAX}, O
Output: *output_List*
// lista de palavras-chave ordenada de acordo com a ontologia O (bottom-up)
KW_List = OrderList (*KW_List*, O)
FOR EACH KW_i **OF** *KW_List* **DO**
  $\mathrm{l}_i \leftarrow 0$
  FOR EACH KW_j **OF** *KW_List*, $(j > i)$ **DO**
    $\mathrm{KW}_{LCA} \leftarrow \mathrm{LCA}(\mathrm{KW}_i, \mathrm{KW}_j)$
    // calcula a distância $d_O(KW_i, KW_j)$
    $\mathrm{l}_{LCA} = \mathrm{MAX}(\mathrm{d}(KW_i, KW_{LCA}), \mathrm{d}(KW_i, KW_{LCA})) + \mathrm{l}_i$
    IF $(\mathrm{l}_{LCA} \leq D_{MAX})$ // verifica se satisfaz o parâmetro definido pelo usuário
      // substitui as palavras-chave originais por $LCA_O(KW_i, KW_j)$
      KW_List $\leftarrow$ *KW_List* - $\{\mathrm{KW}_i, \mathrm{KW}_j\}$
      $\mathrm{KW}_i = \mathrm{KW}_{LCA}$
      $\mathrm{l}_i = \mathrm{l}_{LCA}$
    END IF
  END FOR
  Add KW_i TO *output_List*
END FOR

Complex Cube

O modelo multidimensional em [BOUKRAA; BOUSSAID; BENTAYEB, 2010] usa o conceito do *Complex Object* (*CO*) proposto em [BOUSSAID et al., 2007] que é uma solução para a integração de dados complexos (várias fontes de dados, formatos heterogêneos, diversas estruturas, etc). De acordo com os autores, o *CO* é uma entidade física ou abstrata composta por um ou mais sub-documentos. Cada sub-documento pode representar um texto simples ou texto com *tags*, um visão relacional, uma imagem ou dados temporais.

No modelo em questão, o *CO* é utilizado para representar tanto fatos como membros de dimensões. Os autores também expõem cinco definições para compor seu modelo multidimensional de objetos complexos. Eles propõem um conjunto de operadores de

OLAP para: (1) a construção de cubos de dados complexos (*cubic projection*) e (2) para visualizar os cubos de dados (*visualization operators*).

A fim de validar o modelo e operadores multidimensionais, foi desenvolvido o núcleo de uma estrutura de armazenamento e análise. A transferência dos elementos da modelagem conceitual em níveis lógicos e físicos foi realizada utilizando XML, que é a principal fonte do *Data Warehouse*.

Os módulos da arquitetura funcionam da seguinte maneira: (1) o módulo de ETL lê o XML, carrega os dados para os arquivos XML do *Data Warehouse* e por fim esses arquivos são armazenados em um banco de dados XML nativo; (2) o módulo de especificação do cubo implementa o operador da projeção do cubo. Ele lê o arquivo de metadados, assim como os arquivos de dados, e produz um arquivo de metadados e um conjunto de documentos XML que contêm os dados reais.

Análise *Document-Centric* em Documentos XML

O estudo descrito em [PUJOLLE et al., 2011] utiliza o conceito do modelo *Galaxy*, [RAVAT et al., 2007], associado com um processo adaptado para uma análise específica em documentos XML. Os requisitos de usuário são representados no modelo.

A proposta destaca a diferença entre os tipos de análise existentes para documentos XML. Uma é centrada em dados dos documentos (*data-centric*), onde os dados são altamente estruturados e a ordem dos dados não é determinante, o que é diferente da análise de centrada em documentos (*document-centric*), abordada pelos autores, onde a estrutura é mais dispersa justamente por conter mais textos e onde a ordem dos elementos é muito importante (ordem dos parágrafos em um artigo, por exemplo).

Os autores descrevem os detalhes das etapas de projeto conceitual de um banco de dados multidimensional de documentos XML. Com a utilização de um modelo conceitual multidimensional adaptado, o processo de concepção permite a integração de dados extraídos do documento de texto XML dentro de um sistema adaptado OLAP.

O processo é composto por cinco fases diferentes: (1) a princípio é realizada uma análise paralela dos requisitos do usuário definidos por meio de um esquema conceitual; (2) também são definidas as fontes de dados, no caso, um documento de texto XML; (3) a seguir ocorre uma etapa de confronto, garantindo a compatibilidade entre as fontes de dados e o banco de dados multidimensional; (4) se o dicionário de sinônimos não puder ser usado para facilitar o processo, pode ocorrer a incompatibilidade, e será necessário revisar os requisitos ou as fontes de dados; (5) ocorre a iteração do processo até se encontrar a melhor compatibilidade e por fim a estrutura do banco de dados multidimensional é criada e é feito o carregamento com dados extraídos dos *data sources*.

MDOs - *Multidimensional Ontologies*

Em [NEUMAYR; ANDERLIK; SCHREFL, 2012] é proposto um modelo de OLAP baseado em ontologia, onde as dimensões e fatos são enriquecidos por definições conceituais como semânticas capturadas termos relevantes aos negócios.

O conceito de *Multidimensional Ontologies* (MDOs) é descrito como um modelo multidimensional com conceitos determinados (*views*) e a subsunção do raciocínio sobre conceitos definidos. Uma hierarquia de conceitos é interpretada com um conjunto de nós pertencentes à mesma hierarquia. Um nó é membro de uma hierarquia de conceitos se este é um descendente do nó raiz e para cada restrição de nível ancestral, ele ou um de seus antepassados satisfaz a restrição do conceito de subsunção. Se um nó é membro de uma hierarquia de conceitos, então todos os seus nós descendentes também são membros da hierarquia de conceito.

O modelo, portanto, possui uma ontologia atuando como uma camada conceitual que é inserida entre os analistas de negócios e os dados multidimensionais. Usando as ontologias tradicionais neste contexto, os autores relatam que é difícil de capturar os conceitos hierárquicos e multidimensionais dos analistas de negócios, por isso são propostas ontologias hierárquicas e multidimensionais para melhor capturar essas especificidades estruturais do modelo.

Document Warehousing

Os autores de [CEMBALO; PISANO; ROMANO, 2012] definem os elementos fundamentais do ciclo de vida de um sistema de *Document Warehousing* com base na abordagem tradicional de dados estruturados.

Um *Document Warehousing* é um sistema avançado de análise que permite ao usuário simplesmente consultar um grande número de documentos (*documentary corpus*), usando a tecnologia OLAP, para obter uma sumarização das informações, o que inclui o conteúdo dos documentos, metadados de documentos e agrupamento dos documentos. Este sistema de armazenamento também favorece a economia de memória, já que o documento continua armazenado no repositório original, enquanto que apenas a informação extraída a partir dele e um ponteiro para o documento são armazenados no *Document Warehouse*.

O trabalho destacou a necessidade de extrair informações estruturadas de fonte de dados não estruturadas (base textual) e ainda assim manter as informações relevantes em cada documento. O modelo multidimensional estabelece três tipos de dimensões: (1) comum, como um conjunto de palavras-chave; (2) dimensões de metadados e (3) dimensão categoria, como uma hierarquia ou uma ontologia. As medidas de fatos permanecem sendo numéricas.

3.2 Personalização e Recomendação OLAP

Sistemas recentes consideram o uso de técnicas de recomendação para sugerir consultas para o *Data Warehouse* e ajudar o analista a prosseguir a sua exploração [THOLLOT; KUCHMANN-BEAUGER; AUFAURE, 2012].

O objetivo das consultas OLAP é justamente resumir grandes volumes de dados históricos de acordo com múltiplas dimensões em vários níveis de granularidade. Para fornecer aos usuários apenas os dados relevantes da enorme quantidade de informações disponíveis no modelo multidimensional, sistemas OLAP podem considerar as preferências dos usuários nos eixos da agregação de sua consulta, bem como os níveis de agregação. Esses elementos determinam os dados que serão agregados e se sua configuração não for adequada pode-se obter um conjunto de dados inútil e insatisfatório [JERBI et al., 2008].

Framework para OLAP Personalizado

Na literatura são encontrados alguns trabalhos sobre a personalização de consultas OLAP. Os autores de [BELLATRECHE et al., 2005] oferecem um *framework* de personalização para calcular e visualizar o mais interessante subconjunto de fatos baseados no perfil dos usuários. As preferências dos usuários são pré-determinadas sobre os membros de todas as dimensões e também por *visualization constraints* (usado para definir precisamente a estrutura de um cubo).

Em suas definições, o *framework* possui algoritmos para encontrar o mais interessante sub-cubo de um cubo que pode ser visualizado e para testar se um cubo pode ou não ser visualizado. A principal consequência de se considerar os perfis de usuário para personalização de consultas é a adição de uma nova seleção e condições de união nas consultas iniciais.

Personalização OLAP - Modelo Rok

Para permitir a personalização OLAP, a proposta apresentada em [BENTAYEB; FAVRE, 2009], modelo *Rok*, consiste em integrar uma nova informação ou conhecimento dentro do *Data Warehouse*. São considerados dois tipos de conhecimento: o conhecimento explícito expresso pelos próprios usuários e o conhecimento extraído a partir dos dados usando técnicas de mineração de dados.

O trabalho está dividido em quatro principais processos, são eles: (1) a aquisição do conhecimento; (2) a integração de conhecimento no *Data Warehouse*; (3) o modelo evolutivo do esquema de *Data Warehouse* e (4) a personalização da consulta OLAP. A personalização é composta por duas técnicas denominadas de adaptação e recomendação, onde cada técnica segue as quatro etapas acima citadas.

O objetivo do modelo é recomendar novos níveis de hierarquia na dimensão. As agregações são baseadas em *K-means* para valores numéricos e categóricos. O modelo permite a personalização de usuário para consultas OLAP, especificando o número de classes e de atributos envolvidos no processo do *K-means*.

Framework para recomendação baseada em *logs*

A recomendação de consultas OLAP usando o que o usuário fez durante a sua antiga exploração no cubo de dados é apresentado em [GIACOMETTI; MARCEL; NEGRE, 2008]. Um *framework* genérico permite recomendar consultas OLAP com base em registros (*logs*) de solicitações ao servidor OLAP. O termo genérico é usado no sentido de que o usuário ao mudar seus parâmetros altera a forma de calcular as recomendações.

O processo ocorre da seguinte forma: (1) particionar o *log* em grupos de consultas que são semelhantes e computar sessões generalizadas; (2) gerar recomendações candidatas por intermédio de uma primeira busca na sessão de *log* atual e prevendo consultas futuras que possam interessar ao usuário; (3) classificando as consultas candidatas, de modo a presentar ao usuário as consultas mais relevantes primeiro.

Recomendação baseada em preferências e contexto

Em [JERBI et al., 2008], a consulta inicial é reforçada com as preferências do usuário integrando a dimensão relevante e/ ou atributos para a agregação. A abordagem da personalização é baseada sobre a manutenção de um conjunto de preferências para cada usuário, cuja estrutura está intimamente relacionada com as características do modelo de dados multidimensional.

Uma extensão de um modelo multidimensional é definida a fim de integrar as preferências do usuário. A extensão consiste em adicionar as preferências qualitativas nas dimensões, bem como entre as dimensões.

A linguagem algébrica utilizada permite a apresentação da análise de dados em uma tabela multidimensional, que mostra dados de um fato e duas de suas dimensões vinculadas. As operações OLAP permitem a manipulação e recuperação de dados a partir das preferência baseadas na constelação.

Os mesmo autores, em [JERBI et al., 2009], consideram também que as preferências são sensíveis ao contexto, no entanto, a essência da sensibilidade ao contexto é que os itens têm diferentes relevâncias, pesos, dependendo do contexto do utilizador. Por esta razão, exploram como capturar diferentes graus de relevância de um mesmo elemento em diferentes contextos.

O processo de recomendação ocorre em duas etapas distintas: (1) construção das recomendações, onde o sistema gera nós de análise úteis, e (2) classificação das recomenda-

ções, onde os nós candidatos são classificados em relação às preferências pontuadas, logo apenas a melhor recomendação selecionada é entregue ao usuário.

Recomendação de Relatórios OLAP baseado em preferências

Uma outra visão em relação aos relatórios gerados pelas ferramentas de *Data Warehouse* é abordado em [KOZMINA; SOLODOVNIKOVA, 2011]. Neste trabalho, é apresentada uma ferramenta específica para relatórios OLAP, que considera e processa as preferências do usuário, as quais são úteis para gerar recomendações sobre os relatórios que são potencialmente de interesse do usuário.

Existem camadas de metadados tanto em nível lógico quanto físico. A camada de metadados lógica é usada para descrever o esquema do *Data Warehouse*. A camada física descreve o armazenamento no *Data Warehouse* na base de dados relacional. Os metadados semânticos descrevem os dados armazenados e os elementos de uma forma compreensível ao usuário. Os metadados das preferências OLAP guardam definições das preferências dos usuários na estrutura dos relatórios e seus dados.

Já os metadados dos relatórios descrevem a estrutura de relatórios sobre elementos de *Data Warehouse*. Basicamente, os relatórios são planilhas que contêm itens de dados definidos por cálculos, que especificam fórmulas computacionais de parâmetros e colunas da tabela que normalmente correspondem ao esquema elementos (medidas e atributos). Os relatórios também consistem de condições definidas pelo usuário e são unidas entre as tabelas.

O processo de preferência para criação e transformação do relatório é descrito em cinco passos, são eles: (1) descrição inicial das preferências; (2) normalização das preferências; (3) classificação e reformulação das preferências; (4) indicação da importância das preferências e (5) processamento das preferências e geração dos relatórios recomendados.

Expansão de consultas personalizadas

O uso de técnicas de recomendação são utilizadas para ajudar o usuário a obter informações mais relevantes em relação a sua consulta, como acontece em [THOLLOT; KUCHMANN-BEAUGER; AUFAURE, 2012]. O objetivo do estudo é auxiliar o usuário na fase de concepção da sua consulta oferecendo sugestões de medidas e dimensões.

A expansão da consulta personalizada leva em consideração preferências explícitas e implícitas, combinadas linearmente. As explícitas são, por exemplo, *feedbacks* recebidos do usuário, como classificações atribuídas às medidas, no caso dimensões. Preferências implícitas podem ser derivadas a partir de fontes variadas, como na análise de *logs* de consultas executadas nas sessões do usuário. A expansão também conta com a semântica de modelo de domínio multidimensional e as estatísticas de uso derivadas da coocorrência

de medidas e dimensões em repositórios de documentos BI.

Recomendações semânticas por meio de grupos

Em [AHMED et al., 2014], o sistema *SMART* é apresentado para recomendações semânticas multidimensionais em grupos com o intuito de melhorar a formulação de consulta. O objetivo é inferir o comportamento dos analistas, tomadores de decisão, sobre conceitos ontológicos, utilizando uma estratégia de soma ponderada.

A necessidade da recomendação em grupo ocorre quando existe um grande número de perfis de usuários. Logo, o desafio é descrever a semântica de recomendação do grupo, já que existem diferentes relevâncias para diversos membros de um grupo, revelando divergências e conflitos entre eles.

O sistema de recomendação possui duas camadas: (1) camada de modelagem semântica do grupo, que descreve os interesses do grupo usando conceitos de uma ontologia e (2) camada para geração de recomendações, a qual deriva as recomendações semânticas para a construção dos perfis ontológicos.

Quando o tomador de decisão realiza consultas OLAP selecionando esquemas do *Data Warehouse*, o perfil ontológico do seu grupo é atualizado e todas as anotações associadas de conceitos são ajustadas pela propagação das atividades recentes dos usuários.

3.3 Resumo Qualitativo

Para se obter uma síntese dos trabalhos relacionados neste capítulo foram propostos critérios qualitativos para avaliá-los. Os estudos estão sumarizados em cinco categorias como mostradas na Tabela 1. Cada categoria é marcada quando o tópico em questão está presente no estudo.

Categoria	**Descrição**
A	Modelagem multidimensional com novo modelo orientado a documentos
B	Semântica por meio de ontologia
C	Novas operações OLAP
D	Personalização
E	Recomendação

Tabela 1 – Categoria e descrição dos assuntos

As categorias correspondem aos seguintes assuntos abordados nos estudos:

- **Categoria A** - *Modelagem multidimensional com novo modelo orientado a documentos*: estudos que focam em novas soluções conceituais para modelagem multidimensional para dados textuais. Os estudos desta categoria trazem modelos que

permitem armazenar documentos no ambiente do *Data Warehouse* usando técnicas distintas e privilegiando o acesso aos dados textuais, sejam com técnicas de RI, medidas probabilísticas, estruturas de árvores ou índices.

- **Categoria B** - *Semântica por meio de ontologia*: estudos que usam ontologias para acrescentar semântica à análise de documentos. Os estudos desta categoria incluem abordagens de modelagem multidimensional que buscam de alguma forma representar o conteúdo dos documentos por intermédio de uma representação hierárquica dos conceitos, ou seja, enriquecendo a base de dados com o uso da semântica.

- **Categoria C** - *Novas operações OLAP*: estudos que possuem a descrição de novas funções para realizar operações OLAP para textos. Os estudos desta categoria possuem abordagens que permitem realizar operações OLAP por meio de cubos de decisão constituído de dados textuais. As operações pode ser as operações tradicionais OLAP adaptadas para documentos ou diferenciadas, as quais são exclusivas para documentos de texto.

- **Categoria D** - *Personalização*: estudos que possuem formas de personalização de consultas de acordo com a preferência ou perfil do usuário final. Os estudos abrangem propostas que integram as escolhas do usuário no processamento de consultas OLAP de forma a atingir um resultado mais satisfatório.

- **Categoria E** - *Recomendação*: estudos que possuem formas de recomendar novas agregações OLAP, oferecendo ao usuário final resultados mais relevantes para suas consultas. Os trabalhos incluem propostas que utilizam *logs*, contexto, grupos ou preferência, sugerindo medidas e dimensões das consultas.

Alguns estudos foram apontados em mais de uma categoria, pois apresentam contribuições não apenas sobre um assunto específico. Foram definidos mais dois critérios com o objetivo de se obter uma visão geral da qualidade dos estudos. Os estudos podem se encaixar nas seguintes situações: (1) apresenta evidência de implementação (EI), ou seja, aqueles que apresentam um protótipo além de uma proposta teórica e (2) apresenta um estudo de caso (EC), que relata explicitamente uma pesquisa empírica com a sua implementação e os resultados experimentais.

A fim de oferecer uma descrição geral dos objetivos e características de cada estudo, na Tabela 2, observa-se a distribuição dos estudos de acordo com a sua categoria. As colunas EI e EC, representam respectivamente, trabalhos que possuem evidência de implementação e que aqueles que possuem estudo de caso.

Os percentuais da Tabela 2 são calculados a partir do número de estudos que correspondem ao critério dividido pelo número total dos estudos (18). A categoria A, relacionada as novas formas de modelagem multidimensional para documentos, representa 61,1% dos

Estudo	Categoria					EI	EC
	A	B	C	D	E		
1. [RAVAT et al., 2007], [RAVAT et al., 2008]	X		X			Sim	Não
2. [LIN et al., 2008]	X		X			Sim	Sim
3. [ZHANG et al., 2009]	X					Sim	Sim
4. [PÉREZ; BERLANGA; ARAMBURU, 2009]	X					Sim	Sim
5. [CHEN; GARCIA-ALVARADO; ORDONEZ, 2010]	X					Sim	Não
6. [JANET; REDDY, 2011]	X		X			Sim	Sim
7. [RAVAT; TESTE; TOURNIER, 2007]	X	X	X			Sim	Não
8. [BOUKRAA; BOUSSAID; BENTAYEB, 2010]	X					Sim	Não
9. [PUJOLLE et al., 2011]	X					Sim	Não
10. [NEUMAYR; ANDERLIK; SCHREFL, 2012]	X	X				Sim	Não
11. [CEMBALO; PISANO; ROMANO, 2012]	X	X				Sim	Não
12. [BELLATRECHE et al., 2005]				X		Sim	Não
13. [BENTAYEB; FAVRE, 2009]				X	X	Sim	Não
14. [GIACOMETTI; MARCEL; NEGRE, 2008]					X	Sim	Sim
15. [JERBI et al., 2008], [JERBI et al., 2009]					X	Não	Sim
16. [KOZMINA; SOLODOVNIKOVA, 2011]					X	Não	Sim
17. [THOLLOT; KUCHMANN-BEAUGER; AUFAURE, 2012]				X	X	Não	Sim
18. [AHMED et al., 2014]					X	Sim	Sim
Total	11	3	4	3	6	15+	9+
(%)	61,1	16,7	22,2	16,7	33,3	83,3	50

Tabela 2 – Resumo da análise qualitativa

temas abordados nos estudos. Isso indica que os estudos estão concentrando seus esforços na captura da natureza não estruturada dos dados em seus modelos multidimensionais adaptados. Além disso, são descritos nestes trabalhos operadores OLAP que sejam capazes de manipular dados textuais (categoria C). Dessas propostas (estudos 1 a 11) apenas quatro (estudos 1, 2, 6 e 7) possuem a idealização de operações OLAP para textos, o que ainda demonstra que é um desafio criar operações que manipulem documentos de texto de forma adequada.

Três propostas diferentes (estudos 7, 10 e 11) obtêm resultados em seus experimentos utilizando a ontologia para representar o conhecimento de um domínio específico e assim poder extrair informações dos documentos de texto (categoria B). Desses trabalhos, apenas dois possuem algum estudo de caso para demonstrar as suas propostas.

Sobre o tema personalização (categoria D), apenas três trabalhos destacam maneiras de auxiliar o tomador de decisão pela adaptação de consultas que sejam adequadas as suas necessidades de informação. Já os últimos seis trabalhos (estudos 13 à 18) apresentam estudos relacionados a recomendação de consultas OLAP (categoria E).

Na Tabela 2, nota-se que 83,3% dos estudos têm evidências de implementação

e 50% relataram ser explicitamente uma pesquisa empírica com implementação e seus resultados experimentais. A falta de estudos de caso é um reflexo do desenvolvimento atual da área de pesquisa, pois ela ainda é recente. Investimentos futuros em termos de pesquisa irão trazer resultados mais conclusivos por meio da realização de experimentos. A presença de empresas ou indústrias do ramo nas pesquisas pode justamente aumentar o número de resultados experimentais e as patentes de *software* nessa área.

Outro aspecto verificado por meio da análise qualitativa é que os trabalhos de personalização e recomendação propõem estudos que tratam em sua essência de dados estruturados, ou seja, não há propostas específicas para recomendação e personalização destinadas para documentos de textos (dados não estruturados). Da mesma forma que os trabalhos que tem foco em dados não estruturados não possuem tópicos relacionados à personalização e recomendação.

Por essa razão o modelo multidimensional proposto neste trabalho visa justamente trabalhar no sentido de integrar a personalização/recomendação e a semântica, com a tentativa de fornecer respostas de maneira mais precisa e eficiente, trabalhando com bases de texto. As definições do novo modelo orientado a documentos utiliza uma ontologia para obter a operação OLAP de agregação para texto. O usuário, tomador de decisão, pode personalizar os parâmetros da agregação, sejam parâmetros utilizados na análise da ontologia ou relacionado ao cubo de consulta. Mediante a operação de agregação é possível então obter um recomendador de documentos com base na personalização definida pelo usuário.

4 Consultas OLAP Personalizadas

O presente capítulo apresenta as definições do modelo multidimensional para documentos de textos. Essas definições serão utilizadas como suporte para o algoritmo de recomendação. A arquitetura do sistema é descrita para a compreensão do modelo proposto, o POQT (*Personalized OLAP Queries for Text*), bem como detalhes sobre a operação de agregação OLAP textual que pode ser personalizada. As possibilidades de expansão da consulta OLAP também serão abordadas no final deste capítulo.

4.1 Definições do Modelo Conceitual

As operações OLAP se tornaram ao longo dos anos a principal ferramenta para a exploração de dados de um *Data Warehouse*. A *modelagem multidimensional* é a técnica estruturada usada para obter modelos de dados de simples entendimento e alto desempenho de acesso aos dados, sendo o modelo *estrela* um dos mais usuais.

A tabela fato do modelo estrela é composta por um conjunto de medidas ou fatos. As medidas são tradicionalmente numéricas e podem ser aditivas (como a função `SUM`), semiaditivas (como as funções `AVG`, `MIN` e `MAX`) e genéricas (como as funções `COUNT` e `LIST`). Dependendo do tipo de medida encontrada em uma dimensão nem todas as funções de agregação OLAP podem ser usadas para análises específicas [RAVAT et al., 2008].

No entanto, a análise de dados textuais exige medidas textuais que se enquadrem nas categorias não numéricas e não aditivas. Em [RAVAT; TESTE; TOURNIER, 2007], os autores apresentam uma função de agregação de palavras-chave que permite a agregação de dados textuais em ambientes OLAP, justamente como funções tradicionais fazem com dados numéricos. A função *AVG_KW* utiliza uma ontologia para agrupar palavras-chave a partir de uma palavra-chave que possa definir ou representar essas palavras.

O estudo de caso apresentado neste trabalho baseia-se na ontologia utilizada pela PubMed e em um conjunto de artigos extraídos dessa base de dados para ilustrar gradualmente os conceitos e realizar experimentos em relação ao algoritmo proposto. O contexto ao qual se insere o estudo de caso é o da PubMed descrito no Apêndice A.

O OLAP textual requer um modelo multidimensional adequado para armazenar dados textuais. O modelo POQT estende o tradicional modelo multidimensional para representar dados textuais, novos tipos de medidas e dimensões, como descritos a seguir:

No modelo multidimensional proposto existem *medidas* numéricas e não numéricas. Uma *medida não-numérica* deverá ser encaixar em uma das categorias abaixo:

Definição 1.1: uma *medida bruta* é a medida que corresponde a qualquer tipo de dados não estruturados.

Definição 1.2: uma *medida de descrição* é a medida que descreve o conteúdo de um dado não estruturado por meio de uma representação estruturada.

No caso de dados textuais, dizemos que um documento é uma *medida não numérica bruta*. Por exemplo, os documentos têm atributos de metadados e palavras-chave extraídas a partir do texto. Estes tipos de atributos podem ser uma *medida não numérica de descrição*. Para determinar o que representa uma dimensão, define-se que:

Todo aspecto que caracterize um objeto de análise é considerado uma *dimensão*, mesmo essa seja modelada como uma medida em outro cenário [PÉREZ-MARTÍNEZ et al., 2008].

Esta definição é genérica para englobar qualquer tipo de perspectiva para analisar um fato. Além disso, foi definido um tipo especial de dimensão sem interferir na definição anterior.

Definição 2: a *dimensão hierárquica* é um tipo de dimensão que contém uma descrição de um fato, como ocorre com as palavras-chave pertencentes a uma representação hierárquica de conceitos.

As hierarquias podem ser definidas pelo usuário ou podem ser hierarquias de domínio específico, como a árvore MeSH da Biblioteca Nacional de Medicina dos Estados Unidos [NATIONAL LIBRARY OF MEDICINE, 2014].

É necessário observar que é possível que haja mais do que uma dimensão hierárquica, já que os dados não estruturados podem ter múltiplas descrições para diferentes características e pontos de vista. A dimensão hierárquica precisa ser ligada aos conceitos de uma representação hierárquica.

Definição 3: uma *representação hierárquica* é uma estrutura de dados de árvore, O, relacionada a uma árvore n-ésima marcada com uma raiz T, para o qual cada nó tem no máximo n filhos e um rótulo único que pertence a um vocabulário de domínio. Formalmente, é composta por um quádruplo $O = (T, D, V, R)$, onde D é o universo de

discurso, V é o vocabulário que descreve D e R um conjunto de relações conceituais entre os nós v_i e v_j no espaço de domínio $< D, V >$, onde v_i, $v_j \in V$.

Assim, se uma palavra-chave (descritor) em uma dimensão hierárquica é representado por KW_i então $KW_i \in V$. A representação hierárquica pode ser uma ontologia de domínio, taxonomia ou qualquer outra representação de conceitos sob a forma de uma estrutura de árvore.

As relações entre os conceitos são modeladas por intermédio de uma árvore. Cada nó da árvore representa um descritor de domínio e cada aresta representa o relacionamento de especialização e generalização entre os descritores. O conteúdo dos documentos são expressos pelo vocabulário de descritores de uma representação hierárquica de domínio. Como é o caso, por exemplo, de dois documentos, onde um possui como descritor o termo *'bacteremia'* e outro possui o termo *'pneumonia bacteriana'*, ao realizar a junção entre esses termos há um generalização dos mesmos, ou seja, os termos podem ser definidos por *'infecções bacteriais'*, que é o assunto geral de ambos os artigos.

Com o objetivo de calcular a relação entre duas palavras-chave, como [RAVAT; TESTE; TOURNIER, 2007], utiliza-se o conceito do LCA, que corresponde ao menor antecessor comum.

Definição 4: o *Lowest Common Ancestor* (LCA) entre dois nós v e w em O é expresso por $LCA_O(v, w)$. O LCA é uma função que retorna $v_{LCA} \in V$, o antecessor comum de maior profundidade entre v e w, isto é, o primeiro nó que é um antepassado comum de ambos os nós, v e w. Esta função retorna o nó v ou w no caso de um ser o antecessor do outro.

Definição 5: a *distância* entre v e w em O, denotado por $d_O(v, w)$, é uma função recursiva que retorna o maior número de nós entre o $LCA_O(v, w)$ e v ou w, isto é,

$$d_O\texttt{(v,w)= max(}d_O\texttt{(v,}LCA_O(v,w)\texttt{),}\ d_O\texttt{(w,}LCA_O(v,w)\texttt{))}.$$

Dessa forma a $d_O(v, w)$ retorna a maior distância (maior número de nós) entre o último antecessor comum e um dos dois nós, no caso v ou w. Se a distância entre v e $LCA_O(v, w)$ for maior que a distância entre w e $LCA_O(v, w)$, então está será a distância calculada (d_O), caso contrário a distância entre w e $LCA_O(v, w)$ será a maior e corresponderá ao valor de d_O.

No estudo de caso proposto, uma representação hierárquica foi construída a partir de sub-árvores MeSH com uma raiz artificial `AAA`. Na Figura 6 é possível visualizar as últimas definições do modelo proposto, pois mostra uma linha de hierarquia MeSH (O_{MeSH}). Os

documentos são os artigos científicos da PubMed e a dimensão hierárquica contém os termos MeSH que os descrevem. Então, V é composto pelos termos MeSH e R contém as relações entre eles.

Para simplificar os cálculos de cada termo, a base de dados da PubMed possui um rótulo associado a cada termo, o qual descreve o seu caminho a partir da raiz. Na ontologia utilizada na PubMed existe a possibilidade do mesmo termo pertencer a mais de uma subárvore, podendo então ter um ou mais rótulos de caminho (posições) para o mesmo termo MeSH.

Para exemplificação é considerado que os termos a seguir só tenham um rótulo correspondente, no caso do termo MeSH *'Meningitis, Bacterial'*, ele é rotulado com `AAA.C01.252.200.500`, o que significa que seu nó está no quinto nível da árvore, contando a partir da raiz, e seu pai é o nó rotulado como `AAA.C01.252.200`, o *'Central Nervous System Bacterial Infections'*.

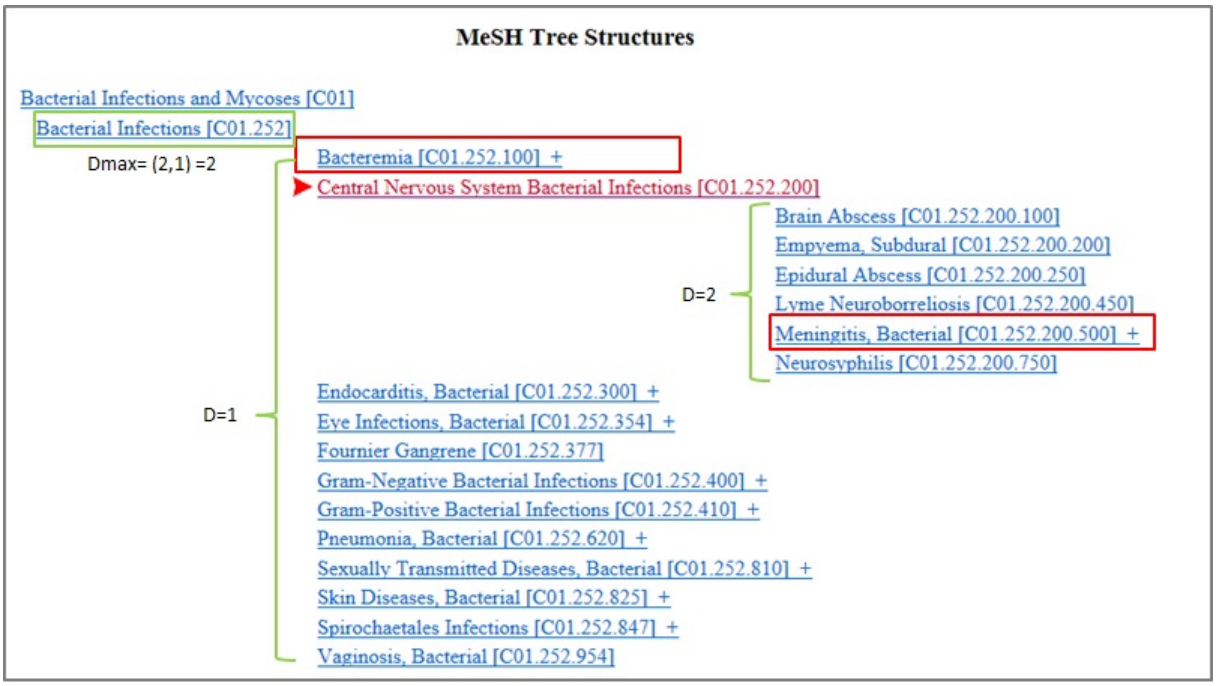


Figura 6 – Árvore MeSH - *Bacterial Infections*

Supondo-se que um primeiro documento possui como termo MeSH o *'Meningitis, Bacterial'* e um segundo documento é descrito pelo termo *'Bacteremia'*. O menor antecessor comum entre esses dois descritores é v_{LCA} = *'Bacterial Infections'*. Em relação ao domínio da ontologia sua posição na subárvores *Diseases* `[C]` são respectivamente: `AAA.C01.252.200.500` e `AAA.C01.252.100`, logo os termos MeSH podem ser agregados de acordo com o seu último antecessor comum, o nó `AAA.C01.252`, que corresponde ao termo *'Bacterial Infections'* (Figura 6). A distância entre eles é calculada da seguinte forma:

d_{MeSH}({Meningitis, Bacterial},{Bacteremia})
= max (d_{MeSH}({Meningitis, Bacterial},{Bacterial Infections}),
d_{MeSH}({Bacteremia},{Bacterial Infections}))
= max (2,1) = 2.

Isto é, o termo *'Bacteremia'* é um nó filho direto de *'Bacterial Infections'* então d_{MeSH}(Bacteremia,Bacterial Infections)= 1 e o *Meningitis, Bacterial* está a uma distância de dois nós do nó LCA, logo d_{MeSH}(Meningitis, Bacterial,Bacterial Infections)= 2. Neste caso, a distância (d_{MeSH}) será 2, que corresponde a maior distância calculada.

Para facilitar o entendimento, neste exemplo são relacionados dois documentos e cada um é descrito por apenas um termo MeSH. Porém, cada documento armazenado no *Data Warehouse* pode possuir um ou mais termos MeSH que destacam os principais assuntos abordados pelos autores.

4.2 *Framework* da Recomendação

Para realizar o OLAP textual, é proposta a integração semântica entre termos textuais para permitir a tomada de decisão centrada no usuário. O uso de ontologias permitirá agregações semânticas bem-sucedidas. Além disso, para garantir uma decisão eficaz centrada no usuário, é fornecida a personalização da agregação com base em parâmetros ajustáveis.

Para que se atinjam resultados significativos no processo de recomendação é necessário que dois tipos de usuários distintos realizem interações com o sistema, os usuários são: o *analista* e o *gestor* da área de domínio. O analista é um especialista na árvore ontológica e o profissional capaz de realizar a configuração da função responsável pela recomendação. Já o gestor da área de domínio é o tomador de decisão que se beneficiará da recomendação de documentos operando a *interface de recomendação*.

Para expressar consultas multidimensionais, o usuário gestor deve especificar os atributos da dimensão para obter resultados que satisfaçam as suas consultas. Uma vez que o resultado da consulta é apresentado, o usuário pode explorá-lo aplicando operações OLAP. Para auxiliar os usuários nesse passo, no contexto da operação de agregação do OLAP textual, é proposto um algoritmo de recomendação com base no algoritmo de agregação descrito em [RAVAT; TESTE; TOURNIER, 2007], que inclui a especificação das necessidades ou preferências do utilizador.

O *framework* geral de agregações OLAP é composto por duas fases: a fase de pré-processamento e a fase do OLAP textual interativo. O pré-processamento da recomendação

é realizado utilizando as escolhas do usuário analista e a hierarquia de conceitos. Já o OLAP textual é utilizado para a recomendação de documentos para o usuário gestor. O conjunto de documentos (*conjunto de itens*) inclui os documentos em conformidade com os parâmetros já definidos e a consulta multidimensional realizada. A partir desse ponto, o processo se torna dinâmico, ou seja, se adequando à necessidade de cada usuário.

4.2.1 Parâmetros da Personalização de Agregações OLAP

Os parâmetros para agregação são essenciais para permitir que os melhores resultados sejam obtidos a partir da recomendação semântica de documentos candidatos. Os documentos a serem recomendados podem ser submetidos a um novo nível de granularidade na hierarquia dimensão para executar operações de agregação sempre que necessário.

Dessa forma, os usuários analistas, especialistas na árvore ontológica, podem definir os parâmetros do algoritmo até que o conjunto de documentos candidatos corresponda ao objetivo da consulta inicial do usuário gestor. Antes de iniciar a descrição dos algoritmos, existem alguns termos relacionados à parametrização do pré-processamento e ao processo de recomendação que devem ser definidos. Esses parâmetros serão responsáveis pela determinação da similaridade de dois documentos. O termo similaridade denota uma maneira de mostrar computacionalmente, a partir de um valor, o quanto dois objetos, neste caso documentos, são semelhantes entre si.

Definição 6: o *fator de ontologia*, D_{MAX}, é a distância máxima que é considerada para concluir que dois nós são semanticamente relacionados na representação hierárquica de conceitos. O *fator de ontologia* é definido pelo usuário analista, ou seja, o especialista na ontologia.

O usuário analista, que está familiarizado com o conceito de hierarquia utilizado, ajusta o valor de D_{MAX} na parametrização da agregação. Espera-se que ele possua uma noção da similaridade entre os documentos, pois seu conhecimento da ontologia permite que as recomendações recuperadas tenham maior precisão.

Na ausência do usuário analista, o valor de cada parâmetro foi obtido de forma experimental, como realizado nesse trabalho e descrito no Capítulo 5. Os melhores valores para a parametrização da recomendação do estudo de caso foram verificados de forma experimental pelos professores especialistas.

No caso da árvore MeSH, o número máximo de D_{MAX} é 12, uma vez que esta é a altura da árvore. O valor mais restritivo é o zero, caracterizando uma correspondência exata de palavras-chave (termos na ontologia). Portanto, o D_{MAX} é uma métrica de similaridade utilizada para determinar a similaridade mínima entre palavras-chave dos documentos.

Para estabelecer a similaridade entre dois documentos que são descritos por várias palavras-chave cada um, foi utilizada uma medida objetiva que envolve a análise estatística dos dados, a frequência.

Definição 7: um *evento* é a comparação que ocorre entre palavras-chave no contexto desse modelo multidimensional proposto.

Definição 8: a *frequência* de um evento i é o número n_i de vezes que o evento ocorreu em um experimento ou estudo.

Definição 9: a *frequência relativa* (f_i) é, portanto, a frequência normalizada pelo número total de eventos (N):

$$f_i = n_i/N,$$

onde $0 \leq f_i \leq 1$.

Nesta proposta, um evento é uma combinação de duas palavras-chave, kw_i e kw_j, que satisfaz o parâmetro de D_{MAX} do pré-processamento, onde kw_i descreve $D_{DocumentoAlvo}$ e kw_j^d descreve $D_d \in$ *Conjunto de itens*. O número total de eventos N é o número total de combinações. O conceito básico da similaridade é receber um par de objetos e retornar uma pertinência entre [0,1] que indica o quão similares são os parâmetros recebidos.

Definição 10: a *similaridade mínima* é um valor ou métrica definido pelo usuário para estabelecer a frequência mínima utilizada na recomendação. O algoritmo considera apenas os eventos com pontuações superiores a este valor.

A *similaridade mínima* é uma métrica de similaridade utilizada para determinar a semelhança mínima entre os documentos. Para o usuário, isso significa estabelecer um limite mínimo de semelhança entre dois documentos que corresponda às suas necessidades e preferências. Por exemplo, 75% de similaridade mínima implica em uma frequência $f_i \geq 0,75$, o que significa que para a recomendação ocorrer, pares de palavras-chave dos dois documentos são semanticamente relacionados pelo menos 75% das vezes.

Considera-se as duas palavras-chave, kw_i e kw_j^d, semanticamente relacionadas se a distância calculada por $d_O(kw_i, kw_j^d)$ é inferior ou igual a um dado valor de D_{MAX}. Se o usuário analista definir a *similaridade mínima* como zero significa que qualquer possibilidade de agregação será aceitável, ou seja, todos os documentos da base serão considerados como documentos recomendados. Desta forma, se a *similaridade mínima* foi

definida de forma coerente obtêm-se o *conjunto de itens* de documentos obtidos a partir do *documento alvo*.

4.2.2 Definições Complementares para a Recomendação de Documentos

Para iniciar o processo de recomendação de documentos, o usuário analista deve apontar um documento existente no *Data Warehouse*.

Definição 11: o *documento alvo* é o documento selecionado para realizar a recomendação. Esse documento é escolhido pelo usuário gestor para iniciar o processo de recomendação, ou seja, é o documento que se tem o interesse de conhecer quais seriam os seus documentos recomendados. O *documento alvo* selecionado pode ser qualquer documento armazenado no *Data Warehouse* que tenha sido submetido ao pré-processamento da recomendação.

O usuário gestor define o *documento alvo* no intuito de conhecer outros documentos, que estão armazenados no *Data Warehouse* e que possuem semelhanças mínimas de acordo com os parâmetros definidos na personalização e a consulta multidimensional realizada.

Definição 12: um *conjunto de itens* é uma lista de todos os documentos que são recomendados a partir do *documento alvo*. Todos os documentos candidatos são pré-processados e um conjunto específico de documentos é obtido a partir da verificação da similaridade com o *documento alvo*.

Para obter melhores resultados, o usuário gestor, pode optar por definir dimensões em sua consulta. Nesse caso, o *conjunto de itens* além de satisfazer os parâmetros de *similaridade mínima* e D_{MAX}, considera os outros atributos incorporados a consulta pelo usuário gestor.

A seguir é apresentada a arquitetura do modelo POQT, ou seja, descrevendo os detalhes sobre as etapas do processo de recomendação e como são utilizados os parâmetros da personalização de agregações OLAP neste sentido.

4.3 Arquitetura

Uma maneira de analisar um conjunto de documentos é conferir quais são palavras-chave que os descrevem, isto é, os descritores relacionados com o assunto tratado. As palavras-chave permitem que o tomador de decisão possa ter uma visão sobre os principais temas tratados em um conjunto de artigos científicos, seja do mesmo autor ou não, e a partir destes dados extrair tendências em relação à área ou tema desses artigos.

Caso os autores sejam analisados de forma individual pode-se inferir qual a sua área de atuação ou interesse. Ao passo que se conhece um artigo pode-se a partir dele verificar quais são os seus artigos recomendados, a fim de aprofundar seu conhecimento nessa área de pesquisa. A possibilidade de personalizar como serão recuperados os artigos recomendados, pode trazer mudanças significativas no conjunto de artigos sugeridos, já que este será mais adequado as expectativas do tomador de decisão. Esses são alguns exemplos de como a recomendação semântica de documentos pode auxiliar o tomador de decisão por meio da personalização de agregações OLAP.

Utilizando as definições do modelo POQT para realizar o processo de agregação e recomendação, foi elaborada a arquitetura do sistema, como mostrada na Figura 7. Como o trabalho visa a recomendação semântica de documentos, o primeiro passo consiste em armazenar as informações dos documentos dentro da estrutura do modelo lógico proposto. O pré-processamento dos dados, neste caso, envolve as seguintes entradas:

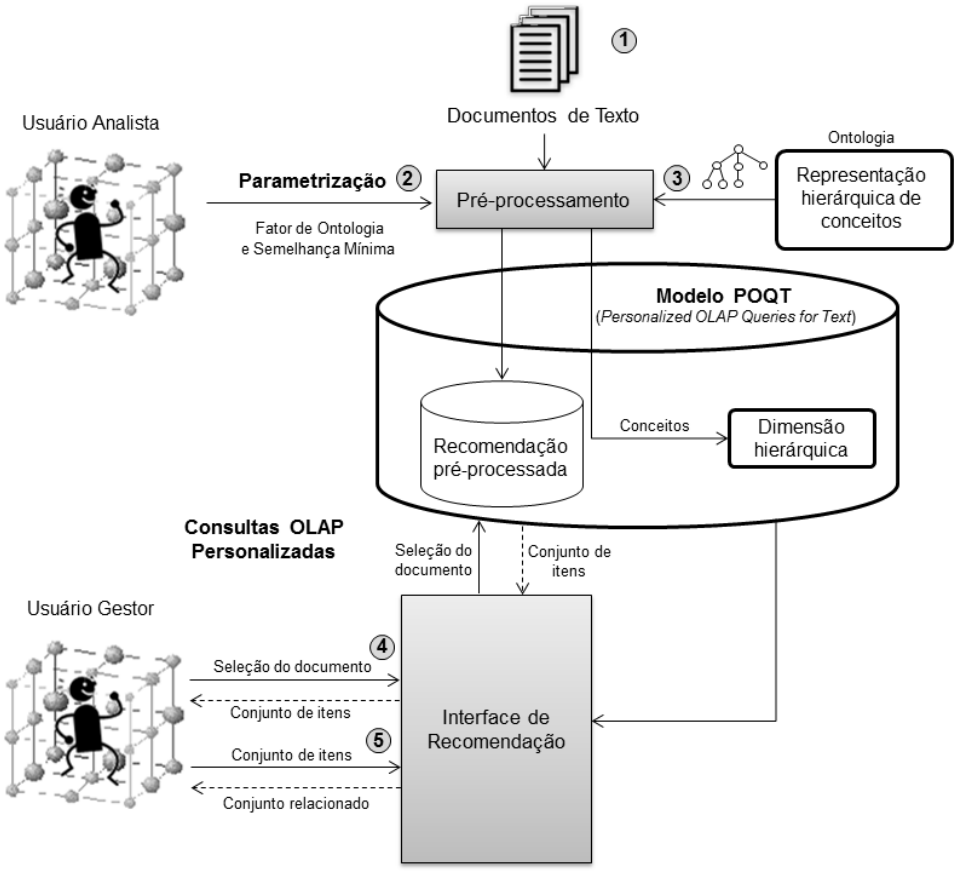


Figura 7 – Arquitetura do modelo POQT

1. *Armazenamento dos documentos de texto*: são realizadas operações de ETL para que os conceitos relacionados a cada documento sejam corretamente indexados dentro do *Data Warehouse* (item 1 da Figura 7). O processo visa selecionar os dados,

dimensões específicas de cada documento e inserí-los de forma correta no sistema de armazenamento, neste caso denominado de *Modelo POQT*.

2. *Armazenamento da representação hierárquica de conceitos*: a *ontologia* de domínio deve ser utilizada para permitir o armazenamento correto das informações do *Data Warehouse*, pois os dados relativos ao assunto dos documentos, as palavras-chave, são organizadas dentro desta *dimensão hierárquica* (item 2 da Figura 7).

3. *Parametrização do pré-processamento*: uma vez que os dados relativos aos documentos e representação hierárquica já estão armazenados no *Data Warehouse*, é necessário definir qual será a parametrização adequada para a determinada ontologia (item 3 da Figura 7). A *similaridade mínima* é um parâmetro considerado no pré-processamento, assim como D_{MAX}. Esses parâmetros são definidos pelo analista e são responsáveis pela configuração do algoritmo de recomendação. No entanto, apenas a *similaridade mínima* pode ser alterada na execução das consultas, ou seja, o usuário gestor pode buscar no momento da consulta os documentos recomendados que possuam um valor de *similaridade mínima* maior do que o valor da parametrização.

O conjunto de documentos armazenados é submetido a uma análise e cálculos de distância para a extração de valores relacionados a similaridade entre os documentos. Para isso, são verificadas as palavras-chave dos diversos documentos e a partir da função de agregação, é possível verificar suas semelhanças navegando pela árvore ontológica. Dessa forma, após pré-processar os documentos de texto armazenados, obtém-se a *recomendação pré-processada*, a qual também é armazenada no *Modelo POQT*.

O segundo passo em relação a recomendação de documentos é o OLAP interativo, o qual permite consultas OLAP personalizadas. Esse processo ocorre a partir das entradas do usuário gestor, onde a sua primeira interação com a *interface de recomendação* consiste em selecionar um documento disponível no *Data Warehouse*. Por meio do documento selecionado, é verificado junto a *recomendação pré-processada* quais os documentos serão recomendados para o determinado documento.

A partir do momento em que o *conjunto de itens* foi definido e mostrado na *interface de recomendação*, o gestor tem a possibilidade de realizar duas ações distintas que podem ou não ocorrer concomitantemente. Essas ações permitem a personalização do resultado da recomendação de acordo com o interesse do gestor. Essas ações envolvem:

1. Consulta OLAP por intermédio do *cubo de consulta* e *similaridade mínima*: o usuário gestor pode refinar o *conjunto de itens* a partir da escolha de um cubo de consulta disponível na *interface de recomendação* (item 4 da Figura 7). Esses *cubos de decisão* são baseados nas dimensões dos documentos armazenados no *Data Warehouse* e o valor de *similaridade mínima* corresponde ao mesmo parâmetro da parametrização.

Para que a consulta retorne documentos com um maior nível de semelhança, o gestor pode aumentar o valor de *similaridade mínima* para obter uma consulta mais refinada. O *cubo de consulta* e a *similaridade mínima* podem ser separadamente utilizados em uma consulta ou utilizados na mesma consulta.

2. Expansão da recomendação para obter *documentos relacionados*: após obter o *conjunto de itens*, a partir da seleção do documento ou após a submissão do *conjunto de itens* ao *cubo de consulta*, existe uma forma expandir os resultados da recomendação (item 5 da Figura 7). A possibilidade de estender os documentos relacionados é interessante quando o usuário gestor deseja conhecer os documentos recomendados a partir dos conjuntos de documentos já definidos (*conjunto de itens*). Dessa forma a *interface de recomendação* retorna para o gestor o conjunto chamado de *conjunto relacionado*.

Por meio das duas ações acima citadas, o usuário gestor é capaz de: (1) reduzir o resultado da sua consulta, trazendo documentos mais estritamente recomendados, ou (2) expandir a consulta trazendo documentos que são também relacionados ao documento recomendado. No Capítulo 4 são mostradas como essas ações podem ser realizadas na interface do protótipo, a qual permite a recomendação de documentos.

É importante salientar que uma vez que a configuração foi realizada pelo usuário analista e os dados enviados para a base de *recomendação pré-processada*, não se torna necessário processar novamente a recomendação. Caso o gestor que está utilizando a *interface de recomendação* tenha a necessidade de realizar um refinamento em relação a recomendação, essa ação é permitida por intermédio do *cubo de consulta* e variação do valor de *similaridade mínima* maior do que o valor expresso na parametrização. Porém, caso seja necessário alterar o valor de D_{MAX} ou diminuir o valor *similaridade mínima*, um novo pré-processamento deverá ser realizado.

4.4 Algoritmos para Agregação e Recomendação

O objetivo do algoritmo de recomendação é calcular os valores de similaridade entre os documentos armazenados no *Data Warehouse*, ou seja, destina-se a obter a *recomendação pré-processada*. Após essa rotina de pré-processamento, torna-se possível recomendar documentos semelhantes a um documento selecionado pelo usuário, sem que necessite esperar o processamento completo das comparações em busca de documentos similares. Portanto, para selecionar o *documento alvo* e buscar por seus documentos recomendados, *conjunto de itens*, é necessário aplicar o Algoritmo 1, para obter o pré-processamento das recomendações.

O algoritmo de agregação *AVG_KW* (Seção 3.1) tem como saída uma lista de palavras-chave resultante da agregação. Para atingir o objetivo da recomendação proposta, no entanto, não se torna interessante saber o termo resultante da junção das palavras-chave, mas sim saber se os documentos são semelhantes ou não, e assim efetuar a recomendação. Dessa maneira, este trabalho não se restringe à agregação, nesse caso ela é o apoio da recomendação semântica. Se a condição que verifica a distância calculada por meio do LCA é menor ou igual D_{MAX} é verdadeira, existe um contador para indicar a quantidade de vezes em que o evento da semelhança entre os termos ocorreu.

O Algoritmo 1 - **Generate_recommendation** tem a finalidade de encontrar o *conjunto de itens* (*itemset*) de um *documento alvo* armazenado na base de dados. Todos os documentos candidatos são avaliados para se obter o índice que indica o seu potencial de relação semântica.

Para encontrar o valor de frequência de cada par de documentos, a princípio, são armazenadas as palavras-chave do documento alvo (*getKeywords(targetdoc)*) na lista *target_lKWs* e o respectivo tamanho de elementos na lista é armazenado na variável *ntarget*. Na sequência, inicia-se o processo de comparação entre o *documentos alvo* (*targetdoc*) e todos os documentos armazenados na base de documentos candidatos (*Candidates*).

Algoritmo 1: Generate_recommendation

Entrada: *targetdoc*, *Candidates*, D_{MAX}, *min_similarity*

Saída: *itemset*

```
// Todo documento deve ser submetido a este processo
itemset ← {}
target_lKWs ← getKeywords(targetdoc)
ntarget ← sizeOf(target_lKWs)
FOR EACH candidate IN Candidates
    lKWs ← getKeywords(candidate)
    itemcount ← 0
    FOR EACH KW_i IN target_lKWs
        FOR EACH KW_j IN lKWs
            listKW_i ← getVariantsKeywords (KW_i)
            listKW_j ← getVariantsKeywords (KW_j)
            IF Calculate_distanceLCA (listKW_i, listKW_i, D_MAX)
                itemcount++
            END IF
        END FOR
    END FOR
    fdoc = Frequency_doc (itemcount, ntarget, sizeOf(lKWs))
    IF ( fdoc ≥ min_similarity)
        itemset ← itemset ∪ candidate
        STORE (targetdoc, candidate, fdoc)
    END IF
END FOR
```

As palavras-chave de cada documento candidato são extraídas (*getKeywords(candidate)*) e armazenadas na lista *lKWs*. Para cada palavra-chave do documento alvo (KW_i) e para cada palavra-chave do documento candidato (KW_j) é necessário extrair uma lista de nós que

possuem a mesma descrição da palavra-chave na árvore ontológica (*getVariantsKeywords(*KW_i*)* e *getVariantsKeywords(*KW_j*)*).

A descrição da palavra-chave pode ocorrer apenas uma vez na árvore ontológica, ou seja, quando ela possui apenas um rótulo de caminho na ontologia toda. Se essa for a situação do termo MeSH, a lista de nós que contém a mesma descrição é composta de apenas um rótulo de caminho. Caso o mesmo termo pertença a subárvores diferentes, todos os seus diferentes rótulos de caminhos (posições) são armazenados nessa lista.

A função *Calculate_distanceLCA* recebe as duas listas de nós que correspondem a todas as ocorrências do mesmo termo MeSH na árvore ontológica e toda vez que na execução do código o parâmetro de configuração D_{MAX} (*fator de ontologia*) for satisfeito, tem-se a indicação que o par de documentos é similar.

Para verificar que os documentos são semelhantes, a função *Frequency_doc* recebe como entrada: os valores das comparações entre palavras-chave que eram menores que D_{MAX} (*itemcount*), a quantidade de palavras-chave do documento alvo (*ntarget*) e do documento candidato (*sizeOf(lKWs)*). Assim a função pode efetuar o cálculo que revelará se a frequência mínima que define que dois documentos são semelhantes foi encontrada.

Caso seja verificado que um documento candidato é semelhante ao *documento alvo* sua frequência (*fdoc*) é armazenada juntamente com a identificação do par de documentos (*targetdoc* e *candidate*) na base destinada as recomendações já pré-processada.

Os algoritmos a seguir são utilizados no Algoritmo 1, são eles:

- Algoritmo 2 - **Calculate_distanceLCA**;
- Algoritmo 3 - **Check_distanceLCA** e
- Algoritmo 4 - **Frequency_doc**.

A função descrita no Algoritmo 2 tem o objetivo de percorrer as listas de posições ($List\_KW_{Target}$ e $List\_KW_{Candidate}$) entre as palavras-chave do documento alvo ($KWt_{specific}$) e um documento candidato ($KWc_{specific}$). Caso o valor encontrado seja menor ou igual ao D_{MAX}, o retorno é verdadeiro, caso contrário, é falso.

Algoritmo 2: Calculate_distanceLCA

Entrada: $List\_KW_{Target}$, $List\_KW_{Candidate}$, D_{MAX}

Saída: *boolean*

```
FOR EACH KWt_specific in List_KW_Target
    FOR EACH KWc_specific in List_KW_Candidate
        IF ((Check_distanceLCA (KWt_specific, KWc_specific, D_MAX, 0)) ≤ D_MAX)
            return true
        END IF
    END FOR
END FOR
return false
```

A distância do LCA é verificada na função do Algoritmo 3. A primeira verificação é realizada para saber se no conjunto de posições de uma palavra-chave do documento alvo ($KW\_setTarget$) é igual a uma das posições da palavra-chave do documento candidato ($KW\_setCandidate$). Desta forma, é possível verificar se em alguns destes termos das listas (KW_{Target}) e ($KW_{Candidate}$) existe algum nó que possui a mesma descrição. Se for o caso, ocorreu a correspondência dos termos MeSH.

Na primeira vez que esta função é chamada, o $distance\_level$ é igual a 0, pois esse é o valor que foi definido inicialmente como zero no Algoritmo 2. O $distance\_level$ indica a distância entre as palavras-chave na árvore e inicia-se com zero para fazer a verificação direta entre os termos, ou seja, nesta primeira chamada cada lista possui apenas uma posição do termo (KW_{Target} e $KW_{Candidate}$) e portanto compara exatamente as descrições entre estas duas palavras-chave. Caso o retorno da comparação seja verdadeiro, significa que não há distância entre os termos e por consequência eles são termos iguais.

Caso os termos comparados não sejam iguais, o próximo passo será inserir em cada conjunto o nó pai correspondente a cada termo. Sendo assim, é adicionado ao conjunto de último ancestral de cada *documento alvo* ($lastKW\_setTarget$) e de cada documento candidato ($lastKW\_setCandidate$), a sua palavra-chave ancestral, ou seja, o nó acima na ontologia ($KW\_setTarget$.Last() e $KW\_setCandidate$.Last()). Se o $distance\_level$ atual for menor que a distância máxima definida (D_{MAX}) e se o último nó ancestral de cada conjunto não for a raiz (*root*) da árvore, o $distance\_level$ é incrementado e passa-se a chamar recursivamente a função para comparar se os pais de cada nó são iguais, ou se um dos nós comparados é o pai do outro.

A condição que verifica se o $distance\_level$ é menor que D_{MAX} tem o objetivo de continuar a execução da função enquanto essa distância ainda não ultrapassou o limite definido pelo D_{MAX} na parametrização da recomendação no Algoritmo 1. Se já ultrapassou o limite, o valor é retornado para o Algoritmo 2.

Em síntese, a função compara recursivamente as palavras-chave e suas ancestrais até encontrar duas palavras-chave iguais ou chegar na raiz da ontologia. Ao terminar, retorna a distância entre as duas palavras-chave, a qual será comparada ao D_{MAX} para avaliar se é válida ou não.

Algoritmo 3: ***Check_distanceLCA***

Entrada: $KW\_setTarget$, $KW\_setCandidate$, D_{MAX}, $distance\_level$

Saída: $distance\_level$

```
// Verifica se as palavras-chave são iguais
FOR EACH KW_Target in KW_setTarget
    FOR EACH KW_Candidate in KW_setCandidate
        IF (KW_Target.description == KW_Candidate.description) // the same term!
            return level
        END IF
    END FOR
END FOR
//Obter a palavra-chave antecessora para verificar se não é a raiz (root)
lastKW_setTarget = KW_setTarget.Last();
lastKW_setCandidate = KW_setCandidate.Last();
//Expandir o nível para encontrar a próxima palavra-chave antecessora
IF(distance_level ≤ D_MAX && (lastKW_setTarget ≠ root || lastKW_setCandidate ≠ root))
    distance_level++
    IF (lastKW_setTarget ≠ root)
        KW_setTarget.Add(getAncestorKw(lastKW_setTarget));
    END IF
    IF (lastKW_setCandidate ≠ root)
        KW_setCandidate.Add(getAncestorKw(lastKW_setCandidate));
    END IF
    return Check_distance(KW_setTarget, KW_setCandidate, D_MAX, distance_level)
ELSE
    return distance_level
END IF
```

O próximo algoritmo, o Algoritmo 4, é chamado diretamente no Algoritmo 1. Essa função realiza o cálculo da semelhança entre dois documentos. A frequência é calculada para cada documento do *conjunto de itens*. Após o cálculo, no Algoritmo 1, o seu valor é comparado com o parâmetro definido pelo usuário final, a *similaridade mínima*. Se a condição for verdadeira, o documento candidato torna-se uma recomendação.

Algoritmo 4: ***Frequency_doc***

Entrada: *itemcount*, *doc1_cardinality*, *doc2_cardinality*

Saída: frequency

```
// Cada documento possui pelo menos uma palavra-chave associada
// Cálculo da frequência (similaridade)
frequency = itemcount / (doc1_cardinality * doc2_cardinality)
return frequency // para comparação com o valor do usuário
```

Durante a consulta, a *similaridade mínima* também pode ser utilizada para refinar as *recomendações pré-processadas*. Neste caso as *recomendações pré-processadas* recebem um cubo de consulta que permite mostrar apenas os documentos recomendados que satisfizerem o valor determinado. Ao aumentar o valor da *similaridade mínima* em relação ao definido no pré-processamento, se restringe o conjunto de documentos retornados pela consulta.

Com o intuito de exemplificar o processo de agregação e consequente recomendação, o Artigo 1 é o *documento alvo* e o Artigo 2 é um documento candidato armazenado no

Data Warehouse. Os termos MeSH que descrevem cada artigo são:

Artigo 1 - *'Meningitis, Bacterial'*, *'Tuberculosis, Meningeal'* e *'Otitis'*
Artigo 2 - *'Bacteremia'* e *'Meningitis, Pneumococcal'*

Para a recomendação foi definido, por exemplo, que a *similaridade mínima* será de 0,75 e $D_{MAX} = 4$. Como a combinação uma a uma de palavras-chave dos dois artigos resulta em seis possibilidades diferentes de combinação, o número total de eventos para o cálculo da *frequência relativa* será de $N = 6$.

Para calcular a quantidade de vezes em que a combinação das palavras-chave foi bem sucedida (n_i), ou seja, satisfazendo o critério de D_{MAX}, é necessário avaliar as combinações uma a uma por intermédio do algoritmo de recomendação. A distância é calculada em cada combinação de termos, como mostrado abaixo:

```
Combinação 1:{Meningitis, Bacterial} e {Bacteremia}
Combinação 2:{Meningitis, Bacterial} e {Meningitis, Pneumococcal}
Combinação 3:{Tuberculosis, Meningeal} e {Bacteremia}
Combinação 4:{Tuberculosis, Meningeal} e {Meningitis, Pneumococcal}
Combinação 5:{Otitis} e {Bacteremia}
Combinação 6:{Otitis} e {Meningitis, Pneumococcal}
```

As palavras-chave dos artigos, no caso os termos MeSH dos artigos na PubMed, são comparadas a fim de verificar se elas possuem uma distância na ontologia que seja menor ou igual a distância máxima pré determinada, D_{MAX}. Cada termo MeSH pode ocorrer em mais de uma subárvore e é por esta razão que é verificado a cada posição daquele termo MeSH se alguma distância calculada é válida dado ao valor máximo permitido. O objetivo é combinar todos os termos MeSH de documentos diferentes para realizar a agregação.

A *frequência* obtida para saber se o Artigo 2 é parte do *conjunto de itens* é 0,83, pois a quantidade de combinações que satisfazem o critério da distância é 5 em um total de 6 combinações. Logo esta *frequência* é superior ao valor estipulado de *similaridade mínima* (0,75), portanto o Artigo 2 é um artigo recomendado para o *documento alvo* (Artigo 1). Desta forma, pode-se afirmar que todo o documento que satisfizer a condição de que $f_i \geq$ *similaridade mínima* estará no *conjunto de itens*. No caso apresentado, apenas o último cálculo de distância (Combinação 6) obteve um resultado que extrapola o valor máximo, definido como 4, já que as distâncias máxima em relação aos nós é 5 porque os termos MeSH estão muito distantes entre si na árvore ontológica.

4.4.1 Expansão dos Resultados da Recomendação

A expansão da recomendação pode ser utilizada quando o gestor deseja conhecer mais documentos que são relacionados ao *documento alvo*, ou seja, quando se tem a necessidade de aumentar a abrangência do assunto para uma melhor análise. A expansão é opcional ao usuário gestor, isto é, o gestor tem o controle do que será exibido na *interface de recomendação*. Quando a exibição dos *documentos relacionados* não é requisitada, é exibido apenas o conjunto dos documentos recomendados.

Definição 13: a *expansão da consulta* é o procedimento utilizado para encontrar documentos que possam ser recomendados a partir dos documentos que compõem o *conjunto de itens*. A possibilidade de aumentar os documentos recomendados é justamente capturar outros documentos que podem ser relacionados ao *documento alvo*, mas que por meio dos parâmetros definidos não foram recomendados.

Definição 14: o *conjunto relacionado* é uma lista com todos os documentos relacionados ao *documento alvo*. O *conjunto relacionado* é obtido por meio da expansão do resultado da consulta de documentos recomendados.

Para se obter os documentos do *conjunto relacionado* é necessário verificar a partir de cada documento do *conjunto de items* quais são os seus documentos recomendados. Sendo assim, o *conjunto relacionado* é composto da união do *conjunto de itens* de cada documento do *conjunto de itens*. Este novo conjunto de documentos é gerado a partir do uso do mesmo algoritmo de recomendação e considera os mesmos parâmetros definidos pelo usuário, desta forma o resultado se manterá íntegro em relação ao *documento alvo*.

Definição 15: O *nível de recomendação* é a classificação da relevância dos resultados obtidos na expansão da recomendação. A definição dos níveis permite facilitar a apresentação dos documentos recuperados nas recomendações. A quantidade de *níveis* do *conjunto de itens* que se deseja explorar pode ser definida pelo usuário. O primeiro nível é o caracterizado pelos documentos recomendados a partir de cada elemento do *conjunto de itens*.

Os documentos que são diretamente recomendados ao *documento alvo*, o *conjunto de itens*, farão parte do nível zero do agrupamento de documentos e na sequência serão verificados documentos que são semelhantes aos documentos que compõem este primeiro conjunto de documentos, o *conjunto relacionado*.

Os documentos semelhantes serão verificados a cada nível até que entre todos os

documentos da base de dados não se encontre mais nenhum *documento relacionado*, ou seja, quando nenhum outro documento possua algum grau de semelhança com o *documento alvo*.

É importante lembrar que os documentos já relacionados em um nível superior não são replicados, caso sejam também recomendados no próximo nível de relacionamento. Quando não se puder encontrar novos documentos relacionados ao *conjunto relacionado*, isto é, quando os documentos recuperados já fazem parte do *conjunto relacionado* é finalizado o processo da expansão da recomendação. Desta forma, a partir da união dos documentos do *conjunto de itens* e do *conjunto relacionado* a base de dados de documento é dividida em dois grupos distintos por meio da agregação, de um lado estão agrupados os documentos que possuem semelhança com o *documento alvo* e do outro um grupo de documentos que não possui relação com o *documento alvo*.

O Algoritmo 5 é responsável por encontrar todos os possíveis documentos recomendados. Trata-se de um tipo de operação que é denominado como fechamento recursivo, também chamado de clausula recursiva. Esta operação é utilizada em relacionamentos recursivos. No caso, recursivamente busca-se por meio dos documentos recomendados (F), os quais são os seus recomendados (Q), o que por meio da união de novos documentos ao *conjunto relacionado*, produz recursivamente o conjunto final de documentos relacionados ($F+$).

Algoritmo 5: ***Clausura dos Documentos Recomendados***

Entrada: F : conjunto de documentos recomendados

Saída: F+ : conjunto relacionado (de todos os documentos obtidos a partir do conjunto F (ou que satisfazem F)

```
F+ = F
REPEAT
    OldF+ = F+
    FOR EACH d in F+
        // documentos na lista de relacionados, inicialmente os recomendados
        Q = conjunto de documentos recomendados a partir de d
        F+ = F+ ∪ Q
UNTIL (OldF+ == F+)
```

Para exemplificar é descrito a seguir o caso mostrado na Figura 8. Se o Artigo `A` possui como artigos recomendados os artigos `B`, `D` e `E`, por intermédio dos critérios de *similaridade mínima* e D_{MAX}, pode-se dizer que o *conjunto de itens* de `A` é formado por `A={B,D,E}`. O próximo passo será então verificar os artigos que tem semelhança a estes três artigos do grupo de recomendações do Artigo `A`.

Dado que os recomendados de `B` são `{A,C,F,E}`, de `D` são `{A,F}` e de `E` são `{A,B}`, pode-se verificar que entre os recomendados de cada artigo, que naturalmente o Artigo `A` faz parte do agrupamento de todos os três artigos. Sendo assim é verificado em cada agrupamento novos documentos que possam fazer parte do agrupamento de `A`, desta forma,

o grupo do Artigo `A` é atualizado para o nível 1 do conjunto de documentos relacionados: $R1_A$= `{B,D,E,C,F}`. Foram adicionados os artigos `C` e `F` ao agrupamento de `A`, pois o Artigo `F` estava agrupado tanto com o artigo `B` quanto com o Artigo `D`, ou seja, se o artigo foi recomendado por um ou mais artigos do *conjunto recomendado* este fará parte do agrupamento do nível 1 do Artigo `A`. O Artigo `C` foi recomendado apenas pelo artigo `B`.

Observando-se o nível 1 do grupo de `A`, passa-se a investigar se existem mais documentos na base que possam fazer parte deste grupo, ou seja, que tenha alguma semelhança e que possam ser de interesse ao tomador de decisão. Desta forma, são analisados os artigos recomendados dos dois novos artigos inseridos no grupo de `A`, neste caso obtêm-se, por exemplo, que `C` recomenda os artigo `{X,B}` e `F` os artigos `{B,C,D}`. Dos novos grupos, temos que `B`, `C` e `D` já fazem parte do agrupamento de `A`, logo apenas o artigo `X` passa a fazer parte do agrupamento de `A`, ou seja, $R2_A$= `{B,D,E,C,F,X}`.

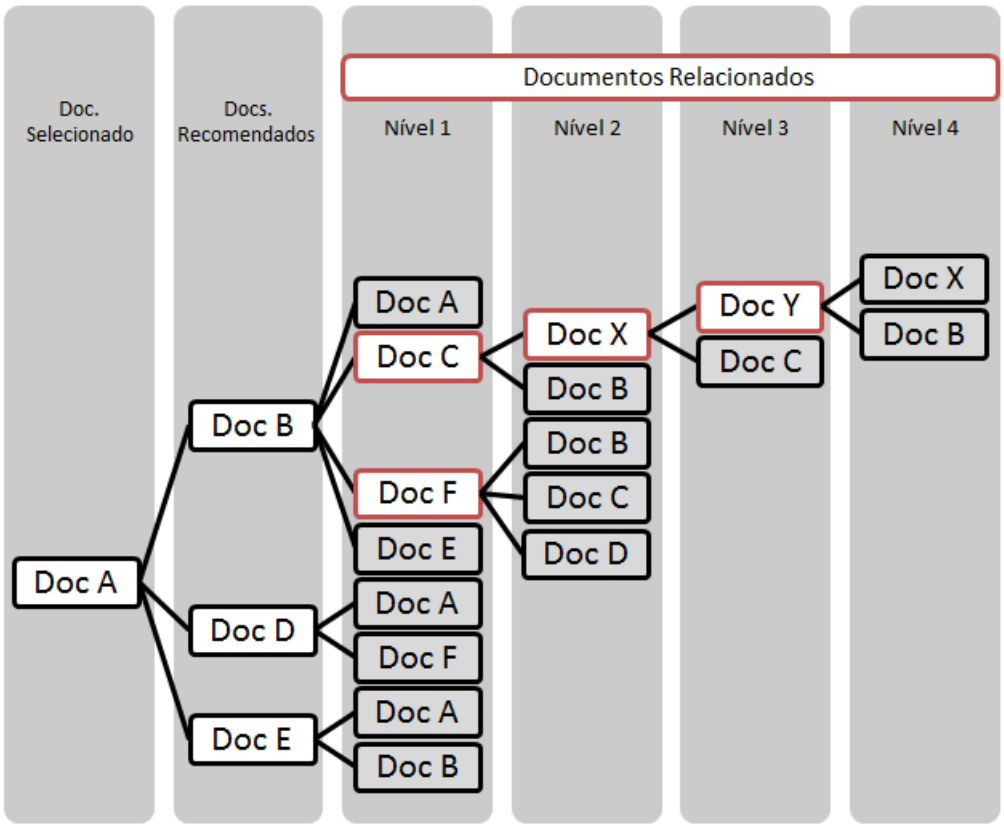


Figura 8 – Caso ilustrativo dos documentos relacionados em níveis

O próximo passo é verificar os documentos recomendados do Artigo `X` e neste caso o conjunto é formado apenas pelos artigos `{Y,C}`. O agrupamento de `A` recebe assim o novo documento Y, $R3_A$= `{B,D,E,C,F,X,Y}`. A próxima iteração é justamente verificar os recomendados de `Y`,`{X,B}`, e neste caso as iterações em busca do grupo `A` terminam já que `X` e `B` já fazem parte do grupo de `A`, finalizando o processamento de artigos relacionados ao Artigo `A`. O agrupamento resultante de documentos relacionados ao Artigo `A` é `{B,D,E,C,F,X,Y}`.

4.5 Sumário

Neste capítulo foram apresentadas as definições do modelo conceitual, cujo objetivo é obter os melhores resultados às consultas de agregação OLAP personalizadas para textos. Essas definições são importantes para estabelecer conceitos que foram abordados no decorrer da arquitetura e dos algoritmos utilizados no processo de agregação e recomendação.

A arquitetura do modelo POQT mostra o funcionamento da recomendação de documentos de texto e os usuários envolvidos em cada etapa do processo, assim é possível estabelecer quais são os objetivos do protótipo PAMDES que será abordado no capítulo a seguir.

Os conceitos utilizados pelos autores [RAVAT; TESTE; TOURNIER, 2007] foram estendidos nos algoritmos de recomendação da proposta, ou seja, o conceito da agregação foi utilizado, mas a maneira de realizar a agregação, sua personalização e como tornar a agregação um método para a recomendação dos documentos, foram desenvolvidos nesta proposta. O conjunto de algoritmos permitiu que o processo de recomendação e agregação fosse estruturado de forma que o procedimento pudesse ser também reproduzido independente da tecnologia a ser utilizada.

A primeira mudança em relação ao que foi definido em [RAVAT; TESTE; TOURNIER, 2007] teve como objetivo permitir a personalização do processo de agregação OLAP. No trabalho dos autores, o D_{MAX} era restrito e nesta proposta este valor é definido de acordo com o critério do usuário gestor. Além disso, foram utilizadas medidas de similaridade no processo de agregação e recomendação de documentos, permitindo a personalização da consulta.

A ontologia de domínio é utilizada com o objetivo de verificar o quanto os termos que definem o assunto de cada documento são similares, ou seja, é o D_{MAX} que define a distância máxima estabelecida entre os termos, para que esses possam ter um medida de semelhança aceitável. A outra medida de similaridade utilizada, a *similaridade mínima*, permite que usuário analista já defina o valor de corte na parametrização, ou seja, estabelece um limite para considerar quais são os documentos semelhantes e quais não são semelhantes. Após a verificação de todos os possíveis documentos semelhantes ao *documento alvo*, isto é, após o processo de expansão da agregação, é possível verificar que a agregação cria dois grupos distintos de documentos, aqueles que são semelhantes e aqueles que não são semelhantes ao *documento alvo*.

A recomendação com base no modelo POQT utiliza um algoritmo de LCA desenvolvido para uma árvore, onde os termos (nós) podem se repetir em subárvores diferentes. O identificador do termo MeSH na ontologia é composto por elementos que indicam quais são todos os nós antecessores em relação ao nó raiz da árvore ontológica, como mostrado na Figura 6. Para aprimorar o pré-processamento da recomendação passou-se a utilizar os

rótulos de caminhos (posições dos nós) na ontologia para se encontrar a distância entre os termos na ontologia, como citado na Seção 4.4. Sendo assim, não se verifica o último ancestral comum pela estrutura inicial de árvore, mas sim pela estrutura do caminho do termo MeSH, onde ambas as posições são comparadas a fim de se encontrar o antecessor de forma mais eficiente. O cálculo da distância máxima também foi modificado, verificando entre o antecessor e o termo MeSH analisado aquele que apresenta a maior distância por meio da posição na ontologia.

O protótipo do sistema de recomendação, a princípio, não considerava que os mesmos termos MeSH podem ser identificados em mais de uma subárvore, porém seus nós são sempre distintos (caminhos diferentes). Logo, todas as comparações entre termos MeSH estavam restritos às comparações realizadas pelo caminho do nó. Para retornar respostas mais precisas, a comparação de cada par de termo MeSH passou a ser a partir do termo MeSH e suas posições, ou seja, não se analisa apenas uma posição do nó do termo, mas sim todas as suas posições nas subárvores.

Essa alteração eliminou também um outro revés encontrado no algoritmo, onde não era permitido que documentos com termos MeSH iguais fossem agrupados, se estes estivessem em posições muito distantes em relação ao caminho na ontologia. Se dois artigos, por exemplo, possuem dois termos MeSH então o número total de eventos (N) é igual a 4. A dupla de termos MeSH de cada artigo é igual, ou seja, em dois eventos a distância dos termos MeSH será nula, comparando os termos iguais, mas quando se comparam os dois termos diferentes estes podem ser muito distantes na ontologia, ou seja, não atingindo D_{MAX} mínimo. Neste caso, o valor da frequência relativa f_i será de 50% e se o valor de similaridade mínima for 75%, o artigo com os mesmos termos MeSH não será recomendado para o artigo dado. Ao se considerar também as posições dos termos nas subárvores, existe então uma probabilidade maior de encontrar os termos distintos do mesmo artigo na mesma subárvore ou em nós mais próximos na ontologia.

A implantação da expansão da recomendação, onde é possível recuperar artigos em outros níveis de recomendação, também é válido para recuperar artigos que são similares a um ou mais artigos recomendados diretamente com o documento selecionado, ou seja, sem esta funcionalidade não era possível realizar a distinção de dois grupos de artigos existentes na base de dados: agrupamento dos artigos que possuem uma *similaridade mínima* com o documento selecionado e aqueles que não possuem nenhuma semelhança e portanto não são de interesse do usuário gestor. Essa funcionalidade permite recuperar artigos com termos MeSH iguais, mas muito distantes, uma vez, que são esgotadas todas as possibilidades de recomendação com os parâmetros definidos pelo usuário.

5 Protótipo: PAMDES

Este capítulo visa descrever o objetivo, a modelagem dos dados, as funcionalidades, o desempenho da recomendação e o desenvolvimento do protótipo PAMDES. Os cenários ilustrados por meio das interfaces de parametrização e de recomendação permitem entender o funcionamento do protótipo.

5.1 Objetivo do Protótipo

O cenário utilizado para mostrar o funcionamento do protótipo PAMDES (*PubMed Article Multidimensional Recommendation System*) é baseado no estudo de caso descrito no Apêndice A. O objetivo do desenvolvimento do protótipo é avaliar a seguinte pergunta realizada por um usuário gestor: *é válido continuar avançando na linha de pesquisa do seu trabalho atual?*

Para tentar responder a esta pergunta, o usuário gestor pode melhorar a sua análise encontrando estudos com objetivos similares, assim poderá observar os trabalhos que estão sendo realizados nesta área, avaliando se é uma área promissora ou não.

Para esse propósito, o gestor pode considerar como uma boa ideia agregar artigos que possuem assuntos relacionados. Logo, o algoritmo de recomendação proposto torna-se útil para criar um grupo de artigos semanticamente semelhantes dado um determinado documento.

5.2 Modelagem de Dados

O modelo estrela utilizado na modelagem do *Data Warehouse* é apresentado na Figura 9. A tabela fato, *Articles*, armazena os artigos e número de citações. As dimensões para análise são: D_Author, D_Date, $D\_Journal$ e as palavras-chave são armazenadas na $D\_MeshList$.

Como o relacionamento cardinal entre os termos MeSH e os artigos é muitos para muitos, foi introduzida uma tabela chamada de $DBridge\_MeshArticle$. Este recurso permite resolver a representação de múltiplas palavras-chave por documento (como em [KIMBALL; ROSS, 2002]) e para analisar os resultados usando SQL padrão. Uma chave primária foi criado para cada termo MeSH do artigo e foi utilizada como chave estrangeira na tabela fato. Se um artigo tiver quatro descritores MeSH, logo o artigo tem uma referência para cada dimensão $DBridge\_MeshArticle$ contendo cada uma das quatro palavras-chave, que fazem referência a tabela $D\_MeshList$.

A árvore de termos MeSH está disponível em arquivo de texto no site da PubMed e seus dados foram inseridos na tabela $D\_MeshList$, que possui 55612 registros de termos da ontologia. Foram armazenados 132 artigos na tabela fato, dos quais contemplaram 238 termos MeSH distintos.

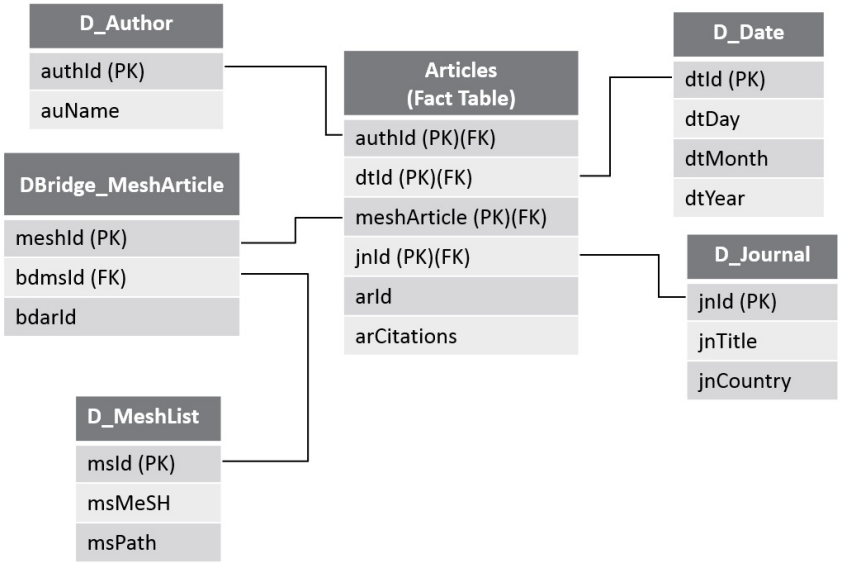


Figura 9 – Modelo multidimensional do estudo de caso

Para sistemas que realizam a gestão de um grande volume de informações, existe a possibilidade de carregar diretamente os dados para o *Data Warehouse* no formato XML, como é o caso da PubMed que permite o *download* dos artigos XML que descrevem cada publicação contida na base. As informações dos artigos utilizadas no experimento foram inseridas de forma manual, já que o processo de ETL não é o foco do trabalho.

5.3 Funcionalidades do Sistema

A construção do protótipo permite que sejam analisados e testados os algoritmos que compõem a recomendação personalizada de documentos. O protótipo permite: (1) que sejam definidos os parâmetros da personalização; (2) o processamento das recomendações ocorra de acordo com os parâmetros; (3) a visualização dos documentos armazenados na base de dados; (4) a seleção de um documento para a verificação de documentos semelhantes; (5) a seleção das dimensões e refinamento da recomendação; (6) a busca de artigos relacionados em toda a base de documentos, a partir da expansão da recomendação e (7) a visualização do resultado final da recomendação, bem como a visualização dos termos MeSH de cada artigo selecionado na interface.

O protótipo PAMDES foi implementado utilizando o Microsoft Visual Studio 2010 e os dados do estudo de caso foram armazenados no SGBD MS SQL Server 2008. O

objetivo do *Data Warehouse* é a análise de artigos científicos e as decisões mais relevantes são temporais e relacionadas aos periódicos (*journals*), autores e aos temos MeSH.

5.4 Implementação da Expansão da Recomendação

O procedimento para se encontrar o conjunto de documentos relacionados ao *documento alvo*, como descrito no Capítulo 4, utiliza um algoritmo de clausura recursiva. Na sequência, é mostrado como foi realizada a implementação desta funcionalidade.

O Algoritmo 5 é uma consulta que fornece uma forma incremental de selecionar os documentos relacionados. A ideia de ser incremental vem do fato de que a cada nova consulta, a partir de um determinado conjunto de documentos, existe a união de elementos em um só conjunto. Desta maneira o conjunto de documentos relacionados aumenta gradualmente e por etapas regulares (níveis).

Na tabela fato *Articles* está relacionado a identificação do artigo ($arId$), seus autores, sua data de publicação, o *journal* e as suas palavras-chave. Já na tabela *tbrcRecommendation* além da identificação da recomendação ($rcId$), possui a identificação do documento alvo ($rcId$), a identificação do documento candidato ($rcarIdRecomendado$) e o valor de frequência de similaridade entre eles.

Algoritmo 5: ***Implementação da Clausura dos Documentos Recomendados***

Entrada: *Output_Recom*

Saída:*Output_Rel*

```
// Buscar pelos possíveis documentos recomendados do Output_List
// Não são inclusos documentos repetidos
SELECT DISTINCT * FROM Articles
WHERE arId IN
(
    //Tabela de pré-processamento
    SELECT rcarIdRecomendado FROM tbrcRecommendation
    //Identificação de documentos relacionados
    WHERE rcarId IN (SELECT rcarId FROM @id)
)
```

A instrução da consulta busca os documentos pelo parâmetro *@id* que pode ser: (1) um documento apenas, quando o *documento alvo* possui apenas um documento recomendado, ou seja, o *conjunto de itens* é composto por apenas um documento, ou (2) pode ser um *conjunto de itens* com mais de um documento. Quando se obtém o primeiro *conjunto relacionado*, pode-se finalizar a consulta ou ainda estender seus resultados aos níveis de agregação (documentos similares aos documentos já relacionados), ou seja, dependendo do nível de interesse do usuário realiza-se uma nova consulta utilizando como parâmetro os documentos do *conjunto relacionado*, expandindo a consulta em mais um nível de documentos relacionados.

No exemplo que será demonstrado na seção a seguir (Seção 5.5), é possível observar a recomendação mostrando dois níveis de documentos relacionados. A recomendação é obtida pelo uso do Algoritmo 5 e revela os artigos que são relacionados a partir do conjunto de artigos que são recomendados.

5.5 Funcionamento do Protótipo

Nesta seção será descrito o funcionamento do protótipo PAMDES por meio de exemplos de configuração, recomendação e expansão da recomendação. O objetivo do pré-processamento é fundamentado pelo tempo de resposta da recomendação, ou seja, ele é essencial para um bom desempenho das agregações OLAP. Para realizar o pré-processamento da recomendação dos artigos armazenados no *Data Warehouse*, é necessário o processo de parametrização que é realizado pelo usuário analista. Após essa fase, o usuário gestor é responsável pelas interações com o sistema de recomendação, definindo consultas OLAP de acordo com o seu interesse.

5.5.1 Desempenho da Recomendação

Inicialmente, o sistema de recomendação tinha um tempo de resposta alto, pois a recomendação era realizada no ato da solicitação do usuário gestor, ou seja, o algoritmo de recomendação computava os valores das distâncias termo a termo de cada artigo da base, buscando seus nós correspondentes na árvore ontológica e calculando a frequência. No decorrer destes cálculos de distância máxima e similaridade, o usuário gestor deveria aguardar o processamento do resultado.

O tempo de resposta da recomendação era cerca de 9,79 segundos por artigo para a base do estudo de caso que era composta de 132 artigos, sendo que cada artigo tinha em média três termos MeSH. Foi verificado que quanto maior o número de termos MeSH maior a quantidade de comparações entre as palavras-chave, aumentando a quantidade de comparações e, por consequência, sendo maior o tempo de resposta. Para o usuário gestor, tomador de decisão, esperar por uma recomendação um tempo igual a este ou ainda superior não é viável, até porque quanto maior a base de dados maior será o tempo de resposta também. A base de dados da PubMed atualmente possui 24 milhões de artigos e para poder realizar a recomendação citada acima seriam necessários aproximadamente 20 dias de processamento, considerando que o algoritmo não está utilizando nenhum tipo de paralelismo para processar a recomendação.

Para o sistema de recomendação retornar os artigos recomendados em tempo real, foi desenvolvida uma rotina de pré-processamento que utiliza os parâmetros de *similaridade mínima* e D_{MAX} determinados pelo usuário analista. Como parâmetro de partida para a realização dos testes foi considerado o D_{MAX} igual a 3 e 4, que foi a sugestão heurística de

[RAVAT; TESTE; TOURNIER, 2007]. Ao acionar a rotina de pré-processamento para cada artigo no *Data Warehouse* são verificadas as relações entre cada documento na base com o primeiro documento. Se o valor mínimo de similaridade for maior que o valor definido de frequência, a recomendação é armazenada em uma tabela específica. A solução é eficaz já que o resultado da recomendação é mostrado assim que a recomendação é acionada pelo o usuário gestor.

A seguir, é demostrado como ocorre o processo de parametrização do pré-processamento da recomendação. O usuário analista é responsável pela configuração da recomendação para a base de documentos e para a ontologia específica. Após a configuração e o pré-processamento da recomendação, o usuário gestor pode acessar a *interface de recomendação* e verificar possíveis recomendações para um documento selecionado.

5.5.2 Processo de Parametrização

O processo de parametrização é a base da realização do pré-processamento da *similaridade mínima* entre os documentos armazenados no *Data Warehouse*. Antes de iniciar a parametrização, é importante carregar a árvore ontológica na memória e na sequência o usuário analista insere os parâmetros na tela de configuração, conforme ilustrado na Figura 10. No exemplo mostrado, o analista atribui uma *similaridade mínima* de 0,50 e valor de *Fator de ontologia* (D_{MAX}), igual à 4.

Após a configuração adequada, o analista aciona o botão *"Configure"*. Deste forma, o pré-processamento será realizado como mostrado na Figura 11. Apesar das telas do sistema serem de um protótipo, sempre são mostrados textos explicativos para ajudar o usuário a entender o significado dos parâmetros. Para visualizar a ajuda é necessário apenas acionar o botão *"?"*.

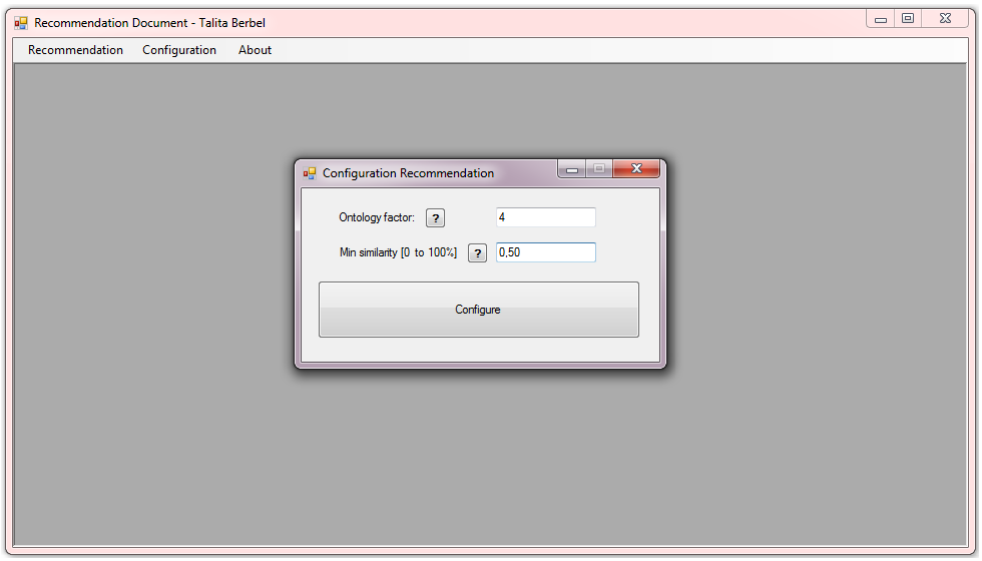


Figura 10 – Interface de configuração dos parâmetros da recomendação

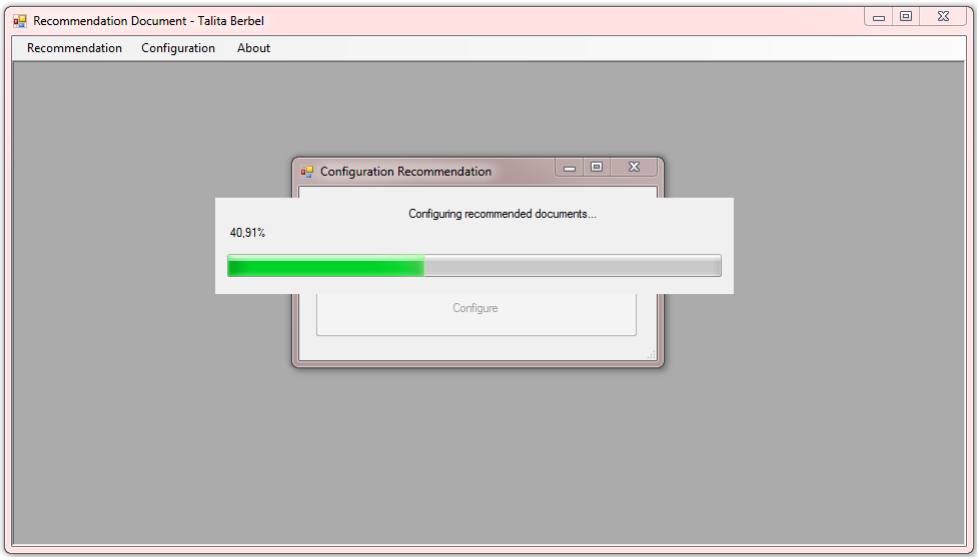


Figura 11 – Pré-processamento com os parâmetros da recomendação

Para fins de análise, no estudo de caso todos os cálculos de similaridade entre os documentos foram calculados e armazenados, porém se o volume de dados na base for maior, recomenda-se apenas armazenar na tabela de recomendações aquelas superiores ou iguais ao valor de similaridade mínimo definido na parametrização. Desta forma, na próxima etapa da utilização do sistema, que consiste na consulta a recomendação por meio da *interface de recomendação* do PAMDES, o usuário gestor terá a possibilidade de aumentar a similaridade para refinar sua consulta, ou seja, mostrar apenas os documentos com a similaridade determinada na interface. O usuário também poderá diminuir o valor de similaridade até o limite da *similaridade mínima* na configuração da parametrização.

5.5.3 Cenário 1: Recomendação de Documentos

Após o processo de parametrização, inicia-se o processo de recomendação e da expansão da recomendação. A Figura 12 mostra a *interface de recomendação* do sistema desenvolvido para a recomendação de documentos por meio da ontologia. No lado esquerdo são mostrados os artigos armazenados no *Data Warehouse*, seus respectivos autores, ano da publicação e *journal*.

A recomendação de documentos começa a partir da seleção de um artigo, o qual se deseja saber as suas recomendações. O primeiro artigo que possui como título *"Acute myocarditis mimicking acute myocardial infarction associated with pandemic 2009 (H1N1) influenza A virus"* foi selecionado para análise.

O cubo de visualização da consulta envolve, a princípio, apenas a dimensão de palavras-chave. Essa dimensão não é apresentada na interface, porque o objetivo do protótipo é a agregação implícita por palavras-chave, ou seja, pode-se agregar documentos com base no tempo, *journals*, autores, porém estas agregações ocorrem baseadas nas

palavras-chave.

Na parte central da interface são mostrados os atributos de dimensões disponíveis (autor, ano, *journal*), além da similaridade mínima entre os documentos que poderá ser ajustada. Para o exemplo, a *similaridade mínima* foi ajustada para 0,75 para refinar os resultados, ou seja, é possível fazer esta alteração já que o valor é superior a *similaridade mínima* da parametrização (0,50).

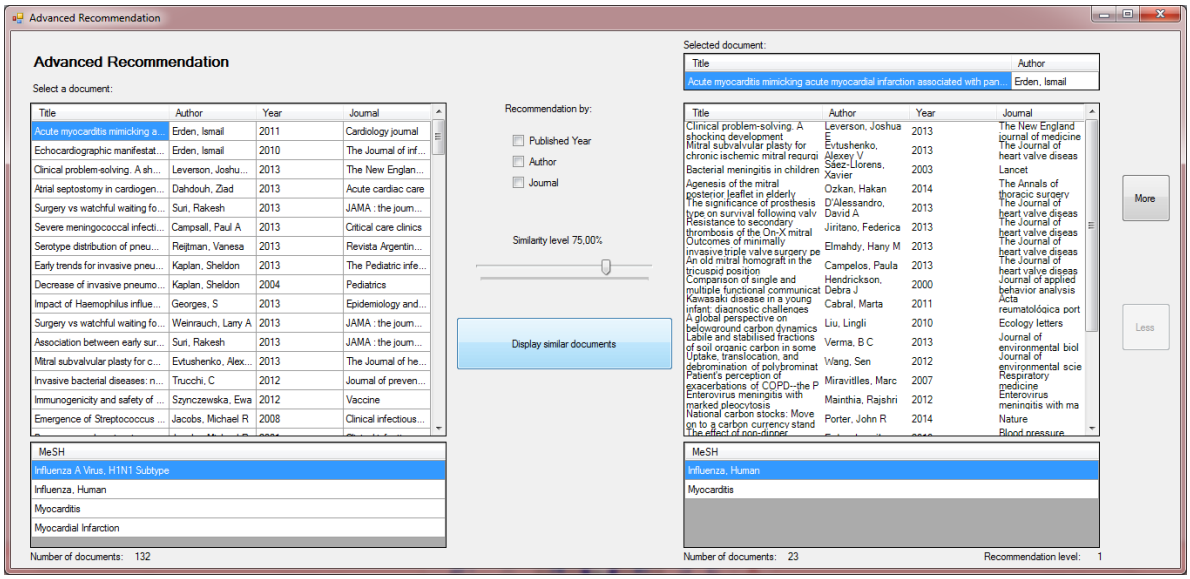


Figura 12 – Interface PAMDES: Recomendações a partir de um dado artigo - Cenário 1

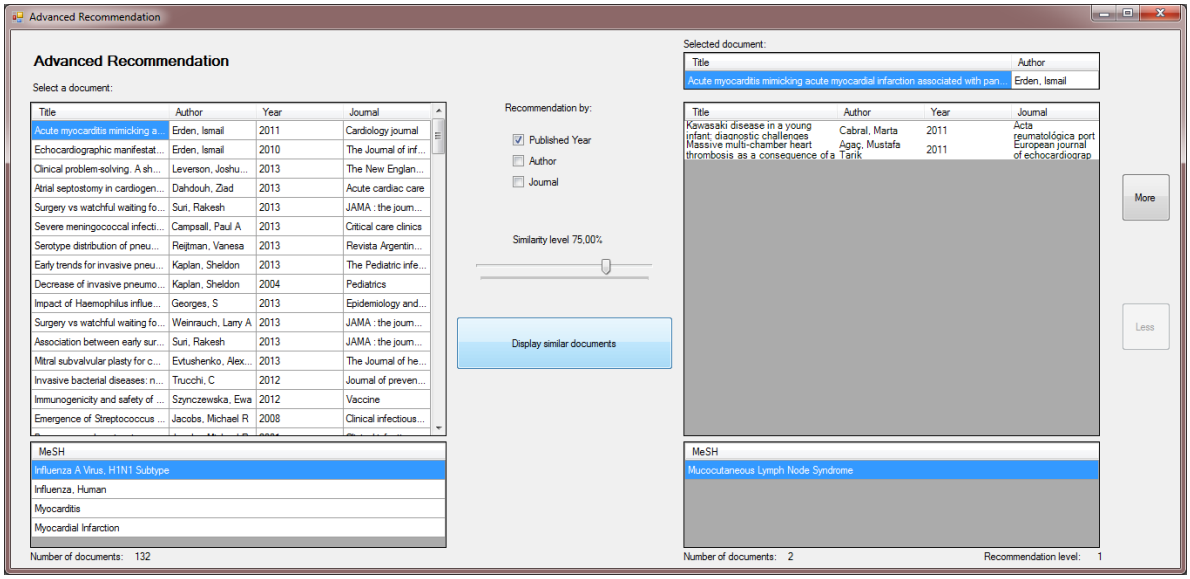


Figura 13 – Interface PAMDES: Artigos recomendados após seleção de uma dimensão - Cenário 1

Após clicar no botão central, *"Display similar documents"*, são mostrados no lado direito da tela, os artigos que são recomendados e cada uma de suas palavras-chave para fins de verificação, como mostrado na Figura 12. No exemplo mostrado no

Cenário 1, o artigo selecionado possui os seguintes termos MeSH: *'Influenza A Virus, H1N1 Subtype'*,*'Influenza, Human'*, *'Myocarditis'* e *'Myocardial Infarction'*. Para efeitos de verificação, na recomendação, o primeiro artigo apontado tem os seguintes descritores: *'Influenza, Human'* and *'Myocarditis'*. Desta forma, é possível ver facilmente que os descritores são semanticamente relacionados. O lado direito da tela mostra os 23 artigos recuperados como artigos recomendados e clicando sobre cada um dos artigos recomendados pode-se observar os termos MeSH associados a eles que são mostrados na parte inferior da tela.

Caso o usuário gestor tenha a necessidade de limitar o seu cubo de visualização, este pode selecionar uma das dimensões disponíveis. Ao selecionar a dimensão ano de publicação (*Published year*), por exemplo, e pressionar o botão central, um novo conjunto de artigos recomendados é processado, ou seja, o resultado foi filtrado (Figura 13), trazendo dois artigos que são do mesmo ano do artigo selecionado e que podem ser de interesse do usuário.

Se os artigos mostrados ao final do processo de recomendação são artigos muitos restritivos, pode-se optar por diminuir o valor de similaridade mínima para 0,50, como mostrado no exemplo do caso 2. A Figura 14 mostra os sete artigos que satisfazem a este novo valor de similaridade e a partir da análise destes artigos pode-se filtrar os resultados ou expandir a recomendação como na Figura 12.

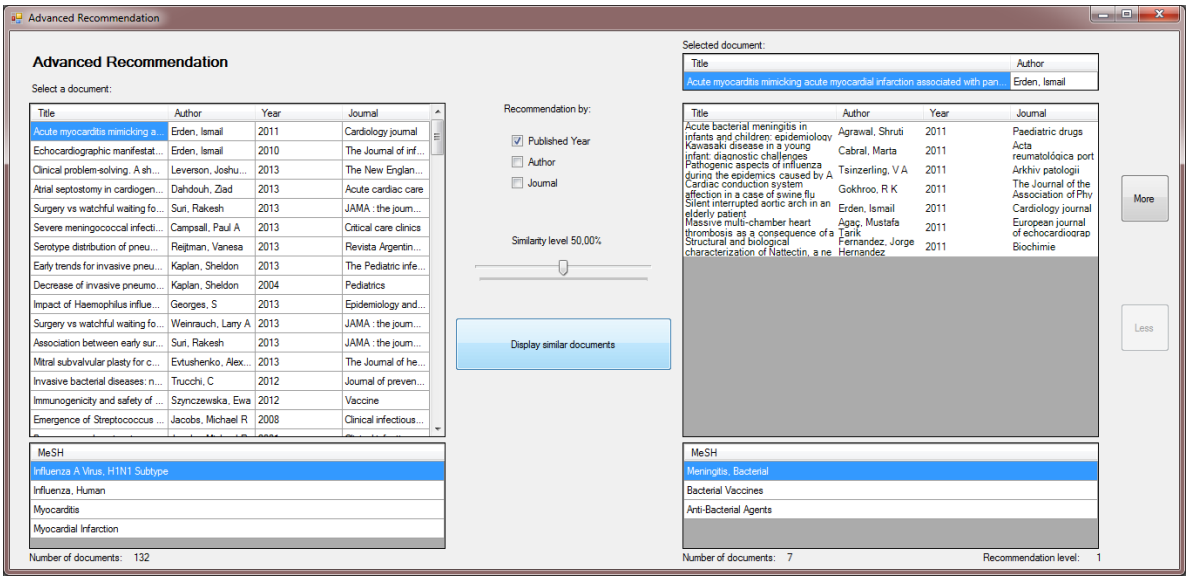


Figura 14 – Interface PAMDES: Artigos recomendados a partir do ajuste da similaridade - Cenário 1

5.5.4 Cenário 2: Expansão da Recomendação de Documentos

No Cenário 1, após verificar o conteúdo dos dois artigos mostrado na Figura 13, pode ser interessante analisar os artigos relacionados a eles (extensão da recomendação). Neste

caso o usuário gestor pode acionar o botão lateral *"More"* e uma nova seleção de artigos é mostrada (Figura 15), agrupando os artigos relacionados aos artigos recomendados.

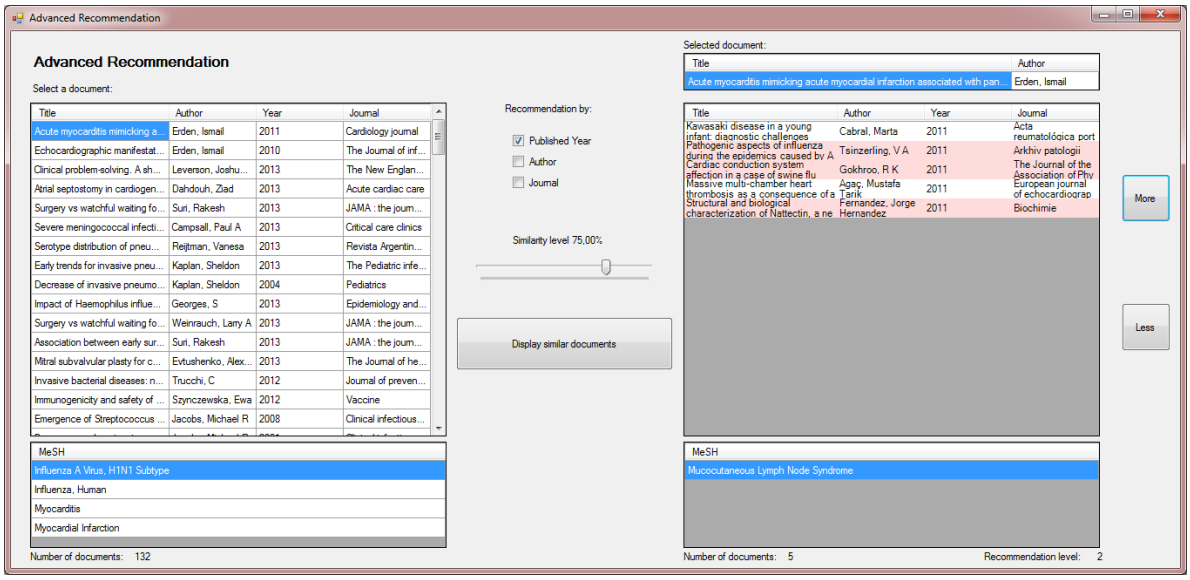


Figura 15 – Interface PAMDES: Artigos relacionados - Cenário 2, Ex.1

Se o usuário acionar a extensão da recomendação é mostrado no canto inferior direito qual o nível de recomendação corrente (*"Recommendation level"*). Para facilitar a distinção dos artigos, a listagem de artigos relacionados passa a ter uma gradação de cor diferenciada, ou seja, cada vez que o botão é pressionado a função pode encontrar ou não mais artigos relacionados.

As gradações de cores são usadas para que o usuário possa identificar na recomendação, quais são os artigos recomendados e quais são os relacionados. No conjunto de documentos recomendados (lado direito da tela) os artigos recomendados possuem um fundo branco, o primeiro grupo de artigos relacionados, primeiro nível da extensão, é mostrado em tom claro de vermelho e para os demais níveis a cor vai aumentando a sua intensidade. Quando o gestor preferir voltar a seleção inicial de artigos basta acionar o botão lateral *"Less"*.

No Exemplo 1 do Cenário 2 (Figura 15) apresentado, ao se buscar novos artigos relacionados a partir do grupo existente não existem novas recomendações, finalizando assim o processo. Desta forma, pode-se dizer que, no final do processo de recomendação os documentos armazenados no *Data Warehouse* são divididos em dois grupos: os associados e aqueles sem nenhuma relação com o dado documento.

Para ilustrar melhor a função que estende os artigos recomendados para um grupo maior, os artigos relacionados, é mostrado o Exemplo 2 do Cenário 2 nas telas a seguir. Ao selecionar o artigo *"Silent interrupted aortic arch in an elderly patient"* e acionar a recomendação são recuperados seis artigos recomendados, como mostrado na Figura 16.

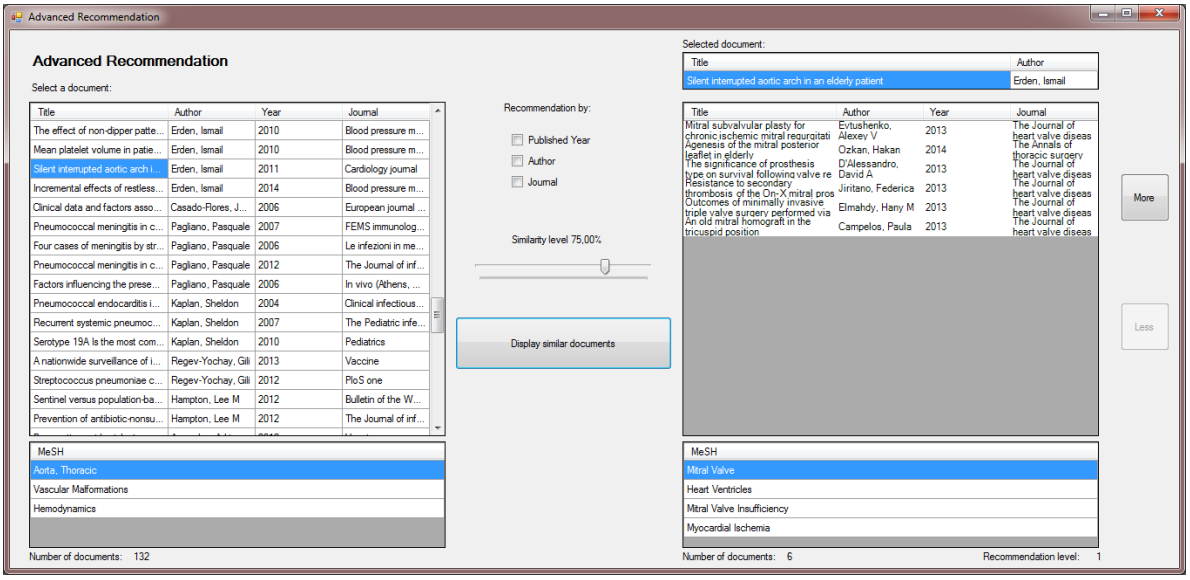


Figura 16 – Interface PAMDES: Artigos recomendados - Cenário 2, Ex.2

Na sequência ao solicitar mais artigos relacionados no botão *"More"*, são agrupados mais 35, pois foram encontrados esses artigos no primeiro nível de artigos relacionados, ou seja, totalizando 42 artigos (Figura 17). Se o grupo de artigos recomendados ainda não for suficiente para análise do usuário, este pode acionar novamente o botão *"More"* e verificar o segundo nível de artigos relacionados e diferenciados por uma tonalidade mais forte da cor na lista de artigos. A Figura 18 mostra os 53 artigos relacionados ao artigo selecionados e seus respectivos termos MeSH.

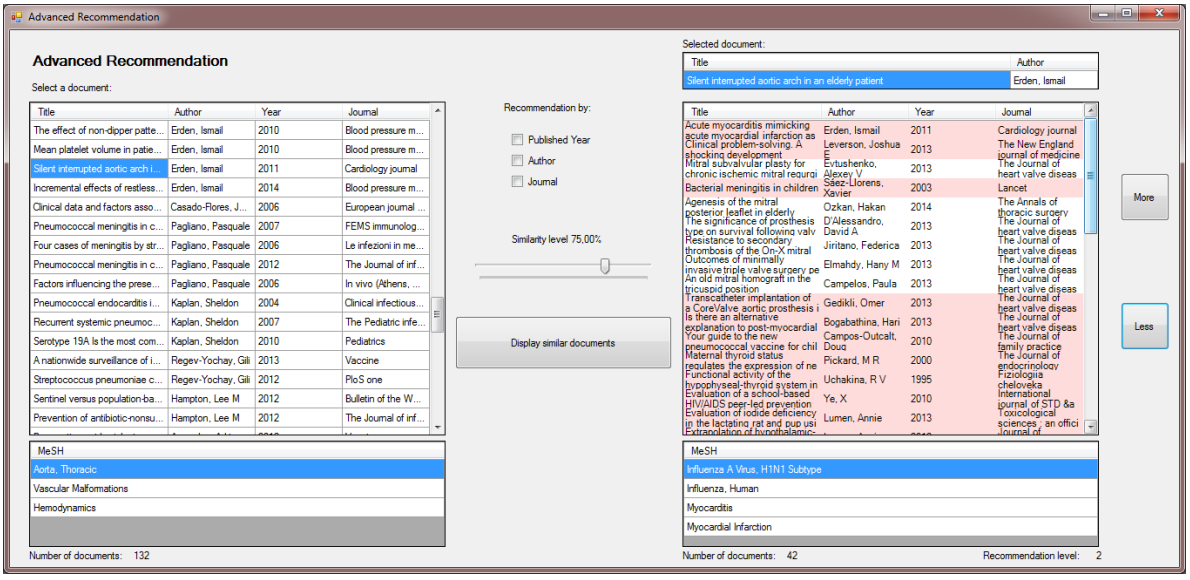


Figura 17 – Interface PAMDES: Artigos relacionados no 1º nível - Cenário 2, Ex.2

Após a apresentação do funcionamento geral do protótipo, ilustrando os algoritmos propostos no Capítulo 3, seguem os experimentos realizados a fim de validar a proposta da recomendação.

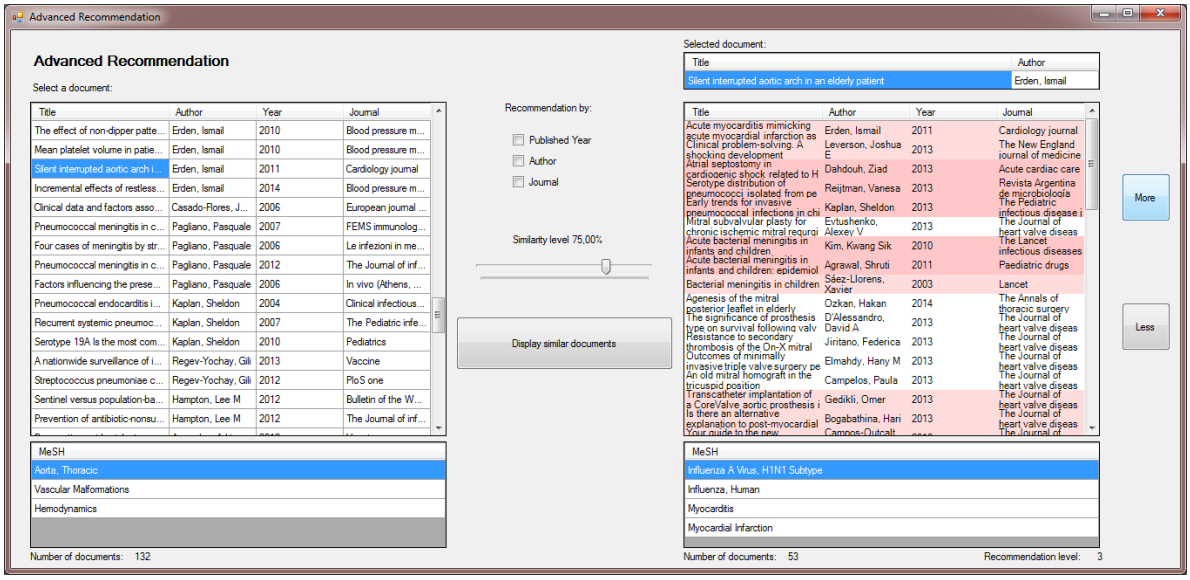


Figura 18 – Interface PAMDES: Artigos relacionados no 2º nível - Cenário 2, Ex.2

5.6 Sumário

O protótipo PAMDES foi desenvolvido utilizando os conceitos estabelecidos no modelo conceitual, com a finalidade de obter informações de dados não estruturados, documentos de textos, de modo que mediante a personalização de agregações OLAP fosse possível a recomendação de documentos.

O PAMDES também se tornou um sistema funcional composto por dados e documentos reais e faz o uso a mesma ontologia de domínio utilizada pela PubMed. A descrição do funcionamento do protótipo permite verificar quais são as possibilidades de personalização da consulta, bem como a atuação de cada tipo de usuário no processo da recomendação e da expansão da recomendação de documentos.

Como melhoria no protótipo, a sua interface poderia incluir a possibilidade do usuário gestor poder expressar a sua opinião sobre o resultado da recomendação, ou seja, apontar se os artigos recomendados satisfazem as suas expectativas. Sendo assim, seria possível estudar maneiras de se usar este *feedback* como um novo critério para realizar a recomendação.

6 PAMDES: Experimentos

Este capítulo descreve as configurações relacionadas aos experimentos, detalhando os perfis dos especialistas envolvidos nas validações, as métricas de avaliação e ajuste da medida *F-measure*. Na sequência são apresentadas as validações sobre a recomendação de documentos, as quais foram realizadas com o auxílio de especialistas em diferentes áreas da biologia e genética.

6.1 Configuração dos Experimentos

A seguir será detalhado o processo de configuração dos experimentos realizados para parametrização empírica da recomendação e a validação da recomendação do protótipo PAMDES. As métricas de avaliação permitem que a parametrização da recomendação seja avaliada e ajustada, a fim de se verificar quais são os melhores valores dos parâmetros para a recomendação semântica de documentos do estudo de caso.

6.1.1 Perfil dos Especialistas

Após o armazenamento dos documentos e da ontologia no *Data Warehouse*, o próximo passo para executar o procedimento experimental foi solicitar o auxílio de quatro professores doutores da Universidade Federal de São Carlos para que estes pudessem ajudar a construir o conjunto de dados de avaliação. As áreas de pesquisa dos doutores são mostradas na Tabela 3.

Professor	**Área de Pesquisa**
1	Estrutura Genética de Populações Florestais
2	Genomas Mitocondriais de Artrópodes
3	Biologia Computacional
4	Ecologia

Tabela 3 – Área de pesquisa dos especialistas

Os especialistas participaram das três validações que serão exploradas na sequência: (1) Ajuste da medida *F-Measure*, (2) Validação do conjunto de artigos recomendados e (3) Validação da expansão da recomendação.

6.1.2 Métricas para a Avaliação

Para avaliar os resultados obtidos com o algoritmo de recomendação foram utilizadas as tradicionais métricas de RI: *recall*, *precision* e *F-measure*. Os experimentos foram conduzidos para avaliar a qualidade das recomendações geradas em termos de *precision* e *recall* [HAN; KAMBER, 2006].

Para realizar a análise dos resultados da recomendação foi utilizada a matriz de confusão, a qual permite analisar os acertos e os erros de um algoritmo. A Tabela 4 mostra a sua estrutura e a seguir são apresentados os seus termos:

	Pred=pos	Pred=neg
Classe=pos	TP	FN
Classe=neg	FP	TN

Tabela 4 – Definição da matriz de confusão

- **TP** (*true positive*): especialista humano e algoritmo estão de acordo.
- **FN** (*false negative*): especialista humano apontou uma categoria e o algoritmo não.
- **FP** (*false positive*): algoritmo apontou uma categoria, mas o especialista humano discordou.
- **TN** (*true negative*): especialista humano e algoritmo concordaram que não pertencia a uma categoria.

	Pred=pos	Pred=neg
Classe=pos	17	11
Classe=neg	06	98

Tabela 5 – Matriz de confusão do exemplo

A Tabela 5 apresenta a matriz de confusão do exemplo mostrado na Figura 12. O modelo apresentado fez 115 previsões corretas (TP+TN) e 17 previsões incorretas (FP+FN), desta maneira, o modelo do exemplo conta com 132 casos. A taxa de precisão ((TP+TN)/Total) correspondente é de 0,8712.

Precision e *recall* são duas métricas amplamente usadas em aplicações nos casos em que a detecção bem sucedida de uma das classes é mais significativa que outras [TAN; STEINBACH; KUMAS, 2009], ou seja, para casos em que os grupos são desbalanceados. Abaixo são mostradas respectivamente as fórmulas de *precision* (6.1) e *recall* (6.2):

$$Precision, P = \frac{TP}{TP + FP} \qquad (6.1) \qquad\qquad Recall, R = \frac{TP}{TP + FN} \qquad (6.2)$$

Nesta proposta o *recall* é definido como a razão entre todos os documentos relevantes recuperados pelo sistema e todos os documentos relevantes disponíveis no banco de dados, ou seja, é a probabilidade de um documento relevante ser efetivamente recuperado na recomendação. Em sistemas de recomendação, o *coverage* é a percentagem de itens que podem realmente ser recomendados, semelhante ao *recall* [THOLLOT; KUCHMANN-BEAUGER; AUFAURE, 2012].

A medida *precision* é a porcentagem de artigos no conjunto de artigos relevantes que são classificados como relacionadas pelo especialista, ou seja, número de artigos relevantes dentro do conjunto de artigos que podem ser ou não ser relevantes.

Ambas as medidas, *precision* e *recall*, podem assumir valores entre 0 e 1. Essas medidas são inversamente proporcionais, isto é, quando o valor de *precision* sobe, o *recall* normalmente cai e vice-versa. Um sistema com alto valor de *recall*, mas com baixo valor de *precision* retorna muitos resultados, mas a maioria de seus documentos previstos estão incorretos quando comparado com a opinião de um especialista. Um sistema com alto *precision*, mas com baixo valor de *recall* é exatamente o oposto, retornando poucos resultados, mas a maioria de seus documentos previstos estão corretos se comparados com a opinião de um especialista [HAUSSER, 2001]. Um sistema ideal, com altos valores de *precision* e *recall*, irá retornar muitos resultados e com todos os documentos recuperados corretamente.

A medida de *F-measure*, também chamada de *F-Score*, considera as medidas de *recall* e *precision* buscando a harmonia entre ambas. Essa medida verifica a eficiência do sistema considerando o erro nas duas medidas. Quando *F-measure* é balanceada, isto é, considerando o mesmo peso para ambas as medidas, ela é chamada de *F1-measure*. No exemplo apresentado na Tabela 5 as medidas de *recall*, *precision* e *F1-measure* são respectivamente; 0,7391; 0,6071 e 0,6667.

6.1.3 Ajuste da Medida *F-measure*

Em uma pesquisa realizada com os professores especialistas, foi investigada se a *F-measure* deveria ser balanceada ou desbalanceada. A *F-measure* combina as características das medidas de *recall* e *precision* em uma métrica conjunta, que permite a aplicação de pesos (ponderação) entre essas medidas.

Cada especialista respondeu a seguinte pergunta e justificou a sua escolha.

O que é mais importante no grupo de artigos recomendados a partir de um de seus artigos?

a) que o grupo de artigos recomendados traga poucos "falsos-positivos", onde falsos-positivos são artigos que vem recomendados mas não são relacionados.

b) que o grupo de artigos recomendados traga poucos "falsos-negativos", onde falsos negativos são artigos que não vem na recomendação e deveriam vir pois são recomendados.

c) as duas considerações anteriores (a e b) são importantes.

Em relação à justificativa a maioria (três especialistas) apontou que não era interessante ter um conjunto de artigos que não possui todos os artigos relacionados, no caso os falsos-negativos, pois trazem insegurança em relação a eficiência da recomendação. Por outro lado, os falsos-positivos também trazem desconforto e incomodam os especialistas, pois são artigos recomendados de forma incorreta.

Como resultado da pesquisa, foi identificado que a maioria dos professores consideram os dois fatores de medidas essenciais, alternativa **c**, por esta razão decidiu-se usar *F1-measure* para calcular a média final harmônica da precisão nos experimentos seguinte. A *F1-measure* é definida na fórmula (6.3) abaixo:

$$F_1 = \frac{2 * (Precision * Recall)}{(Precision + Recall)} \tag{6.3}$$

A seguir são detalhados os processos de validação do algoritmo de recomendação, seus procedimentos e resultados.

6.2 Validação Experimental

O processo de validação experimental tem o objetivo de determinar se a utilização do modelo conceitual da recomendação semântica de documentos mediante a personalização de agregações OLAP permite obter recomendações bem sucedidas. A validação, portanto, consiste em avaliar se o modelo apresentado é adequado para o seu propósito.

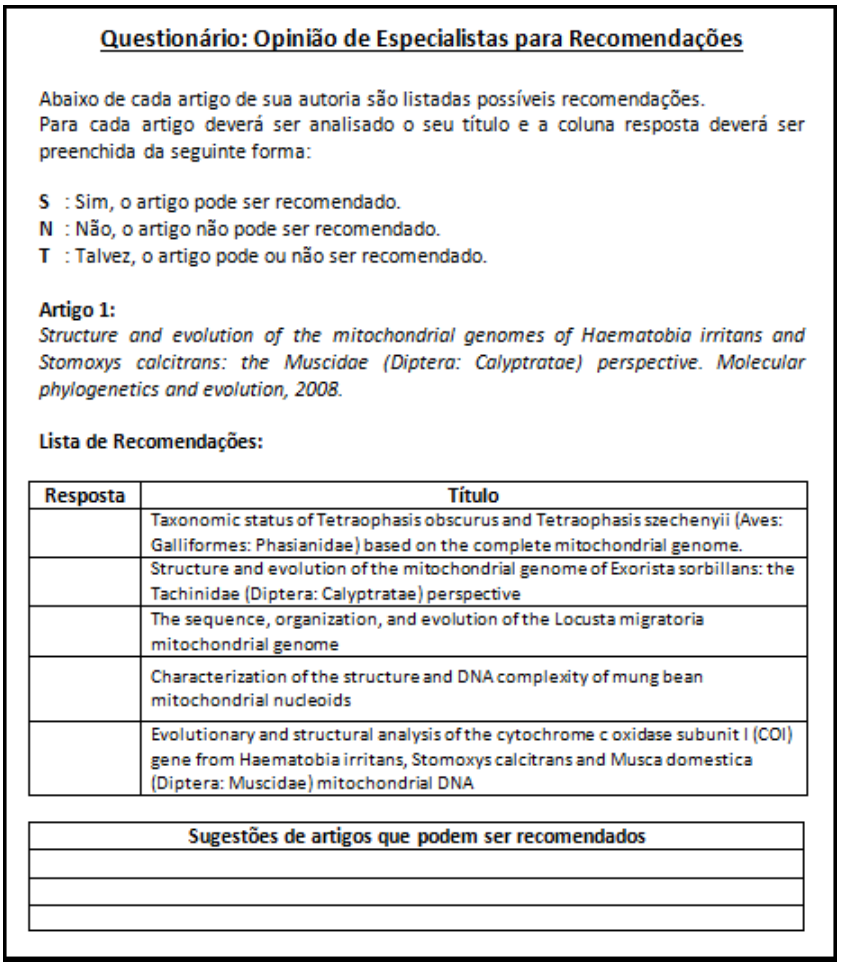
As validações experimentais são divididas em duas partes. A primeira avalia o conjunto de artigos recomendados após a seleção de artigos dos especialistas e por meio desta validação é possível determinar de forma experimental, isto é, sem a presença de

um usuário analista, os melhores parâmetros de personalização para o estudo de caso. A segunda, analisa o conjunto de artigos recomendados após o processo de expansão da recomendação.

6.2.1 Conjunto de Artigos Recomendados

Para cada especialista, foram selecionados ao menos dois artigos de sua autoria indexados pela PubMed e foi proposto que eles classificassem possíveis artigos recomendados a partir de seus trabalhos em duas categorias: (1) um conjunto de artigos conhecidos por eles e que são relacionados com seus trabalhos e (2) um conjunto de artigos que não são relacionados com seus trabalhos.

Para permitir que os especialistas expressassem suas opiniões, foi elaborado o formulário ilustrado na Figura 19. Entre os trabalhos relacionados dos especialistas os artigos mostrados na Tabela 20 do Apêndice B foram usados nas validações realizadas nas próximas sessões.



Questionário: Opinião de Especialistas para Recomendações

Abaixo de cada artigo de sua autoria são listadas possíveis recomendações.
Para cada artigo deverá ser analisado o seu título e a coluna resposta deverá ser preenchida da seguinte forma:

S : Sim, o artigo pode ser recomendado.
N : Não, o artigo não pode ser recomendado.
T : Talvez, o artigo pode ou não ser recomendado.

Artigo 1:
Structure and evolution of the mitochondrial genomes of Haematobia irritans and Stomoxys calcitrans: the Muscidae (Diptera: Calyptratae) perspective. Molecular phylogenetics and evolution, 2008.

Lista de Recomendações:

Resposta	Título
	Taxonomic status of Tetraophasis obscurus and Tetraophasis szechenyii (Aves: Galliformes: Phasianidae) based on the complete mitochondrial genome.
	Structure and evolution of the mitochondrial genome of Exorista sorbillans: the Tachinidae (Diptera: Calyptratae) perspective
	The sequence, organization, and evolution of the Locusta migratoria mitochondrial genome
	Characterization of the structure and DNA complexity of mung bean mitochondrial nucleoids
	Evolutionary and structural analysis of the cytochrome c oxidase subunit I (COI) gene from Haematobia irritans, Stomoxys calcitrans and Musca domestica (Diptera: Muscidae) mitochondrial DNA

Sugestões de artigos que podem ser recomendados

Figura 19 – Formulário para obter as opiniões dos especialistas

Para cada trabalho, o especialista apontou possíveis artigos recomendados que são

relacionados ao artigo de sua autoria. Estes artigos recomendados pelos especialistas foram da mesma forma armazenados no *Data Warehouse* e estão relacionados na Tabela 21 do Apêndice B. Quando o especialista apontou que o artigo não seria recomendado e que talvez fosse recomendado, seus valores de similaridades também foram verificados, para saber se os valores estavam abaixo do limite de similaridade. Em todos os casos, respostas negativas e incertezas sobre a recomendação, o valor estava abaixo da *similaridade mínima*, portanto, os artigos não seriam recomendados, demonstrando assim que havia coerência entre o resultado da recomendação e a percepção dos especialistas.

Discussão dos Resultados do Experimento

A finalidade da experimentação é avaliar o algoritmo, a fim de mostrar a qualidade das agregações realizadas na recomendação. Foram coletados os resultados da função de recomendação de cinco artigos indicados pelos especialistas. Além disso, o objetivo era identificar os parâmetros de *similaridade mínima* e D_{MAX} mais adequados e que oferecessem o melhor resultado de agregação, ou seja, efetuar a calibração do algoritmo para este experimento.

A seguir são apresentadas nas Tabelas 6, 7, 8, 9 e 10 as matrizes de confusão da seguinte configuração: a *similaridade mínima* de 0,50 e *fator de ontologia* igual a 4.

	Pred=pos	Pred=neg
Classe=pos	4	1
Classe=neg	0	127

Tabela 6 – Matriz de confusão do Artigo 1 - artigos recomendados

	Pred=pos	Pred=neg
Classe=pos	4	0
Classe=neg	0	128

Tabela 7 – Matriz de confusão do Artigo 2 - artigos recomendados

	Pred=pos	Pred=neg
Classe=pos	6	0
Classe=neg	3	123

Tabela 8 – Matriz de confusão do Artigo 3 - artigos recomendados

	Pred=pos	Pred=neg
Classe=pos	8	0
Classe=neg	6	118

Tabela 9 – Matriz de confusão do Artigo 4 - artigos recomendados

	Pred=pos	Pred=neg
Classe=pos	7	1
Classe=neg	4	120

Tabela 10 – Matriz de confusão do Artigo 5 - artigos recomendados

Após as experiências realizadas e analisadas nas matrizes de confusão mostradas acima, a tabela abaixo (Tabela 11) foi elaborada com os valores de *precision*, *recall* e *F1-measure*. A configuração mostrada nesta tabela apresenta os melhores resultados. As outras duas melhores configurações são mostradas na Tabela 12 e Tabela 13.

Artigo	**Precision**	**Recall**	**F1-Measure**
1	1,000	0,800	0,889
2	1,000	1,000	1,000
3	0,667	1,000	0,800
4	0,571	1,000	0,727
5	0,636	0,875	0,737
Médias	0,775	0,935	0,831

Configurações	
Similaridade Mínima	0,50
Fator de Ontologia	4

Tabela 11 – Configuração 1 para a recomendação

Artigo	**Precision**	**Recall**	**F1-Measure**
1	1,000	0,800	0,889
2	1,000	1,000	1,000
3	0,750	0,750	0,750
4	0,000	0,000	0,000
5	1,000	0,800	0,889
Médias	0,750	0,670	0,706

Configurações	
Similaridade Mínima	0,50
Fator de Ontologia	3

Tabela 12 – Configuração 2 para a recomendação

Aumentando D_{MAX}, o número de artigos relacionados aumenta, mas também aumenta o número de candidatos no *conjunto de itens* classificados como falsos-positivos. Este fato afeta diretamente o valor de *precision* que é reduzido. À medida que o número de artigos relacionados sobe, o valor de *recall* também aumenta.

Foi observado, também, que para valores de D_{MAX} menores, o valor de similaridade mínima também deveria ser menor. O valor de D_{MAX} sendo menor, a distância entre os

Artigo	Precision	Recall	F1-Measure
1	0,200	0,500	0,286
2	0,167	0,250	0,200
3	0,400	1,000	0,571
4	0,571	0,889	0,696
5	0,538	0,875	0,667
Médias	0,375	0,703	0,484

Configurações	
Similaridade Mínima	0,55
Fator de Ontologia	5

Tabela 13 – Configuração 3 para a recomendação

termos na árvore torna-se mais restrito e, por conseguinte, a relação entre o conjunto de palavras-chave também diminui. Isto provoca um aumento na precisão, pois foi menor o número de falsos-positivos, e maior, o de falsos-negativos.

A intenção é propor para os especialistas sempre a melhor configuração dos parâmetros no caso da hierarquia de termos MeSH. A configuração empírica da parametrização realizada por meio das opiniões dos especialistas trouxe resultados muito interessantes, em especial, a Configuração 1, a qual atingiu os valores de *precision* e *recall* mais próximos de 1. O valor de *precision* alcançou o valor de 77,5% e *recall* 93,5%.

Comparando o resultado da medida *precision* com outros sistemas de recomendação, independente da base de dados utilizada, os valores obtidos são válidos nos experimentos da configuração 1 e 2, respectivamente 77,5% e 75%.

De acordo com os resultados experimentais de [WENG; LIN; CHEN, 2009] seu valor de *precision* médio foi de 77%, já em [ADOMAVICIUS et al., 2005] a medida *precision* tinha uma variação de 41,5% à 75,4%. Em [AHMED et al., 2014] os valores de *precision* estão entre 81% e 23%. [GIACOMETTI; MARCEL; NEGRE, 2008] aponta que 70% é um valor de *precision* suficientemente bom.

Os valores de *recall* em [ADOMAVICIUS et al., 2005] estão entre 27% e 62,3% e em [AHMED et al., 2014] estão entre 19% e 42%. Em termos de *recall* as três configurações propostas são válidas de acordo com os artigos que apontam o resultado desta medida. As medidas de recall calculadas nas três configurações são respectivamente 93,5%, 67% e 70,3%.

6.2.2 Expansão da Recomendação

Após a validação dos parâmetros de configuração da recomendação, é necessário verificar o grupo de documentos formados pelos documentos recomendados e relacionados, a fim de observar se os valores de *precision*, *recall* e *F1-measure* sofreram alterações. O objetivo desta validação é verificar se na expansão da recomendação existe a presença ou não de ruídos que possam prejudicar a resposta referente a recomendação.

A recomendação pela expansão do conjunto de documentos recomendados, traz

um novo conjunto de documentos chamado de documentos relacionados. Para verificar o comportamento do novo conjunto de documentos foi utilizado o mesmo conjunto de documentos relacionados pelos especialistas (Tabela 21 do Apêndice B).

Discussão dos Resultados do Experimento

A seguir são apresentadas nas Tabelas 14, 15, 16, 17 e 18, as matrizes de confusão da seguinte configuração: frequência (%) de 0,50 e D_{MAX} igual a 4 utilizando o primeiro nível de artigos relacionados.

	Pred=pos	Pred=neg
Classe=pos	5	0
Classe=neg	7	120

Tabela 14 – Matriz de confusão do Artigo 1 - artigos relacionados

	Pred=pos	Pred=neg
Classe=pos	4	0
Classe=neg	6	122

Tabela 15 – Matriz de confusão do Artigo 2 - artigos relacionados

	Pred=pos	Pred=neg
Classe=pos	6	0
Classe=neg	8	118

Tabela 16 – Matriz de confusão do Artigo 3 - artigos relacionados

	Pred=pos	Pred=neg
Classe=pos	8	0
Classe=neg	11	113

Tabela 17 – Matriz de confusão do Artigo 4 - artigos relacionados

	Pred=pos	Pred=neg
Classe=pos	7	1
Classe=neg	7	117

Tabela 18 – Matriz de confusão do Artigo 5 - artigos relacionados

Neste experimento, que resultou nos dados da Tabela 19, foi verificado que se o artigo que o especialista apontou não está nos documentos recomendados, certamente ele está nos documentos relacionados, por esta razão o *recall* aumentou (de 0,935 para 0,975) e consequentemente a medida *precision* diminuiu (de 0,775 para 0,433).

O *recall* aumenta já que mais artigos são recomendados, além daqueles que são apontados pelo especialista. Em relação a medida de *recall* da expansão da recomendação, ela possui um valor que está na faixa dos valores de *recall* apontados pelos trabalhos de [ADOMAVICIUS et al., 2005] e [AHMED et al., 2014].

Artigo	**Precision**	**Recall**	**F1-Measure**
1	0,417	1,000	0,588
2	0,400	1,000	0,571
3	0,429	1,000	0,600
4	0,421	1,000	0,593
5	0,500	0,875	0,636
Médias	0,433	0,975	0,598

Configurações	
Similaridade Mínima	0,50
Fator de Ontologia	4

Tabela 19 – Configuração 1 para expansão da recomendação

A expansão da recomendação permitiu que nenhum artigo relacionado seja excluído do grupo de artigos que tem assuntos semelhantes ao documento selecionado. A expansão por outro lado trará muitos artigos com diferentes graus de similaridade. Neste sentido, para facilitar o usuário gestor a distinguir aqueles artigos que pertencem ao conjunto recomendados daqueles que pertencem ao conjunto relacionados, eles são marcado na *interface de recomendação* por diferentes gradações de cores, como mostrado no Capítulo 4.

6.3 Sumário

O protótipo PAMDES, desenvolvido com base nas definições do modelo multidimensional e no algoritmo de recomendação, permitiu a validação de experimentos que mostraram o comportamento da recomendação frente a um estudo de caso real da PubMed.

Os algoritmos de recomendação trouxeram resultados satisfatórios, dado os valores obtidos na análise das métricas de avaliação e que foram comparados com outros trabalhos do mesmo gênero. A expansão da recomendação permitiu que os artigos relevantes fossem recomendados, o que também é um resultado interessante. Os resultados experimentais indicam que os melhores parâmetros iniciais são: *fator de ontologia* igual a 4 e *similaridade mínima* de 0,50. É proposto portanto, que esses valores de parametrização sejam utilizados como parâmetros iniciais para o pré-processamento, no entanto, os parâmetros podem ser ajustados e um novo pré-processamento poderá ser realizado.

Em um caso com um número maior de registros, seria interessante elaborar uma estratégia de processamento distribuído para realizar a primeira carga de dados, já que esta envolveria não só a inserção dos documentos na base de dados, mas também todo o pré-processamento das recomendações. Para as demais cargas de dados, ou seja, quando novos artigos forem inseridos na base de documentos, sendo então cargas de menor número de documentos, ter uma rotina incremental do cálculo de recomendações seria a melhor opção.

7 Conclusão

O principal objetivo deste trabalho foi atingido. Foi proposta uma abordagem OLAP para operação de agregação centrada no usuário, com a finalidade de recomendar documentos baseado em parâmetros definidos pelo usuário.

Com base na revisão sistemática [GONZÁLEZ; BERBEL, 2014] realizada para encontrar o estado da arte sobre OLAP textual, foi possível perceber que os trabalhos não viabilizavam a recomendação e personalização de consultas OLAP para documentos. Já, as propostas sobre recomendação e personalização, no entanto, estavam restritos a trabalhar com dados estruturados.

Para permitir a recomendação, foi elaborado um conjunto de algoritmos que permitem que a agregação OLAP utilize as preferências dos usuários. A recomendação mediante a agregação possibilitou a visualização de um cubo OLAP de documentos de texto, onde a partir da modelagem multidimensional, modelo POQT, foi possível agrupar e recomendar artigos, por meio da utilização de uma ontologia de domínio.

A arquitetura do modelo conceitual foi utilizada para o desenvolvimento do protótipo PAMDES. O estudo de caso foi baseado no banco de dados da PubMed e a semântica é representada utilizando a hierarquia de termos MeSH. Os resultados dos experimentos realizados mostraram que boas recomendações são possíveis utilizando o protótipo desenvolvido. As melhores configurações de recomendação foram apresentadas e podem ser utilizadas como parâmetro da recomendação inicial e posteriormente podem ser adaptados para refletir os objetivos da busca do usuário gestor.

As contribuições deste trabalho são:

1. a proposta de integração entre OLAP Textual e a recomendação personalizada de documentos;
2. uma representação formal dos conceitos de medida e dimensão para compor um modelo multidimensional para dados não estruturados com dados não estruturados como possível medida;
3. o conjunto de algoritmos para a recomendação que funciona usando semelhança semântica entre documentos por meio de uma ontologia de domínio;
4. a expansão da recomendação, o que permite o usuário gestor encontrar documentos que também possam ser relevantes para a sua consulta e
5. desenvolvimento do protótipo PAMDES que é um estudo de caso baseado em dados reais.

7.1 Publicações

Como parte deste trabalho, a autora desenvolveu os seguintes artigos:

O artigo **How to help end users to get better decisions? Personalizing OLAP Aggregation queries through Semantic Recommendation of Text Documents** foi aceito no *International Journal of Business Intelligence and Data Mining* (IJBIDM) em 2014. Atualmente, aguarda-se a publicação (2015).

O artigo **Considering Unstructured Data for OLAP: a Feasibility Study using a Systematic Review** foi publicado na Revista de Sistemas de Informação da Faculdade Salesiana Maria Auxiliadora, Edição Jul-Dez/2014.

Revisão Sistemática sobre OLAP Textual e as propostas de Recomendação e Personalização OLAP, artigo a ser submetido em 2015.

7.2 Trabalhos Futuros

Os trabalhos futuros incluem:

1. fornecer a capacidade de estabelecer mapeamentos semânticos entre conceitos definidos pelo usuário e representações hierárquicas;
2. permitir a adaptação da recomendação para outras áreas de domínio, por meio de diferentes ontologias;
3. desenvolver uma rotina ETL para inserir automaticamente documentos no *Data Warehouse* e tornar incremental o pré-processamento da recomendação;
4. no caso do protótipo, criar uma interface Web com o intuito de disponibilizar a recomendação para os usuários do PubMed e, além disso, permitir que os arquivos dos trabalhos recomendados possam ser baixados;
5. elaborar uma estratégia de processamento distribuído ou paralelo para realizar as cargas de documentos no *Data Warehouse* e o pré-processamento por documento dos parâmetros ajustáveis inicias para as recomendações;
6. criar uma interface para que o usuário gestor possa dar um *feedback* a respeito dos resultados da recomendação;
7. permitir a criação de grupos de documentos semanticamente similares, possibilitando assim agregações OLAP dinâmicas, ou seja, além dos dois grupos já criados (artigos relacionados e não relacionados) e
8. filtrar os resultados da expansão da recomendação de forma a aumentar o *precision*.

Referências

ADOMAVICIUS, G. et al. Incorporating contextual information in recommender systems using a multidimensional approach. *ACM Transactions on Information Systems*, New York, v. 23, n. 1, p. 103–145, jan. 2005. Disponível em: <http://doi.acm.org/10.1145/1055709.1055714>. Acesso em: 8 set. 2014. Citado 2 vezes nas páginas 96 e 98.

ADOMAVICIUS, G.; TUZHILIN, A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, Piscataway, v. 17, n. 6, p. 734–749, jun. 2005. Disponível em: <http://dx.doi.org/10.1109/TKDE.2005.99>. Acesso em: 8 set. 2014. Citado na página 36.

AHMED, E. B. et al. Smart: Semantic multidimensional group recommendations. *Multimedia Tools and Applications*, Dordrecht, p. 1–19, 2014. Citado 4 vezes nas páginas 51, 53, 96 e 98.

ARIYACHANDRA, T.; WATSON, H. Key organizational factors in data warehouse architecture selection. *Decision Support Systems*, Amsterdam, v. 49, p. 200–212, 2010. Citado na página 28.

BARCELLOS, C. et al. Sistema de recomendação acadêmico para apoio a aprendizagem - Universidade Federal do Rio Grande do Sul. Renote: Revista Novas Tecnologias na Educação, Porto Alegre, v.5, n.2, p. 1-10, 2007. Citado na página 37.

BELLATRECHE, L. et al. A personalization framework for OLAP queries. In: *ACM International Workshop on Data Warehousing and OLAP, 8., 2005, New York. Proceedings...* [s.n.], 2005. p. 9–18. Disponível em: <http://doi.acm.org/10.1145/1097002.1097005>. Acesso em: 27 ago. 2014. Citado 2 vezes nas páginas 48 e 53.

BENTAYEB, F.; FAVRE, C. Rok: Roll-up with the k-means clustering method for recommending OLAP queries. In: *Database and Expert Systems Applications*. [S.l.]: Springer Berlin Heidelberg, 2009, (Lecture Notes in Computer Science, v. 5690). p. 501–515. Citado 3 vezes nas páginas 23, 48 e 53.

BEPPLER, F. D. *Um modelo para recuperação e busca de informação baseado em ontologia e no círculo hermenêutico. 2008. 123 f.* Tese (Doutorado) — Universidade Federal de Santa Catarina, Florianópolis, 2008. Citado 2 vezes nas páginas 35 e 36.

BOUKRAA, D.; BOUSSAID, O.; BENTAYEB, F. OLAP operators for complex object data cubes. In: *Advances in Databases and Information Systems*. [S.l.]: Springer Berlin Heidelberg, 2010, (Lecture Notes in Computer Science, v. 6295). p. 103–116. Citado 3 vezes nas páginas 39, 45 e 53.

BOUSSAID, O. et al. Integration and dimensional modeling approaches for complex data warehousing. *Journal of Global Optimization*, Dordrecht, v. 37, p. 571–591, 2007. Citado na página 45.

CEMBALO, A.; PISANO, F. M.; ROMANO, G. An approach to document warehousing system lifecyle from textual ETL to multidimensional queries: A proof-of-concept prototype. In: *International Conference on Complex, Intelligent, and Software Intensive Systems, 6., 2012, Palermo. Proceedings...* [S.l.: s.n.], 2012. p. 828–835. Citado 3 vezes nas páginas 39, 47 e 53.

CHEN, Z.; GARCIA-ALVARADO, C.; ORDONEZ, C. Enhancing document exploration with OLAP. In: *IEEE International Conference on Data Mining, 2010, Sidney. Proceedings...* [S.l.: s.n.], 2010. p. 1407–1410. Citado 3 vezes nas páginas 40, 43 e 53.

COSTA, A. A. L. et al. Recomendação personalizada de conteúdo para suporte à aprendizagem informal no contexto da saúde - Universidade Federal Rural do Semi-Árido. Renote: Revista Novas Tecnologias na Educação, Porto Alegre, v.12, n.1, p. 1-10, 2014. Citado na página 38.

DEHURI, S. *Intelligent Techniques in Recommendation Systems: Contextual Advancements and New Methods*. [S.l.]: IGI Global, 2012. Citado na página 38.

GIACOMETTI, A.; MARCEL, P.; NEGRE, E. A framework for recommending OLAP Queries. In: *International Workshop on Data Warehousing and OLAP, 11., 2008, Napa Valley. Proceedings...* [S.l.]: ACM, 2008. p. 73–80. Citado 3 vezes nas páginas 49, 53 e 96.

GONZÁLEZ, S.; BERBEL, T. Considering unstructured data for OLAP: a feasibility study using a systematic review. *Revista de Sistemas de Informação da Faculdade Maria Auxiliadora*, n. 14, p. 26–35, jul./dez., 2014. Citado 4 vezes nas páginas 23, 24, 40 e 101.

HAN, J.; KAMBER, M. *Data Mining : Concepts and Techniques*. 2nd. ed. [S.l.]: Morgan Kaufmann, 2006. Citado 6 vezes nas páginas 15, 27, 31, 33, 34 e 90.

HAUSSER, R. *Foundations of Computational Linguistics: Human-Computer Communication in Natural Language*. 2nd. ed. [S.l.]: Springer, 2001. Citado na página 91.

INMON, W. H. W.; GLASSEY, J. D.; L., K. *Managing the Data Warehouse*. [S.l.]: Wiley, 1996. Citado na página 27.

ISHIKAWA, H.; OHTA, M.; KATO, K. Document warehousing: A document-intensive application of a multimedia database. In: *International Workshop on Research Issues in Data Engineering - Distributed Object Management, 14., 1998, Heidelberg. Proceedings...* [S.l.: s.n.], 2001. p. 25–31. Citado na página 39.

JANET, B.; REDDY, A. V. Cube index for unstructured text analysis and mining. In: *International Conference on Communication, Computing & Security, 11., 2011, Odisha. Proceedings...* [S.l.: s.n.], 2011. p. 397–402. Citado 2 vezes nas páginas 43 e 53.

JERBI, H. et al. Management of context-aware preferences in multidimensional databases. In: *International Conference on Digital Information Management, 3., 2008, London. Proceedings...* [S.l.: s.n.], 2008. p. 669–675. Citado 3 vezes nas páginas 48, 49 e 53.

JERBI, H. et al. Preference-based recommendations for olap analysis. In: *International Conference on Data Warehousing and Knowledge Discovery, 11., 2009, Linz. Proceedings...* [S.l.: s.n.], 2009. p. 467–478. Citado 3 vezes nas páginas 23, 49 e 53.

JESUS, D. N.; BRITO, P. F. Protótipo de um aplicativo de recomendação de artigos científicos para materiais didáticos. In: *Encontro de Computação e Informática de Tocantins, 15., 2011, Palmas. Anais...* [S.l.: s.n.], 2011. p. 228-237. Citado na página 37.

KANG, J.; CHOI, J. An ontology-based recommendation system using long-term and short-term preferences. In: *International Conference on Information Science and Applications, 2011, Pattaya. Proceedings...* [S.l.: s.n.], 2011. p. 1–8. Citado 3 vezes nas páginas 23, 36 e 37.

KIM, H. H.; PARK, S. S. Mediaviews: A layered view mechanism for integrating multimedia data. In: *International Conference on Object-Oriented Information Systems, 9., 2003, Geneva. Proceedings...* [S.l.: s.n.], 2003. p. 250–260. Citado na página 40.

KIMBALL, R.; ROSS, M. *The Data Warehouse Toolkit. The Complete Guide to Dimensional Modeling*. 2nd. ed. São Paulo: Makron Books, 2002. Citado 4 vezes nas páginas 28, 30, 31 e 77.

KOZMINA, N.; SOLODOVNIKOVA, D. Towards introducing user preferences in olap reporting tool. In: NIEDRITE, L.; STRAZDINA, R.; WANGLER, B. (Ed.). *BIR Workshops*. Springer, 2011. (Lecture Notes in Business Information Processing, v. 106), p. 209–222. Disponível em: <http://dblp.uni-trier.de/db/conf/bir/bir2011w.html#KozminaS11>. Acesso em: 8 set. 2014. Citado 2 vezes nas páginas 50 e 53.

LIN, C. et al. Text cube: Computing IR measures for multidimensional text database analysis. In: *IEEE International Conference on Data Mining, 8., 2008, Pisa. Proceedings...* [S.l.: s.n.], 2008. p. 905–910. Citado 2 vezes nas páginas 41 e 53.

MEDICAL LITERATURE ANALYSIS RETRIEVEL SYSTEM ONLINE. P*ub*M*ed: N*ational C*enter for* B*iotechnology* I*nformation*. 2014. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed>. Acesso em: 20 jan. 2013. Citado 2 vezes nas páginas 26 e 109.

MIDDLETON, S. E.; ALANI, H.; ROURE, D. D. Exploiting synergy between ontologies and recommender systems. *(CoRR*, cs.LG/0204012), 2002. Citado na página 36.

MUSTAPHA, N. et al. Modular ontological warehouse for adaptative information search. In: *Model and Data Engineering*. [S.l.]: Springer Berlin Heidelberg, 2012, (Lecture Notes in Computer Science, v. 7602). p. 79–90. Citado na página 35.

NATIONAL LIBRARY OF MEDICINE. M*e*SH*: M*edical* S*ubject* H*eadings*. 2014. Disponível em: <http://www.nlm.nih.gov/mesh>. Acesso em: 20 jan. 2013. Citado 3 vezes nas páginas 26, 56 e 109.

NEUMAYR, B.; ANDERLIK, S.; SCHREFL, M. Towards ontology-based olap: Datalog-based reasoning over multidimensional ontologies. In: *International Workshop on Data warehousing and OLAP, 15., 2012, Maui. Proceedings...* [S.l.: s.n.], 2012. p. 41–48. Citado 3 vezes nas páginas 35, 47 e 53.

PÉREZ, J. M.; BERLANGA, R.; ARAMBURU, M. J. A relevance model for a data warehouse contextualized with documents. *Information Processing and Management*, v. 45, n. 3, p. 356–367, 2009. Citado 3 vezes nas páginas 40, 42 e 53.

PÉREZ-MARTÍNEZ, J. M. et al. Contextualizing data warehouses with documents. *Decision Support Systems*, Amsterdam, v. 45, n. 1, p. 77–94, 2008. Citado na página 56.

PUJOLLE, G. et al. Multidimensional database design from document-centric XML documents. In: *Data Warehousing and Knowledge Discovery: 13th International Conference, DaWaK 2011, Toulouse, France, August 29-September 2,2011: Proceedings*. [S.l.: s.n.], 2011. p. 51–65. Citado 3 vezes nas páginas 39, 46 e 53.

RAVAT, F.; TESTE, O.; TOURNIER, R. OLAP Aggregation Function for Textual Data Warehouse. In: *International Conference on Enterprise Information Systems,9., 2007, Funchal. Proceedings...* [S.l.: s.n.], 2007. p. 151–156. Citado 11 vezes nas páginas 23, 25, 30, 34, 44, 53, 55, 57, 59, 74 e 81.

RAVAT, F. et al. A conceptual model for multidimensional analysis of documents. In: *Conceptual Modeling: ER 2007: 26th International Conference on Conceptual Modeling, Auckland, New Zealand, November 5-9, 2007: proceedings*. [S.l.]: Springer Berlin, 2007. Citado 5 vezes nas páginas 39, 40, 41, 46 e 53.

RAVAT, F. et al. Top_keyword: An aggregation function for textual document OLAP. In: *Data Warehousing and Knowledge Discovery: 10th International Conference, DaWaK 2008, Turin, Italy, September 2-5, 2008: proceedings*. [S.l.: s.n.], 2008. p. 55–64. Citado 3 vezes nas páginas 41, 53 e 55.

SKOUTAS, D.; SIMITSIS, A. Ontology-based conceptual design of ETL processes for both structured and semi-structured data. *International Journal on Semantic Web and Information Systems*, v. 3, n. 4, p. 1–24, 2007. Citado 2 vezes nas páginas 25 e 35.

TAN, P.-N.; STEINBACH, M.; KUMAS, V. *Introdução ao Data Mining: Mineração de Dados*. [S.l.]: Editora Moderna, 2009. Citado 2 vezes nas páginas 32 e 90.

THOLLOT, R.; KUCHMANN-BEAUGER, N.; AUFAURE, M. aude. Semantics and usage statistics for multi-dimensional query expansion. In: *International Conference of Database Systems for Advanced Applications, 17., 2012, Busan. Proceedings...* [S.l.: s.n.], 2012. p. 250–260. Citado 4 vezes nas páginas 48, 50, 53 e 91.

TSENG, F. S. C.; CHOU, A. Y. H. The concept of document warehousing for multi-dimensional modeling of textual-based business intelligence. *Decision Support Systems, Amsterdam*, v. 42, p. 727–744, 2006. Citado na página 39.

VIEIRA, F.; NUNES, M. Dica: Sistema de recomendação de objetos de aprendizagem baseado em conteúdo. Revista Scientia Plena, v.8, p. 1-10, 2012. Citado na página 37.

WENG, S.-S.; LIN, B.; CHEN, W.-T. Using contextual information and multidimensional approach for recommendation. *Expert Systems with Applications, New York*, v. 36, n. 2, Part 1, p. 1268 – 1279, 2009. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0957417407005787>. Acesso em: 8 set. 2014. Citado na página 96.

ZHANG, D. et al. Topic modeling for OLAP on multidimensional text databases: Topic cube and its applications. *Statistical Analysis and Data Mining*, v. 2, n. 5-6, p. 378–395, 2009. Citado 2 vezes nas páginas 42 e 53.

ZHUOLUN, Z.; SUFEN, W. A framework model study for ontology-driven ETL processes. In: *Proceedings of International Conference on Wireless Communications, Networking and Mobile Computing, 4., 2008, Dalian, Proceedings...* [S.l.: s.n.], 2008. p. 1–4. Citado 2 vezes nas páginas 35 e 36.

APÊNDICE A – Contexto: PubMed

O PubMed é uma ferramenta para pesquisa da literatura científica da área de ciências biológicas, que é disponibilizada pelo NCBI (*National Center for Biotechnology Information*) da Biblioteca de Medicina dos Estados Unidos [NATIONAL LIBRARY OF MEDICINE, 2014] (NLM). Essa ferramenta permite acesso ao banco de dados gratuito de artigos e publicações fornecidos pelo MEDLINE [MEDICAL LITERATURE ANALYSIS RETRIEVEL SYSTEM ONLINE, 2014].

Atualmente a PubMed possui mais que 24 milhões de citações da literatura biomédica do MEDLINE, revista científicas e livros on-line. As citações podem incluir links para o conteúdo completo dos textos da Central da PubMed ou dos *web sites* dos editores. O banco de dados da MEDLINE contém citações a partir do final dos anos 1940 até o presente, e possui também alguns materiais mais antigos.

A primeira lista oficial de títulos de assuntos publicados pela NLM foi em 1954 sob o título de *Subject Heading Authority List*. Ela foi baseada na lista de autoridade interna que tinha sido usada para a publicação do *Current List of Medical Literature*, o qual por sua vez, tinha incorporado os títulos do *Library's Index-Catalogue* e do 1940 *Quarterly Cumulative Index Medicus Subject Headings*. Com o início do *Index Medicus (New Series)* em 1960, um novo e completamente revisado *Medical Subject Headings* surgiu.

Listas categorizadas de termos *Medical Subject Headings* foram impressas pela primeira vez em 1963 e continha treze categorias principais e um total de cinquenta e oito grupos separados em subcategorias e categorias principais. Essas listas categorizadas possibilitaram que os usuários pudessem encontrar muitos termos mais relacionados com o seu termo de pesquisa do que acontecia antes na antiga estrutura de referência cruzada. No mesmo ano, a segunda edição do *Medical Subject Headings* continha 5700 descritores, em comparação com 4400, na edição 1960. Das posições usadas na lista de 1960, 113 foram substituídos por termos mais atualizados. Em contraste com a edição de 2015 de MeSH que contém 27455 descritores.

A seguir são descritos os três tipos básicos de registros MeSH: os descritores, os qualificadores e os registros de conceitos suplementares.

Descritores: também conhecidos como principais títulos (*Main Headings*), descritores são usados para indexar as citações na base de dados MEDLINE da NLM. A maioria dos descritores indica o assunto de um item, isto é, são as palavras-chave que descrevem o conteúdo dos artigos. No caso de um artigo de um periódico, os descritores irão definir do que se trata o artigo. O descritor MeSH (*Medical Subject Headings*), portanto, é

o vocabulário controlado da NLM, que é utilizado para indexações de publicações, catalogação e busca de informações de documentos da biomedicina e saúde no PubMed. Muitos sinônimos, quase-sinônimos e conceitos intimamente relacionados estão incluídos como termos de entrada para ajudar o usuário a encontrar o descritor MeSH mais relevante para o assunto que está procurando. Nos bancos de dados *on-line* da NLM, muitos termos digitados pelos pesquisadores são automaticamente mapeados para descritores MeSH para facilitar a obtenção de informações relevantes.

Qualificadores: há 83 qualificadores, também conhecidos como subtítulos (*Subheadings*), utilizados para a indexação e catalogação em conjunto com os descritores. Qualificadores podem ser um meio conveniente de agrupar as citações que estão preocupadas com um aspecto particular de um assunto. Os qualificadores são atualizados anualmente.

Registros de conceitos suplementares: conhecidos como *Supplementary Concept Records* (SCR), chamado anteriormente *Supplementary Chemical Records*, são usados para indexar produtos químicos, drogas e outros conceitos, tais como, as doenças raras para o MEDLINE. Ao contrário dos descritores, os SCRs não tem números correspondentes a sua posição na árvore, no entanto, cada SCR está ligado a um ou mais descritores, geralmente mais amplo, pelo *Heading Mapped To field* no SCR. SCRs são atualizados semanalmente, ao contrário dos registros de descritores e qualificadores, que são geralmente atualizados em uma base anual. Existem atualmente mais de 200 mil registros SCR, com mais de 505 mil termos SCR.

Os descritores MeSH são usados pela NLM para indexação de artigos de 5.400 revistas biomédicas mais importantes do mundo para o banco de dados da MEDLINE/ PubMed. Eles também são usados em bancos de dados produzidos pela NLM, os quais incluem a catalogação de livros, documentos e audiovisuais adquiridos pela Biblioteca. Da mesma forma, as consultas de busca usam MeSH vocabulário para encontrar itens de um tópico desejado.

A equipe da seção de *Medical Subject Headings* revisa e atualiza continuamente o vocabulário MeSH. Funcionários especialistas da área são responsáveis pelas áreas de ciências da saúde em que têm conhecimento e experiência. Além de receber sugestões dos indexadores, os funcionários: (1) coletam novos termos e como esses aparecem na literatura científica ou em áreas emergentes de investigação; (2) definem esses termos no contexto do vocabulário existente e (3) recomendam a sua adição ao MeSH. Profissionais de várias disciplinas também são consultados sobre as mudanças organizacionais gerais e uma coordenação é mantida com vários vocabulários especializados.

Para permitir a análise de medidas textuais, foi utilizada uma representação hierárquica do domínio de conceitos. A hierarquia de domínio dos conceitos é representada por meio de uma estrutura de árvore, chamada de MeSH *Tree*, onde cada nó representa

um termo MeSH, e cada aresta representa o seu relacionamento com os demais termos da ontologia.

A árvore de termos MeSH é organizada em 16 ramos principais, onde são distribuídos 55611 termos (nós), além de um nó raiz artificial (*Root*). O nó raiz foi criado para unificar as representações hierárquicas de cada ramo (subárvore) que compõem a ontologia. Os termos MeSH são desde termos ligados a anatomia, organismos e doenças até tecnologias, características de publicações e informações geográficas. Cada categoria é dividida em subcategorias e dentro de cada subcategoria, os descritores estão dispostos hierarquicamente do termo mais geral para o mais específicos em doze níveis de hierarquia.

Cada descritor MeSH aparece em pelo menos uma posição da árvore e podem aparecer em quantas posições adicionais forem necessárias. Cada descritor é seguido pelo número que indica sua localização na árvore. Para cada título da PubMed são apresentados um ou mais termos MeSH.

APÊNDICE B – Opinião dos Especialistas

A Tabela 20 relaciona os artigos que foram utilizados nas validações dos algoritmos.

Artigo	Título
1	Structure and evolution of the mitochondrial genomes of Haematobia irritans and Stomoxys calcitrans: the Muscidae (Diptera: Calyptratae) perspective
2	Characterization of the screwworm flies Cochliomyia hominivorax and Cochliomyia macellaria by PCR-RFLP of mitochondrial DNA
3	Evaluation of morpho-anatomical and chemical differences between varieties of the medicinal plant Casearia sylvestris Swartz
4	Structural and biological characterization of Nattectin, a new C-type lectin from the venomous fish Thalassophryne nattereri
5	Genetic structure in fragmented populations of Solanum lycocarpum A. St.-Hil. with distinct anthropogenic histories in a Cerrado region of Brazil

Tabela 20 – Artigos dos especialistas utilizados para análise

A Tabela 21 relaciona as artigos recomendados pelos especialistas para cada artigo da Tabela 20.

Artigo	Recomendados pelos especialistas
1	Characterization of the screwworm flies Cochliomyia hominivorax and Cochliomyia macellaria by PCR-RFLP of mitochondrial DNA
	The sequence, organization, and evolution of the Locusta migratoria mitochondrial genome
	Identification of screwworm species by polymerase chain reaction-restriction fragment length polymorphism
	PCR-RFLP based method for molecular differentiation of sand fly species Phlebotomus argentipes, Phlebotomus papatasi, and Sergentomyia babu found in India
	Structure and evolution of the mitochondrial genome of Exorista sorbillans: the Tachinidae (Diptera: Calyptratae) perspective
2	Structure and evolution of the mitochondrial genomes of Haematobia irritans and Stomoxys calcitrans: the Muscidae (Diptera: Calyptratae) perspective
	The sequence, organization, and evolution of the Locusta migratoria mitochondrial genome
	Identification of screwworm species by polymerase chain reaction-restriction fragment length polymorphism

Continua na página seguinte

Tabela 21 – *Continuação da página anterior*

Artigo	Recomendados pelos especialistas
	PCR-RFLP based method for molecular differentiation of sand fly species Phlebotomus argentipes, Phlebotomus papatasi, and Sergentomyia babu found in India
3	Habitat area and structure affect the impact of seed predators and the potential for coevolutionary arms races
	Functional extinction of birds drives rapid evolutionary changes in seed size
	Are tortoises important seed dispersers in Amazonian forests?
	Diterpenoids from Casearia sylvestris
	Characterization of trace elements in Casearia medicinal plant by neutron activation analysis
	Population genetic structure of the endangered tropical tree species Caryocar brasiliense, based on variability at microsatellite loci
4	Habitat area and structure affect the impact of seed predators and the potential for coevolutionary arms races
	Functional extinction of birds drives rapid evolutionary changes in seed size
	Structure and evolution of the mitochondrial genome of Exorista sorbillans: the Tachinidae (Diptera: Calyptratae) perspective
	Molecular cloning and characterization of a C-type lectin in roughskin sculpin (Trachidermus fasciatus)
	Characterisation of local inflammatory response induced by Thalassophryne nattereri fish venom in a mouse model of tissue injury
	Are tortoises important seed dispersers in Amazonian forests?
	Diterpenoids from Casearia sylvestris
	Genetic structure in fragmented populations of Solanum lycocarpum A. St.-Hil. with distinct anthropogenic histories in a Cerrado region of Brazil
5	Functional extinction of birds drives rapid evolutionary changes in seed size
	Are tortoises important seed dispersers in Amazonian forests?
	Evaluation of morpho-anatomical and chemical differences between varieties of the medicinal plant Casearia sylvestris Swartz
	Population genetic structure of the endangered tropical tree species Caryocar brasiliense, based on variability at microsatellite loci
	Structural and biological characterization of Nattectin, a new C-type lectin from the venomous fish Thalassophryne nattereri
	Diterpenoids from Casearia sylvestris
	Characterization of trace elements in Casearia medicinal plant by neutron activation analysis

Tabela 21 – Artigos recomendados pelos especialistas