Universidade Federal do Rio de Janeiro

Programa de Pós-Graduação em Informática

Fábio Martins Heuseler

Uma abordagem multifacetada para exploração integrada de dados estruturados e não-estruturados em ambientes OLAP

Dissertação de Mestrado, apresentada ao Programa de Pós-graduação em Informática, Instituto de Matemática, Núcleo de Computação Eletrônica, Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Informática.

Orientadoras: Maria Luiza Machado Campos
Vanessa Braganholo Murta

Rio de Janeiro

2010

Universidade Federal do Rio de Janeiro

Programa de Pós-Graduação em Informática

Fábio Martins Heuseler

Uma abordagem multifacetada para exploração integrada de dados estruturados e não-estruturados em ambientes OLAP

Dissertação de Mestrado, apresentada ao Programa de Pós-graduação em Informática, Instituto de Matemática, Núcleo de Computação Eletrônica, Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Informática.

Orientadoras: Maria Luiza Machado Campos
Vanessa Braganholo Murta

Rio de Janeiro

2010

H595 Heuseler, Fábio Martins

Uma abordagem multifacetada para exploração integrada de dados estruturados e não-estruturados em ambientes OLAP / Fábio Martins Heuseler.-- 2010.
131 f.: il.

Dissertação (Mestrado em Informática) – Universidade Federal do Rio de Janeiro. Instituto de Matemática. Núcleo de Computação Eletrônica, 2010.

Orientadoras: Maria Luiza Machado Campos; Vanessa Braganholo Murta

1. Dados não-estruturados. –Teses. 2. Data Warehousing - Teses. 3. Exploração OLAP Multifacetada – Teses. I. Maria Luiza Machado Campos (Orient.). II. Vanessa Braganholo Murta (Orient.). III. Universidade Federal do Rio de Janeiro . Instituto de Matemática. Núcleo de Computação Eletrônica. IV. Título.

CDD.

Fábio Martins Heuseler

Uma abordagem multifacetada para exploração integrada de dados estruturados e não-estruturados em ambientes OLAP

Dissertação de Mestrado, apresentada ao Programa de Pós-graduação em Informática, Instituto de Matemática, Núcleo de Computação Eletrônica, Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Informática.

Aprovada em: Rio de Janeiro, 24 de fevereiro de 2010.

_____

Profª. Maria Luiza Machado Campos, Ph.D. – PPGI/UFRJ

_____

Profª. Vanessa Braganholo Murta, D.Sc. – PPGI/UFRJ

_____

Profª. Jonice de Oliveira Sampaio, D.Sc. – PPGI/UFRJ

_____

Profª. Fernanda Araújo Baião, D.Sc. – DIA/UNIRIO

_____

Prof. Asterio Kiyoshi Tanaka, Ph.D. – DIA/UNIRIO

AGRADECIMENTOS

À Deus, por permitir meu convívio com as pessoas abaixo citadas e também aquelas que ficaram de fora, não por importância menor, mas em razão de esquecimento causado pelas preocupações inerentes ao trabalho desenvolvido.

À minha filha, Maria Fernanda, razão da minha vida.

À minha esposa e melhor amiga, Paloma. Que a humanidade aprenda a amar tanto quanto nos amamos...

Aos meus pais Conceição e Ruben, que me deram ao longo da vida o que tinham de melhor: seu amor.

À minha querida avó Lourdes. Que sua dedicação e força sirvam de exemplos para todos.

Ao grande mestre de minha vida: meu avô Artur. Que Deus me permita ser 10% do Homem que você foi.

Aos meus queridos irmãos, João e Anna.

Aos meus avós paternos, cuja presença sinto até hoje, mesmo não estando mais fisicamente entre nós..

Aos amigos do Serpro, que me deram suporte no início desta caminhada e aos amigos da Petrobras, que me ajudaram a concluí-la.

Aos amigos de toda vida, cujos nomes não precisam ser citados, pois estes sabem que fazem parte desta conquista.

À Maria Luiza e Vanessa, meu muito obrigado pelas horas de dedicação. Obrigado pela paciência e carinho sempre demonstrados.

Às professoras Jonice Sampaio e Fernanda Baião e ao professor Asterio Tanaka, meu agradecimento por aceitarem fazer parte da banca deste trabalho, dedicando seu tempo ao seu aprimoramento.

RESUMO

HEUSELER, Fábio Martins. **Uma abordagem multifacetada para exploração integrada de dados estruturados e não-estruturados em ambientes OLAP**. 2010. 131 f. Dissertação (Mestrado em Informática). - PPGI, Instituto de Matemática, Núcleo de Computação Eletrônica, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2010.

As soluções de Data Warehousing se consolidaram nas últimas décadas como importante estratégia para exploração de dados nas organizações, apoiando a tomada de decisão através da integração de dados oriundos de diversas fontes e suporte adequado a operações analíticas. Estas soluções, de forma geral, têm se concentrado no tratamento de dados estruturados, deixando de considerar um rico acervo de informações de natureza textual, na forma de documentos, e-mails, contratos e outras fontes. Em alguns cenários, a não utilização dessas informações pode acarretar em perdas na capacidade analítica sobre os recursos e atividades corporativas. Esta dissertação tem como objetivo apresentar uma abordagem e uma arquitetura de apoio à exploração conjunta de dados estruturados e não-estruturados, considerando a relação entre o papel das taxonomias facetadas em sistemas de gerenciamento de conteúdo não-estruturado e o papel das dimensões de um esquema dimensional em um universo analítico de dados estruturados. Além disso, os mecanismos propostos por esta arquitetura foram desenvolvidos e testados através da construção de um protótipo e sua aplicação no domínio de clínicas médicas, onde as vantagens de análises envolvendo os diversos tipos de informações foram evidenciadas.

ABSTRACT

HEUSELER, Fábio Martins. **Uma abordagem multifacetada para exploração integrada de dados estruturados e não-estruturados em ambientes OLAP**. 2010. 131 f. Dissertação (Mestrado em Informática). - PPGI, Instituto de Matemática, Núcleo de Computação Eletrônica, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2010.

Data Warehousing solutions have been consolidated in the last decades as an important strategy for data exploration in organizations, supporting the decision process through the integration of data from different sources, giving support to analytical operations. These solutions, in general, have been focusing on structured data, leaving behind a rich world of information, textual by nature, such as documents, e-mails, contracts and other sources. In some situations, not using this information can generate significant loss in analytical capability over corporate resources and activities. This dissertation has the objective of introducing an approach and an architecture that supports an integrated exploration of structured and unstructured data, considering the relation between the role played by faceted taxonomies in unstructured content management systems and the role played by dimensions of a dimensional model, in a structured analytical environment. Besides, the mechanisms proposed by this architecture were implemented and tested in a prototype experimented on an application of a medical clinic domain, where the advantages of analysis involving the different types of information were evidenced.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

API	Application Programming Interface
DCS	Dynamic Category Sets
DW	Data Warehouse
DTD	Document Type Definition
ETC	Extração Transformação e Carga
ETL	Extract Transform Load
HTML	HyperText Markup Language
HTTP	HyperText Transfer Protocol
JSP	Java Server Pages
MDX	Multidimensional Expressions
MVC	Model-View-Controller
OCR	Optical Character Recognition
ODBC	Open Data Base Connectivity
OLAP	On-line Analytical Processing
RI	Recuperação de Informação
RDF	Resource Description Framework
SCF	Sistemas de Classificação Facetada
SGC	Sistemas de Gerenciamento de Conteúdo
SGBD	Sistema Gerenciador de Banco de Dados
SQL	Structured Query Language
SOM	Self Organizing Map
VCR	Voice Character Recognition
XML	Extensible Markup Language

LISTA DE ANEXOS

SUMÁRIO

1. Introdução

A abordagem de Data Warehousing se consolidou no mercado como solução para análises gerenciais sobre os dados das empresas. Ao integrar diversas fontes de dados, de sistemas transacionais diversos e mesmo de origem externa, um sistema de Data Warehousing possibilita uma visão única dos dados da corporação, levando os dirigentes da mesma a tomadas mais rápidas e precisas de decisões (INMON, 2005; WATSON, WIXOM, 2007).

O Data Warehouse (DW), principal componente da abordagem, funciona, como o próprio nome sugere, como um grande armazém, no qual os dados de diversas fontes serão inseridos após passar por um longo e pesado processo de extração, limpeza e transformação. Esse processo é conhecido como Extração, Transformação e Carga (ETC). O banco de dados desse "armazém" é freqüentemente modelado de uma maneira específica, denominada "modelagem dimensional", que tem como propósito disponibilizar a informação de maneira que fatores como desempenho e acessibilidade sejam destacados (KIMBALL, ROSS, 2002).

Usualmente, os dados recuperados e disponibilizados no DW possuem a característica de serem estruturados, ou seja, os sistemas transacionais que os detêm os armazenam em bancos de dados ou estruturas de dados específicas. As informações que não estão neste escopo não são usualmente utilizadas diretamente no processo de tomada de decisão, a exemplo de dados contidos em atas de reuniões, e-mails, páginas HTML da Intranet (e da Internet) e até mesmo ligações telefônicas.

Ao não incorporar os dados semi ou não-estruturados, os dirigentes das companhias estão perdendo informações importantes em relação a sua empresa (TUCKER, 1999; McCALLUM, 2005). Inmon e Nesavich (2008) afirmam que aproximadamente 80% das informações disponíveis nas companhias estão armazenadas de maneira não-estruturada. Ao basearmos nossas operações analíticas somente nos dados estruturados, estamos aproveitando somente 20% das informações que teriam potencial para ajudar no processo decisório. Assim, por exemplo, ao construir um DW de uma companhia onde o atendimento ao cliente é crucial para a evolução desta no mercado, informações trocadas entre os atendentes e os clientes, seja por e-mail ou em uma chamada telefônica, são fontes muito importantes para análises de desempenho da empresa.

Em um hospital, milhares de diagnósticos estão disponíveis nas fichas dos pacientes. Essas informações são escritas, em sua maioria, de forma descritiva textual pelo médico e arquivadas junto com tantas outras. O benefício de se permitir uma visão analítica sobre essas informações seria imenso, uma vez que o médico teria ao seu alcance todo o histórico de diagnósticos e tratamentos, possibilitando ao mesmo uma definição e resolução mais rápida dos problemas de seus pacientes. Cada caso diagnosticado serviria de insumo para casos futuros, fazendo o hospital atingir um nível alto de excelência.

Para capturar estas informações descritivas, quando escritas manualmente, ferramentas de reconhecimento de caracteres (OCR) podem ser aplicadas. Já existem no mercado diversas destas, tais como: SimpleOCR [1], TopOCR [2] e FreeOCR [3]. Esta disponibilidade possibilitaria um tratamento digital e a integração destas informações aos sistemas computacionais corporativos, reduzindo a necessidade de digitação. Já para capturar as informações provenientes de ligações telefônicas, aplicativos de reconhecimento de voz (VCR) de grandes empresas, como IBM, Microsoft e Cisco podem ser utilizados, incorporando este tipo de informação no suporte ao processo decisório.

Ao realizar esta incorporação, necessitamos de mais mecanismos do que simplesmente disponibilizar as informações em meio digital. Como explicitado anteriormente, a natureza dos dados em um DW é estruturada e, portanto, a incorporação de dados semi- ou não-estruturados requer o uso conjunto de técnicas de tratamento de informação textual e de Data Warehousing. Mais especificamente, antes de passar pelo processo de ETC os documentos devem passar por rotinas de tratamento, que irão realizar um pré-processamento destes. Alguns dos processos mais conhecidos na área de Recuperação de Informação e Mineração de Textos devem ser aplicados como, por exemplo, a eliminação de palavras não-significativas e a eliminação de caracteres de formatação.

Paralelamente ao crescimento dos sistemas de apoio à decisão focados em análises sobre dados estruturados, podemos observar que o universo de dados não-estruturados está

[1] http://www.simpleocr.com/

[2] http://www.topocr.com

[3] http://freeocr.co.uk/

cada vez mais em foco nas pesquisas e nos trabalhos sobre técnicas de visualização e recuperação de informação (RI). É cada vez mais crescente a necessidade dos usuários de realizarem análises sobre grandes volumes de informações textuais presentes em um conjunto de documentos, denominado coleção, ou até mesmo sobre a grande massa de documentos que é a Web, ao invés de simples buscas para acesso a conteúdos específicos.

Ao se deparar com um universo repleto de informações, relevantes e não-relevantes para sua necessidade, o usuário precisa ter subsídios que possibilitem uma melhor exploração, levando-o a encontrar resultados satisfatórios para sua necessidade (MARCHIONINI, 2006). Neste ponto, a presença de uma estrutura de classificação eficiente dos documentos irá influenciar em muito o sucesso dos sistemas de busca. Esta classificação irá afetar diretamente os processos de indexação e exploração, sendo fator crucial para o sucesso.

Um esquema classificatório do tipo taxonomia permite organizar os documentos segundo uma hierarquia que possibilite ao usuário mapear o universo de informações presentes nestes em categorias. Segundo Tzitzikas, et. al (2004), uma taxonomia facetada é constituída, na verdade, por um conjunto de taxonomias, cada uma descrevendo um determinado domínio de conhecimento sobre um aspecto diferente. Em uma taxonomia facetada, as categorias são agrupadas em perspectivas macro, criando um nível de classificação mais abrangente e permitindo a redução de níveis de uma hierarquia. Além disto, a faceta possibilita ao usuário uma identificação mais rápida da classificação proposta. Esta abordagem de classificação identifica o conteúdo de um documento sem que seja necessária sua leitura completa para apoiar o processo de descoberta. Por serem de fácil entendimento, as taxonomias vêm sendo muito utilizadas como mecanismos de classificação (BRODER, 2009; BROUGHTON, 2006; SACCO, 2006; TZITZIKAS, ANALYTI, 2007; UDDIN, JANECEK, 2007), principalmente por sítios de comércio eletrônico, como a Amazon[4], por exemplo, nos quais os produtos a serem vendidos são classificados em uma taxonomia de acordo com sua natureza.

Observando o papel desempenhado por taxonomias facetadas na exploração de informações não-estruturadas e tendo identificado a necessidade de se incorporar estas ao processo analítico (apoiado por sistemas de Data Warehousing), aprofundamos nossos estudos

[4] http://www.amazon.com/

de maneira a encontrar um caminho que possibilite uma exploração conjunta destes universos. Grande parte das soluções estudadas para este trabalho está embasada na existência de um DW de documentos, ou seja, necessitam que uma estrutura de apoio seja construída a partir dos documentos que contêm as informações não-estruturadas, para a realização de uma exploração conjunta entre os dois mundos (não-estruturado e estruturado).

Algumas iniciativas de exploração conjunta entre estes dois universos vêm sendo estudadas e novas abordagens estão sendo propostas (INMON, NESAVICH, 2008), (MOREIRA, CORDEIRO, CAMPOS, 2009), (PEREZ, et. al, 2007), (BORDAWEKAR, LANG, 2005) e (McCABE, et. al, 2000). Em algumas das soluções, as análises são feitas separadamente para depois serem compiladas e analisadas de maneira conjunta. A construção desta ligação entre os dois universos é visivelmente um dos pontos críticos no problema em questão. Conjuntamente a esta ligação, deve ser disponibilizada ao usuário uma ou mais formas de visualização conjunta destes dados. A construção, ou o aprimoramento, de uma ferramenta que ofereça apoio para esta nova característica é um problema a ser abordado por trabalhos na área.

É importante ressaltar que os sistemas de gerenciamento de conteúdo, disponíveis atualmente, realizam um apoio à gerência do conteúdo não-estruturado através de mecanismos de indexação e recuperação. No entanto, estes exploram pouco características analíticas, não propiciando ao usuário facilidades de manipulação e visualização, como agregações e detalhamento de informações. Como exemplo, podemos citar iniciativas como (SANKAR, TALWAR, MITRA, 2002) e (PHILLIPS, et. al, 2005), as quais são baseadas em redes neurais e outras técnicas de mineração. Entretanto, estas ainda não estão em geral integradas segundo uma abordagem analítica mais ampla e de simples utilização, carecendo especialmente de uma estratégia de exploração conjunta com os dados estruturados associados.

A solução proposta por este trabalho tem como base a seguinte hipótese: taxonomias facetadas podem realizar um papel exploratório nos dois universos, permitindo uma exploração conjunta dos mesmos. Considerando a possibilidade de utilização de dados não-estruturados em ambientes analíticos e visualizando o potencial exploratório que uma taxonomia facetada pode oferecer em um universo de informações, tanto estruturadas como não-estruturadas, desenvolvemos nossa abordagem.

O objetivo desta dissertação é apresentar esta abordagem de apoio à exploração integrada de dados não-estruturados, presentes em uma coleção de documentos, e dados estruturados, já usualmente tratados em ambientes de DW. Os esforços foram feitos tendo como principal foco a exploração das semelhanças existentes entre o papel das dimensões de um esquema dimensional, em um universo analítico de dados estruturados, e o papel das taxonomias facetadas, em sistemas de gerenciamento de conteúdo não-estruturado, como mecanismos de exploração de perspectivas sobre seus respectivos domínios.

O texto deste trabalho está organizado conforme descrito a seguir. O capítulo 2 apresenta as principais iniciativas de exploração conjunta de dados estruturados e dados não-estruturados em ambientes analíticos, encontradas no estudo realizado para a elaboração deste trabalho. O capítulo 3 é dedicado a taxonomias facetadas, visando um maior detalhamento do assunto, crucial para o desenvolvimento da abordagem proposta, apresentada no capítulo 4. No capítulo 5, apresentamos a arquitetura concebida para a abordagem. Neste capítulo também é apresentado o protótipo construído, de maneira a implementar os mecanismos propostos pela arquitetura. No capítulo 6 apresentamos a aplicação do protótipo no domínio da medicina, mostrando dados sobre o processamento ocorrido e um exemplo ilustrativo de uma exploração conjunta dos universos (estruturado e não-estruturado). Além disto, o capítulo apresenta um estudo comparativo entre a abordagem proposta e outras soluções estudadas. Após este comparativo, no capítulo 7, as conclusões obtidas por este trabalho são apresentadas.

2. Dados estruturados e não-estruturados em ambientes analíticos

Vimos, no capítulo introdutório deste trabalho, que um mecanismo que classifique corretamente um universo de informação pode ajudar muito na exploração das informações por parte dos usuários. Um dos principais representantes de mecanismos classificatórios são as taxonomias facetadas. Podemos considerá-las como uma visão multidimensional sobre os documentos. De acordo com Uddin e Janecek (2007), a previsibilidade e a lógica apresentadas por um sistema que utilize taxonomias facetadas o torna compatível com os requisitos necessários a ambientes exploratórios, servindo de base para todos os métodos de recuperação da informação (BAEZA-YATES, RIBEIRO-NETO, 1999; MANNING, RAGHAVAN, SCHÜTZE, 2008).

Em se tratando de análises por sobre informações estruturadas, temos, na literatura atual, dois grandes nomes de autores que merecem destaque: Bill Inmon (INMON, 2005; INMON, 2009) e Ralph Kimball (KIMBALL, ROSS, 2002; KIMBALL et al, 2008). O primeiro defende a construção do DW de maneira *top-down*, ou seja, todo o negócio da empresa deve ser mapeado e modelado, construindo o DW corporativo. Após a construção deste, os Data Marts seriam então derivados. Data Marts são visões de cada área dentro da companhia, ou seja, um ou mais assuntos que possuam correlação são associados e disponibilizados ao usuário através de um subconjunto de informações passíveis de análise pelo mesmo. De acordo com Jukic (2006), um Data Warehouse combina dados de bases de dados operacionais de toda uma companhia, enquanto um Data Mart é geralmente menor, possuindo um foco em um determinado departamento ou assunto. Ralph Kimball, por outro lado, defende uma abordagem *bottom-up*, na qual seriam construídos primeiramente os Data Marts e, em uma segunda fase, estes seriam integrados, formando o DW da companhia.

Ao utilizar a abordagem proposta por Ralph Kimball, o desenvolvimento fica mais ágil, dando um retorno sobre o investimento de maneira mais rápida para o cliente. Entretanto, corremos o risco de formarmos Data Marts isolados, cuja integração com o todo se torna uma tarefa muito difícil, por vezes impossível. A abordagem de Bill Inmon possui como grande vantagem uma forte integração dos assuntos, uma vez que todos são modelados e

desenvolvidos em conjunto. Entretanto, esta abordagem pode levar um tempo muito grande para ser concluída, levando a uma espera demasiada para o cliente, gerando descontentamento por parte deste, podendo levar o projeto ao fracasso. Em seu trabalho, Jukic (2006) apresenta um estudo sobre estratégias de modelagem para projetos de DW, contendo inclusive uma análise sobre as diferenças entre as abordagens citadas.

A escolha da abordagem mais adequada irá depender das características da empresa e do porte do projeto em questão. Uma abordagem híbrida, na qual se modela o todo e se desenvolve por partes pode ser considerada uma "melhor prática" em relação à maioria dos projetos de DW.

Independente da metodologia adotada, o processo de ETC estará sempre presente. Tendo então os dados carregados no DW, o usuário irá precisar de uma ferramenta de acesso a estes. Existem no mercado diversas ferramentas, denominadas "ferramentas OLAP", que realizam este acesso. Podemos citar como exemplos o BusinessObjects[5], o MicroStrategy[6] e o Cognos[7]. Todas estas possuem características comuns que facilitam a manipulação das informações contidas no DW, possibilitando cruzamentos, sumarizações, detalhamentos, dentre outras operações analíticas.

Para a concepção e elaboração da solução proposta para este trabalho, foi realizado um estudo sobre abordagens e ferramentas, visando a apoiar a análise conjunta entre o universo estruturado e o não-estruturado. Nas seções 2.1, 2.2, 2.3 e 2.4 apresentamos as principais iniciativas encontradas na tentativa de exploração conjunta dos dois universos. A seção 2.5 contém considerações finais sobre as soluções estudadas.

[5] http://www.sap.com/solutions/sapbusinessobjects/index.epx

[6] http://www.microstrategy.com.br/

[7] http://www-01.ibm.com/software/data/cognos/

2.1 Data Warehouse 2.0™

"O DW 2.0™ é o resultado de uma evolução arquitetônica da informação" (INMON, 2007a). Fazem parte dessa evolução a incorporação dos dados não-estruturados ao ambiente analítico, a definição de um ciclo de vida para os dados e a incorporação de metadados, sendo estes componentes fundamentais da arquitetura. Esta solução é uma marca registrada por seu criador, Bill Inmon, com objetivos de mercado, fazendo uso de ferramentas como, por exemplo, o Forest RIM[8] e o SeePower [9].

A arquitetura do DW 2.0™ propõe, conforme mencionado, a incorporação de dados não-estruturados no processo de Data Warehousing. Esses dados podem ser provenientes de e-mails, conversas de telefone, documentos, planilhas, etc. Devido à natureza destes dados, que têm como essência o texto livre, incorporá-los simplesmente no mundo analítico, segundo Inmon, criaria um universo de análise muito pouco produtivo, sendo necessário que estes textos passem por uma série de tratamentos antes de serem incorporados efetivamente. É preciso separar o conteúdo realmente relevante para a empresa.

No processo de obtenção de e-mails, por exemplo, devem-se separar os conteúdos trocados em razão de simples conversas entre amigos daqueles que tratam de temas que sejam realmente de interesse da corporação. Além de sua relevância, é importante determinar qual o conteúdo deste texto, ou seja, dar significado a este para que a análise possa ganhar em riqueza. Nota-se claramente que a grande preocupação neste momento de obtenção dos dados é o do tratamento do conteúdo adquirido, de maneira a tornar este significativo para análises.

Para "entender" estes textos, existem duas abordagens: lingüística e temática. A abordagem lingüística trata o texto de acordo com a língua na qual ele está escrito, sendo sua utilização muito complexa e pouco flexível (uma análise deve ser feita para cada língua). A abordagem temática utiliza a análise de *strings* de caracteres para classificar o texto em determinadas categorias, que são definidas previamente de acordo com o objetivo da empresa. Uma vez tendo sido analisado, o texto pode ser armazenado no DW de diferentes maneiras: ponteiros simples para a fonte, o texto completo, os primeiros *n* caracteres, entre outros.

---

[8] http://www.textual-etl.com/

[9] http://www.compudigm.com/

De modo a realizar esta interpretação do conteúdo não-estruturado, algumas ferramentas e procedimentos são adotados (INMON, NESAVICH, 2008), como a utilização de glossários, processadores de sinônimos, remoção das *stop words*, processo de radicalização, disponibilização de termos alternativos e/ou relacionados, além de realizar a correção da grafia dos termos, caso o usuário cometa algum engano. Estes serão mais bem explicados a seguir.

Identificação de temas

Segundo Dagan e Church (1994), um glossário é uma lista de termos e suas respectivas traduções. No DW 2.0™, um glossário é utilizado para classificar os temas identificados nos dados não-estruturados. Este processo de identificação consiste em capturar e organizar os dados não-estruturados de maneira que estes possam ser acessados e analisados, com algum significado, de acordo com algum tópico criado em um ou mais temas no glossário.

Associação de sinônimos

Um processador de sinônimos irá permitir ao analista a associação de termos de modo a tornar as análises mais significativas. Em um universo médico, por exemplo, os termos “sangue” e “pressão” podem ser combinados em um só termo: “pressão sanguínea”. Desta maneira todas as ocorrências das partes serão analisadas como entradas do termo composto, dando mais significado e poder de análise.

Remoção de *Stop Words*

Stop Words são termos que não são relevantes para análises. Uma palavra que ocorra em 80% dos documentos de uma coleção não tem utilidade em termos de recuperação da informação (BAEZA-YATES, RIBEIRO-NETO, 1999). Na língua portuguesa, termos como “a”, “um”, “algum” não são significativos para análise, podendo ser retirados do processamento do texto. Através da construção de uma lista de termos “não desejáveis”, o analista reduz o escopo de sua pesquisa em uma coleção.

Radicalização

O processo de radicalização (*stemming)* visa o reconhecimento da raiz das palavras, de modo a relacionar termos que, apesar de possuírem grafias distintas, tenham o mesmo significado em uma busca realizada pelo usuário. Através deste processo, a busca por um termo vai retornar ocorrências, além do termo pretendido, de termos com a mesma raiz. Ao realizar uma busca por “correr”, seriam retornados resultados relacionados também a ocorrências de “correndo”, “correu”, “correria”, dentre outros termos que possuam a mesma raiz. Além de enriquecer a capacidade de busca para o usuário, a radicalização acaba por provocar uma redução da estrutura de indexação (BAEZA-YATES, RIBEIRO-NETO, 1999), uma vez que vários termos serão indexados por um único termo “raiz”. Os algoritmos de radicalização irão variar de acordo com a língua na qual o texto a ser processado está escrito. Em seu trabalho, Viera e Virgil (2007) apresentam uma tabela com os algoritmos disponíveis para diversas línguas, além de analisar algoritmos de radicalização disponíveis para a língua portuguesa.

Identificação de termos alternativos e relacionados

Ao realizar uma busca por determinado termo, é muito interessante que seja apresentado aos usuários um conjunto de termos que sejam alternativos e possuam o mesmo significado em um contexto de um conteúdo textual. Uma busca por “enxaqueca” poderia retornar resultados relativos à “dor de cabeça” e “cefaléia”, por exemplo.

Juntamente com os termos alternativos, termos relacionados ao item podem ser apresentados e seriam de extrema importância para a exploração da informação. Seguindo o exemplo anterior, ao buscar por “enxaqueca”, alguns termos como “paracetamol”, “clínica de tratamento” e “doenças mais reportadas” poderiam ser indiretamente disparados ao mecanismo de busca, retornando para o usuário os documentos relacionados. Sempre deixando claro nos resultados os que são decorrentes da busca direta e os que foram conseqüência da aplicação dos mecanismos de tratamento.

Correção da grafia dos termos

Ao digitar o termo em questão, o usuário pode cometer um equívoco em sua grafia. Este pode ser resultante de erro de digitação ou até mesmo de desconhecimento da grafia correta. Principalmente para esta segunda causa, o mecanismo de busca deve oferecer ao usuário opções para os termos digitados. Encontramos facilmente este tratamento nos sítios de buscas atuais. Ao tentar buscar documentos sobre “leny kravitz”, os sítios retornam algumas entradas, mas sugerem que você refaça sua busca pelo termo “lenny kravitz” (cuja grafia do nome está correta).

Diante de todos os processos apresentados, o DW 2.0™ propõe uma nova abordagem, na qual os dados não-estruturados podem e devem ser utilizados em conjunto com dados estruturados em ambientes analíticos. O primeiro passo é a leitura do texto. Uma vez lido, este texto irá para uma área de trabalho, onde serão exibidas as palavras e um índice que irá remeter à fonte destas palavras. O próximo passo caberá ao analista que poderá executar uma série de ações sobre este texto (eliminação de palavras, edição, contadores, etc.). Todo este trabalho é realizado antes do texto estar pronto para a visualização.

Com este trabalho concluído, a arquitetura do DW 2.0 propõe que um componente denominado *Self Organizing Map* (SOM) seja utilizado para processar as palavras e os índices, gerando um mapa de visualização destes (INMON, 2007b). Através deste mapa, o usuário poderá realizar operações como busca de conteúdo, busca de palavras, etc.

A arquitetura de tratamento e disponibilização dos dados não-estruturados no DW 2.0 pode ser observada, de forma geral, na Figura 1. Primeiramente os dados são lidos e, após passarem por técnicas de processamento textual, incorporados ao DW. Os dados são então lidos e apresentados ao usuário por meio de estruturas de visualização (SOMs). Maiores detalhes podem ser obtidos em (INMON, 2007c; INMON, 2008b).

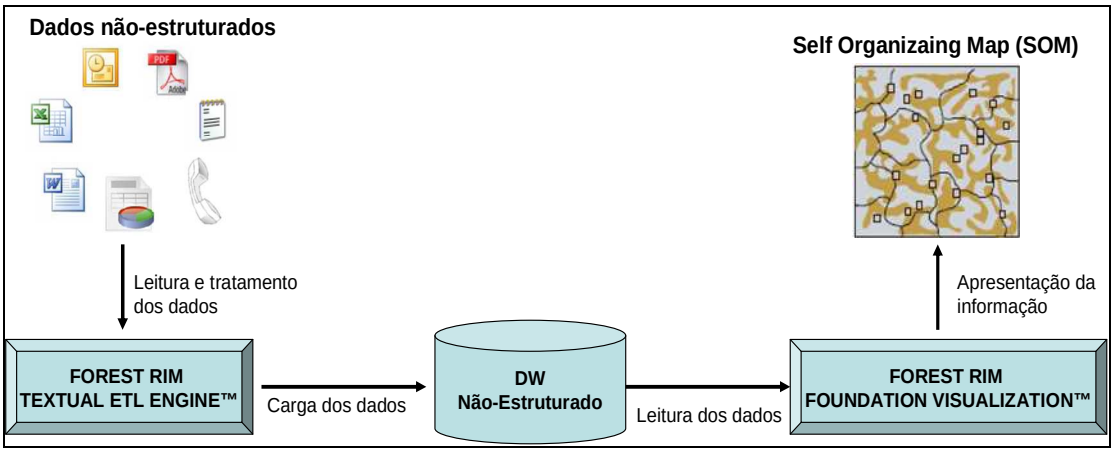

Figura 1. Arquitetura macro sobre o tratamento e disponibilização dos dados não-estruturados no DW 2.0

Analisando a arquitetura, podemos observar que esta possui a grande inovação de tratar os dados não-estruturados, trazendo-os para a tecnologia de DW. Além disto, um novo mecanismo de visualização dos dados é proposto, se apresentando como uma possível solução para problemas de visualização. Através das técnicas mencionadas, teremos como resultado um DW com informações provenientes dos documentos não-estruturados, mas como fazer para integrá-las com as informações presentes no DW estruturado, ou seja, como realmente construir a ponte entre os dois universos não está explicitamente informado na documentação disponível publicamente.

DW 2.0™ possui outras características, como uma nova abordagem no tratamento de mudanças nos dados ao longo do tempo e a construção de um dicionário de metadados corporativo. Estas, entretanto, estão fora do escopo deste trabalho. Maiores detalhes sobre as mesmas podem ser encontrados na documentação presente na seção sobre DW2.0™ disponível no sítio “Corporate Information Factory”[10].

[10] http://inmoncif.com/registration/news/dw2.php

2.2 DoctorOLAP

No trabalho realizado por Moreira, Cordeiro e Campos (2009), um ambiente de análise para informações originárias de prontuários médicos é criado, denominado DoctorOLAP. O principal objetivo do trabalho foi permitir uma análise multifacetada de maneira a atender as demandas analíticas dos médicos sobre as informações disponíveis. Os dados presentes no ambiente possuem natureza heterogênea, podendo ser originários de fontes de dados estruturados, semi-estruturados ou não-estruturados.

De acordo com Moreira, Cordeiro e Campos (2009), a maioria dos Data Warehouses construídos para o domínio de clinicas médicas têm como foco a integração de dados de natureza estruturada, à exceção da arquitetura proposta por Zhou e outros (2008). Nesta o universo não-estruturado é a principal fonte de informações, pois, de acordo com Zhou e outros (2008), este universo é a principal fonte da medicina chinesa tradicional, que constitui o assunto a ser explorado pela arquitetura proposta. Através de uma ferramenta própria, o MEDical Information Extraction (MedIE) (ZHOU et al., 2006), as informações dos prontuários são extraídas, tratadas e carregadas em um modelo dimensional que irá suportar a análise, realizada através da ferramenta OLAP BusinessObjects.

No DoctorOLAP, a fonte de informações é a base de dados de um sistema de prontuários eletrônicos, denominado eDoctor, onde todas as informações pertinentes aos pacientes e suas consultas são registradas pelo médico que está realizando o atendimento. Alguns dados, como nome e data, por exemplo, são armazenados de maneira estruturada, em tabelas relacionais. Entretanto, existe um campo de livre digitação, denominado “Texto da anamnese”, que possibilita ao médico uma descrição completa de todo o atendimento. Algumas informações podem ser inseridas de maneira semi-estruturada, como por exemplo, a pressão arterial, que será sempre precedida da sigla “PA”. Além deste, outros campos com conteúdo não-estruturado estão presentes no modelo.

Ao realizar o levantamento das necessidades analíticas dos médicos, Moreira, Cordeiro e Campos (2009) constataram que nem sempre as respostas poderiam ser obtidas

analisando-se somente os dados de origem estruturada, tendo estes que ser complementados pelos dados presentes nas anotações realizadas de maneira livre.

Para resolver este problema, os autores desenvolveram a arquitetura apresentada na Figura 2. Nela estão previstas como fontes dados de natureza estruturada e não-estruturada. Dois processos de ETC foram desenvolvidos, provendo informações para o mesmo cubo de análise, possuindo, entretanto, focos diferentes. Enquanto um se preocupa com a transformação dos dados, o outro tem como objetivo a classificação dos termos presentes nos campos de livre escrita (não-estruturados). Foi criado, para cada paciente, um documento de texto que expressa os atendimentos ao mesmo, podendo este ser acessado a qualquer momento das análises.

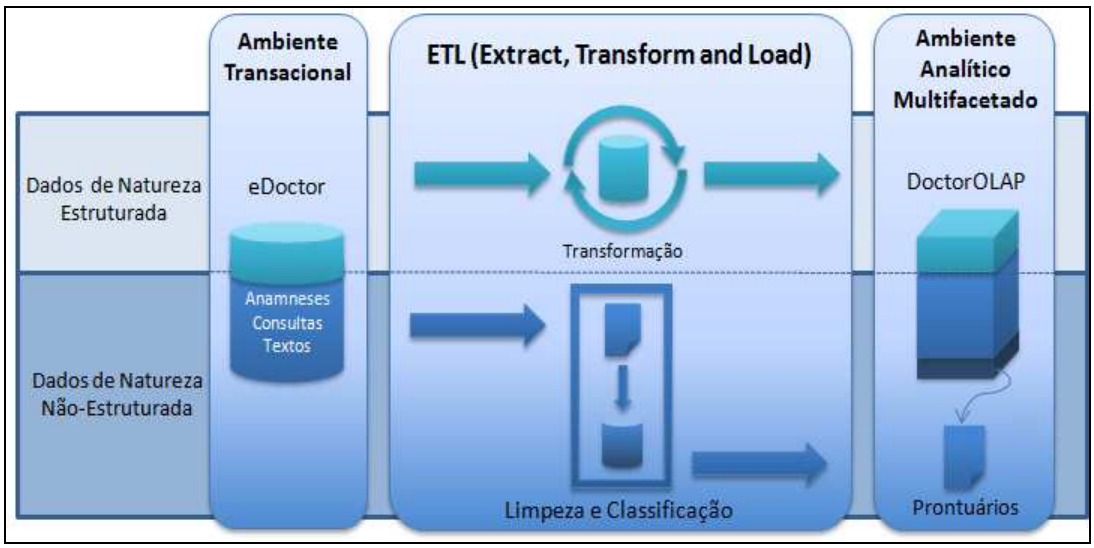

Figura 2. Arquitetura do DoctorOLAP (MOREIRA, CORDEIRO, CAMPOS, 2009)

De maneira a permitir uma análise conjunta dos dados estruturados e não-estruturados, foi desenvolvido o modelo presente na Figura 3. Podemos observar que a ligação entre os dois universos é realizada através da dimensão “Paciente”. Esta ligação ocorre da seguinte maneira: como foi gerado um documento contendo todas as informações pertinentes, para cada paciente, ao se aplicar os procedimentos de tratamento textual o sistema consegue identificar qual o paciente em questão, ou seja, já neste momento está sendo construída uma ligação entre os dois universos.

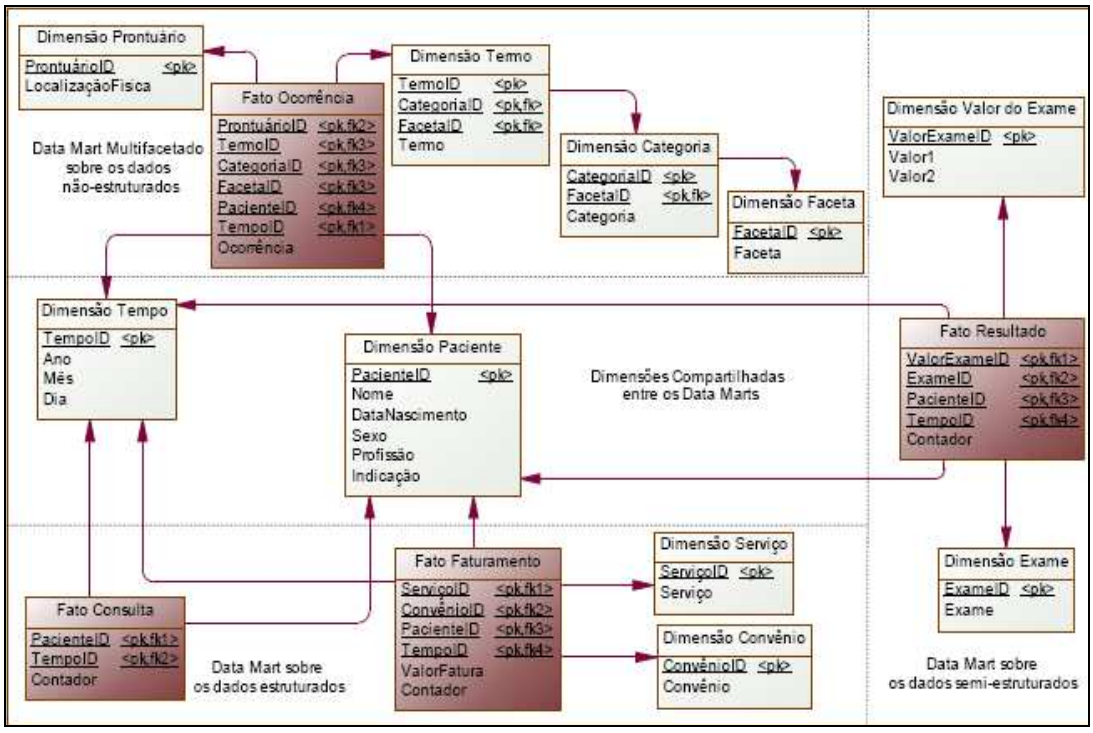

Figura 3. Modelo de dados do DoctorOLAP (MOREIRA, CORDEIRO, CAMPOS, 2009)

Podemos observar, analisando a Figura 3, que cada termo será classificado em uma categoria, que por sua vez estará presente em uma faceta. O processo de ETC Textual, construído por Moreira, Cordeiro e Campos (2009) foi realizado em três etapas: extração dos prontuários médicos, classificação dos termos nas categorias (e conseqüentemente nas facetas) e carga do modelo dimensional. A primeira etapa consiste na extração dos dados dos campos não-estruturados e geração dos documentos de prontuário. Nesta etapa foi utilizada a ferramenta *Foundation*[11], para a remoção de *stop words* e radicalização.

Com a ajuda dos médicos que utilizam o sistema eDoctor, a etapa de classificação foi realizada, tendo como resultado a taxonomia facetada que deu origem às dimensões "Termo", "Categoria" e "Faceta". A última etapa, de carregamento dos dados, foi realizada através de mecanismos especificamente para o trabalho.

Tendo os dados carregados e disponíveis, o DoctorOLAP utiliza a ferramenta OLAP Dundas[12] para a criação das consultas. Maiores detalhes sobre este processo, assim como das

[11] http://www.inmondatasystems.com/foundation.html

[12] http://www.dundas.com/

etapas anteriores, estão disponíveis em (MOREIRA, CORDEIRO, CAMPOS, 2009). Além destas informações, é possível visualizar as telas obtidas com as consultas e os resultados encontrados.

O DoctorOLAP permitiu a realização de análises conjuntas entre dados provenientes de um universo estruturado e dados provenientes de um universo não-estruturado. Entretanto, a solução foi idealizada de maneira a atender especificamente às necessidades de um universo de informações sobre dados clínicos médicos, sendo a dimensão de "Paciente" o elo entre os universos. Novos mecanismos de classificação, navegação e exploração precisam ser desenvolvidos de maneira a generalizar a solução.

2.3 Data Warehouse de documentos XML

Por sua natureza semi-estruturada, os arquivos XML tornam-se os principais meios na tentativa de se estruturar uma informação de natureza não-estruturada. Por possuir mecanismos eficientes e práticos de recuperação das informações existentes em seu corpo como, por exemplo, o XPath[13] e o XQuery[14], XML vem ganhando espaço na construção de base de dados de documentos.

Muitos estudos realizados para a construção deste trabalho (GOLFARELLI, RIZZI, VRDOLJAK, 2001; NASSIS et al, 2005; PARK, HAN, SONG, 2005; PEREZ, 2007; WIWATWATTANA et al, 2007; RUSU, RAHAYU, TANIAR, 2006) apontam a construção de um DW de documentos em XML como solução para a integração entre os dados não-estruturados com os dados estruturados. A construção deste repositório pode ter como origem documentos totalmente não-estruturados, que seriam transformados em documentos XML, ou mesmo documentos originalmente construídos em XML. Estes, cada vez mais encontrados com o visível crescimento de aplicações de comércio eletrônico e da Internet, tornam-se indispensáveis no processo decisório da empresa.

O desafio, conforme exposto por Bordawekar e Lang (2005), está em prover as operações utilizadas nos mecanismos OLAP tradicionais neste repositório de documentos

[13] http://www.w3.org/TR/xpath

[14] http://www.w3.org/TR/xquery/

XML, ou seja, tornar operações já conhecidas entre os gestores da informação, disponíveis nesta nova abordagem. Mesmo tendo disponíveis estas operações, outro desafio está em como integrar estas análises com os dados estruturados. Esta integração será estudada adiante, tendo sido descrita na solução proposta por Perez e outros (2005).

Um aspecto importante ressaltado por Bordawekar e Lang (2005) sobre a natureza das análises do DW de documentos XML consiste em perceber que as análises OLAP tradicionais envolvem dados de negócio, consistindo em valores numéricos (por exemplo, vendas de determinado produto), através de funções de agregações. Como o XML é usado também para especificação de dados que não pertencem ao negócio em si, a análise poderá necessitar de dados numéricos e não numéricos.

Bordawekar e Lang (2005) propõem em seu estudo uma extensão do XQuery, através da incorporação de operadores especiais (GROUB BY, ROLL UP, CUBE e TOPOLOGICAL), construídos pelos autores, para apoiar a natureza de análise do DW de documentos XML, partindo da seguinte premissa:

> "Qualquer documento XML que está sendo analisado é primeiramente analisado e traduzido em uma representação baseada em uma árvore lógica [...]. Esta representação lógica pode ser analisada utilizando-se as linguagens existentes para XML como, por exemplo, XPath, XSLT e XQuery." (BORDAWEKAR, LANG, 2005).

A grande vantagem desta abordagem é a utilização de operações OLAP, familiares aos usuários de sistemas analíticos, na descoberta de informações presentes nos documentos XML. Entretanto, estas técnicas foram aplicadas aproveitando-se a natureza semi-estruturada destes documentos, ficando restrita a estes, deixando de contemplar diversas informações presentes em documentos de texto e e-mails.

Tendo a corporação um DW composto por informações não-estruturadas, relevantes para a empresa, o desafio de integração passa a ser a construção de um mecanismo que possibilite acesso único aos dois ambientes: o DW de informações estruturadas, composto por dados provenientes de sistemas da empresa, ou seja, composto por dados de origem estruturada e o DW de informações não-estruturadas, cujo conteúdo está na forma de

documentos XML, denominado “document warehouse”. Perez e outros (2005) propõem um trabalho que apresenta como solução a construção de uma arquitetura que irá utilizar, para análises, informações destes dois ambientes através de um ponto de entrada único para as consultas a serem realizadas.

A arquitetura proposta unifica os dois Data Warehouses, previamente existentes, criando um Data Warehouse de Contexto (C*ontextualized Data Warehouse)*. Este “novo” DW é estabelecido de forma virtual, ou seja, não serão utilizados procedimentos de armazenamento destes dados em um terceiro ambiente, sendo os dados recuperados e disponibilizados de acordo com a demanda do usuário. Através desta arquitetura, os dados estruturados poderão ser comparados, analisados e expandidos, sendo relacionados com documentos de origem interna (de conteúdo produzido pela própria corporação) ou externa (notícias recuperadas da Web sobre concorrentes, por exemplo) da empresa. A Figura 4 (PEREZ, et. al, 2005) ilustra a arquitetura sugerida.

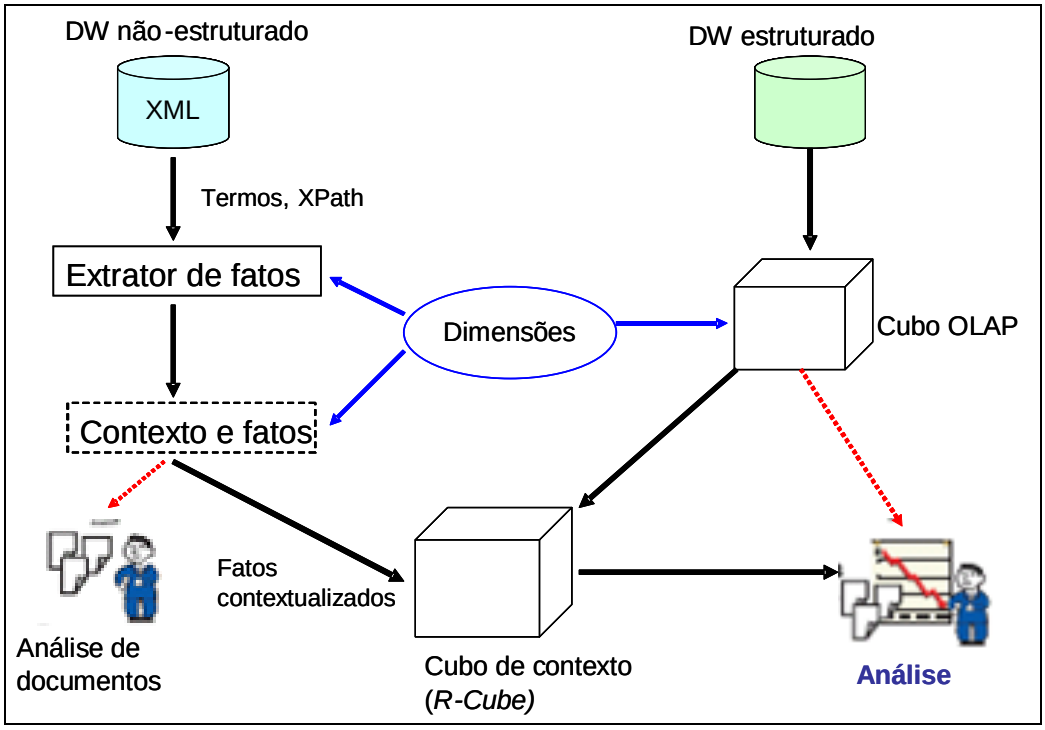

Figura 4. Arquitetura dos R-Cubes (PEREZ, et. al, 2005)

O estudo define como integração a capacidade de se visualizar, em uma mesma resposta de consulta, os dados não-estruturados e os dados estruturados. A consulta deverá então possuir características heterogêneas, englobando aspectos de:

- Consultas em ambientes OLAP estruturados;
- Consultas de RI, realizadas em documentos contidos em coleções;
- Técnicas de consulta a documentos XML, uma vez que os documentos carregados são transformados para um formato XML.

Para atender a estas especificidades, uma consulta deverá ser formada por:

- Uma seqüência de termos (como definimos em RI);
- Uma expressão *XPath* (que irá indicar onde procurar os termos nos documentos XML);
- As condições que serão avaliadas no DW estruturado, sendo estas escritas em MDX[15], de acordo com o modelo dimensional disponível, contendo os atributos, fatos e filtros relevantes para a consulta.

O fluxo de execução de uma consulta será realizado da seguinte maneira: (1) As condições descritas em MDX serão enviadas ao DW corporativo. Este então irá retornar os valores das dimensões e dos fatos que satisfizeram os parâmetros informados. Esta operação não difere das características de consultas OLAP utilizadas usualmente nos Data Warehouses atuais. (2) Paralelamente, a seqüência de termos definida será submetida ao DW não-estruturado (XML) e os documentos que contêm os termos, no local definido pelo XPath, serão retornados, juntamente com a relevância dos mesmos em relação aos documentos.

Tendo o retorno dos dois Data Warehouses envolvidos, (3) será realizada a construção de fatos de análise, que irão conter, em uma linha, os dados retornados pela consulta MDX e os documentos que estão relacionados a estes. Neste momento é realizado um cálculo de relevância para cada fato descoberto. Este retorno é chamado de Cubo de Relevância (*R-Cube*), pois irá possuir os valores (métricas) e as informações (atributos) encontrados pelo MDX, assim como a relevância (dimensão relevância do cubo) e uma lista com os documentos retornados (dimensão contexto do cubo).

Para exemplificar o retorno de uma consulta neste ambiente, podemos utilizar o seguinte cenário, conforme explicitado por (PEREZ, et. al, 2005). Uma empresa produtora de óleo vegetal possui, em seu DW estruturado, as seguintes informações:

[15] http://www.microsoft.com/msj/0899/mdx/mdx.aspx

- Fatos: Quantidade Vendida e Custo
- Dimensões: Tempo, Produto e Cliente.

Como DW não-estruturado, a empresa mantém uma base de notícias de jornais, extraídas da Internet no formato XML. Na Figura 5 temos um exemplo de um fragmento de um documento retirado da Web e disponibilizado no DW documentos.

```
<article date=''Dec.1,1998''>
<paragraph>
The financial crisis in Southeast Asian countries,
has mainly affected companies in the food market
sector.  Particularly, Chicken SPC Inc.  has reduced
total exports to $1.3 million during this half of the
year from $10.1 million in 1997.
</paragraph> ...
</article>
```

Figura 5. Exemplo de documento armazenado no DW de documentos. (PEREZ, et. al, 2005)

Ao realizar uma consulta contendo os parâmetros:

- **Termos** =“financial, crisis”
- **XPath** =“/db/business/article/paragraph”
- **MDX** = (Produtos.[comida], Clientes.Pais, Tempo.[1998].Mes, SUM(Métricas.Quantidade) > 0)

teríamos como resultado o cubo apresentado na Tabela 1. A primeira coluna é o fato de contexto, aquele que foi recuperado para a consulta. As colunas “País”, “Mês” e “Custo” contêm informações vindas do DW estruturado e foram recuperadas de acordo com o parâmetro MDX. A coluna “Relevância” contém uma medida da importância do fato de contexto em relação aos parâmetros da consulta submetida. Podemos observar na tabela que os fatos mais relevantes são as vendas feitas para o Japão e para a Coréia durante os meses de Outubro e Novembro de 1998. A coluna “Contexto” lista todos os documentos nos quais os termos foram encontrados, juntamente com a relevância dos mesmos no documento específico. O usuário pode, então, realizar operações de *drill-through* para analisar o conteúdo dos documentos, realizando assim uma expansão dos dados relacionados com a consulta.

Tabela 1. Cubo criado como retorno da consulta. (PEREZ, et. al, 2005)

Fato	Produto	País	Mês	Custo	Relevância	Contexto
f1	fo1	Cuba	1998/03	4.300.000$	0,05	D3(0,005); D7(0,005)
f2	fo2	Japão	1998/02	3.200.000$	0,1	D5(0,02)
f3	fo2	Coréia	1998/05	900.000$	0,2	D4(0,04)
f4	fo1	Japão	1998/10	300.000$	0,4	D1(0,04); D2(0,08)
f5	fo2	Coréia	1998/11	400.000$	0,25	D2(0,08); D6(0,01)

A solução proposta por Perez e outros (2005) apresenta um grande avanço na integração entre dados estruturados e não-estruturados, permitindo que o usuário realize uma consulta aos dois universos, através do acionamento de duas sub-consultas em paralelo, unificando e estabelecendo interligações entre estas na resposta. Entretanto, o universo não-estruturado está restrito a documentos de natureza semi-estruturada, mais especificamente documentos XML. A expansão desta técnica, possibilitando abranger um universo maior de dados, é um caminho que pode ser explorado em futuros trabalhos na área. Além disto, a construção de uma ferramenta para auxiliar o usuário no desenvolvimento das consultas tornaria a técnica mais acessível ao mundo gerencial. Vale ressaltar também que a solução não permite uma maior exploração do universo não-estruturado pelo usuário.

2.4 Cubo de documentos

McCabe e outros (2000) e Lee e outros (2000) apresentam uma abordagem para recuperação de informação baseada nos conceitos de OLAP é apresentada visando à utilização de todas as facilidades de manipulação de informação, presentes no mundo OLAP, nas tarefas de busca de informação em uma coleção de. Para permitir tal abordagem, um modelo dimensional deve ser construído para enquadrar as informações presentes no documento em um Data Mart de navegação. Sendo este Data Mart disponibilizado para o usuário, as buscas seriam então realizadas como consultas OLAP, de maneira que os termos, documentos e outras informações seriam visualizados como objetos OLAP, ou seja, como atributos e métricas.

Assim como em um sistema tradicional de RI, os documentos de uma coleção passariam pelos procedimentos padrão de limpeza de seu conteúdo (eliminação de *stop words*, radicalização, etc.). A diferença desta abordagem consiste em que, após passar por este tratamento, os documentos passam também por um processo de ETC (extração transformação e carga), e são então carregados no modelo dimensional desenvolvido.

O modelo dimensional deve possuir todas as informações necessárias para que o usuário consiga recuperar documentos relevantes em sua consulta. Um possível modelo que satisfaz estas necessidades é apresentado por McCABE e outros (2000) e tem como componentes:

- Tabela de fatos OCORRENCIA_TERMOS. Cada linha registra a ocorrência de um termo, contendo um peso (freqüência) e as chaves das dimensões.
- Dimensões: Tempo, Localização, Documento e Termo. Cada dimensão possui informações que serão utilizadas nas consultas.

A Figura 6 apresenta o modelo dimensional sugerido no estudo realizado por McCABE e outros (2000).

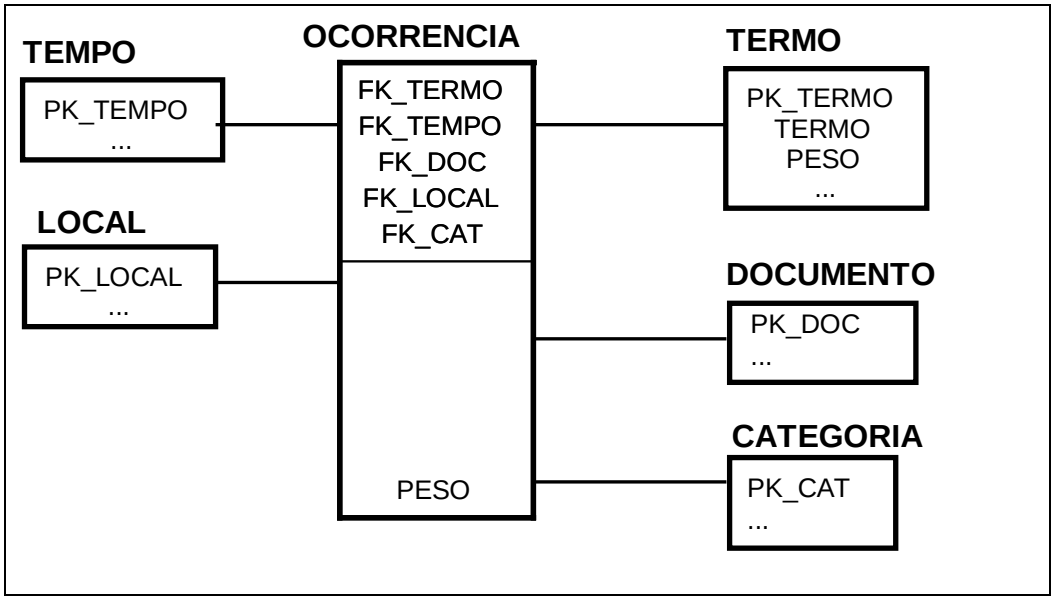

Figura 6. Modelo dimensional sugerido no estudo em (McCABE, et. al, 2000)

Nesta abordagem, a maneira como as consultas são realizadas pelos usuários diferem do comportamento padrão dos mecanismos atuais de RI. As consultas à coleção são realizadas

da mesma forma que consultas em ambientes OLAP, sendo mediadas por uma ferramenta especialista, tal como BusinessObject e MicroStrategy. Atualmente existem diversas ferramentas OLAP no mercado, cada uma com suas especificidades. O detalhamento destas não está no escopo deste trabalho.

Todas as facilidades provenientes destas ferramentas são incorporadas à tarefa de recuperação de informação. Operações como *drill-down, roll-up, drill-across e drill-thorght* podem ser realizados nos documentos da coleção.

Algumas perguntas que podem ser respondidas pela consulta OLAP:

- Encontre todos os documentos sobre florestas, publicadas em Nova Iorque, no primeiro trimestre de 1998.
- Quando foram publicados os documentos da região sul sobre escravidão?
- Quais são os cinco termos que mais aparecem em documentos sobre escravidão?

Esta técnica tem como ponto forte o aproveitamento das facilidades presentes nas ferramentas OLAP existentes no mercado, assim como está preparada para uma eventual ponte entre este cubo de documentos e os cubos de informações estruturadas. Esta ponte poderia ser realizada através de dimensões em comum que os dois universos possuam. Entretanto, técnicas de recuperação de informação estão cada vez mais maduras, possuindo mecanismos eficientes e de bom desempenho para a realização das tarefas que seriam designadas às ferramentas OLAP, ou seja, este caminho pode não ser o mais eficiente para consultas a informações não-estruturadas. Além disto, a solução não oferece um mecanismo exploratório do universo não-estruturado que seja satisfatório ao usuário.

2.5 Considerações finais

Nas soluções analisadas, os dados não-estruturados passam por um processo de ETC e são carregados em um modelo dimensional, passando a fazer parte do DW, tornando-se passíveis de exploração em conjunto com os dados vindos do universo estruturado. Ao realizar a estruturação do dado não-estruturado, estamos alterando sua natureza para poder realizar as análises necessárias, podendo esta alteração afetar a riqueza de informações

disponíveis. Além disto, o processo de estruturação irá geralmente amarrar a solução desenvolvida para um cenário específico, dificultando o caminho para uma solução mais genérica. Nestes cenários, as dimensões desempenham o papel principal na exploração de informações, sendo peças chave dos modelos. A dimensão que irá realizar a ligação entre o universo estruturado e o não-estruturado irá diferir de acordo com a solução desenvolvida para o domínio de informação que está sendo explorado.

Ao analisarmos os dados não-estruturados em sua forma original, precisamos de um mecanismo capaz de desempenhar o papel de auxílio exploratório. Iremos observar, no capítulo 3, que este papel pode ser desempenhado eficientemente por uma taxonomia facetada.

Tanto o DoctorOLAP quanto o DW 2.0™ oferecem mecanismos exploratórios mais satisfatórios, do ponto de vista do usuário, do que as soluções encontradas baseadas em Data Warehouses de documentos. Enquanto o DoctorOLAP foi construído de maneira a atender um domínio específico, o DW2.0™ pode ser aplicado a diferentes domínios, não especificando entretanto como a exploração conjunta dos universos (estruturado e não-estruturado) deve ser realizada.

3. Taxonomias facetadas

Este capítulo apresenta os principais estudos realizados, para este trabalho, sobre taxonomias facetadas, assim como algumas aplicações destas em soluções de exploração de informações. Tzitzikas e Analyti (2004) definem uma taxonomia facetada como um conjunto de taxonomias, cada uma descrevendo o domínio em questão sob um aspecto (ou faceta) diferente. O estudo teve como objetivo destacar as principais características e utilizações deste mecanismo, de maneira a torná-lo a base para a exploração conjunta desejada entre o universo estruturado e o não-estruturado.

Na seção 3.1 são apresentados alguns exemplos de utilização de taxonomias facetadas em sistemas, tanto para classificação quanto para gerenciamento de conteúdo. Um padrão de representação deste mecanismo, o XFML, é apresentado na seção 3.2. A seção 3.3 aborda o papel de taxonomias facetadas como mecanismos de auxílio à exploração. Alguns pontos críticos em relação à utilização são levantados na seção 3.4, sendo algumas considerações finais realizadas na seção 3.5.

3.1 Utilização de taxonomias facetadas em sistemas de classificação e gerenciamento de conteúdo

Podemos denominar de Sistemas de Classificação Facetada (SCFs) aqueles que utilizam taxonomias facetadas como método de descrição multidimensional e agrupamento da informação através de seus assuntos ou atributos. SCFs assumem o fato de que os usuários podem entender uma informação sobre mais de um ângulo, com seus respectivos atributos. Encapsulando estes atributos, ou dimensões, como facetas, o sistema provê aos usuários múltiplas facetas, com diversas categorias de informação, permitindo ao usuário buscas e navegações com grande flexibilidade (UDDIN, JANECEK, 2007).

Uddin e Janecek (2007) apresentam em seu trabalho vários exemplos de sítios da Web que estão utilizando este tipo de classificação para buscas e navegação por seus conteúdos. Como exemplo da vantagem que a utilização de facetas pode apresentar, imaginemos um

sistema com documentos médicos, cuja taxonomia de auxilio à busca foi construída conforme a Figura 7. Caso um usuário queira consultar documentos relativos a hospitais do Rio de Janeiro, a taxonomia irá atender perfeitamente. Agora, imagine que, independente do estado, o usuário queira encontrar documentos, relativos à Ginecologia, que foram criados em Postos de Saúde. Percebemos claramente que a organização proposta não será de grande ajuda para a consulta em questão.

Quando arrumamos a taxonomia em facetas, conforme a Figura 8, o usuário ganha a possibilidade de criar análises combinando os termos "Ginecologia", presente na faceta "Especialidade", e "Posto de Saúde", presente na faceta "Local".

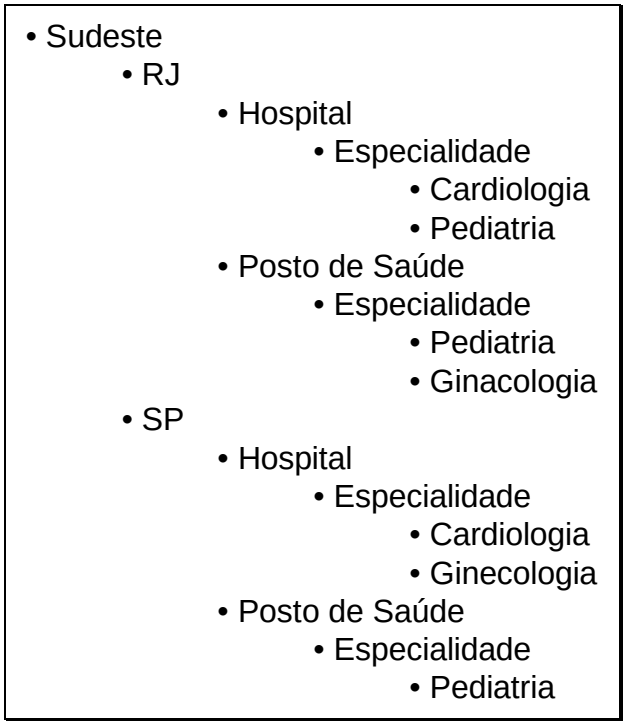

Figura 7. Taxonomia sem facetas

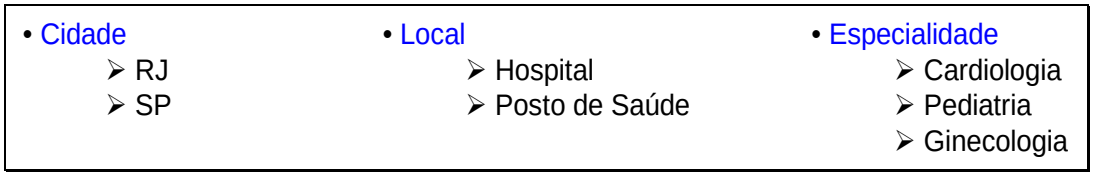

Figura 8. Taxonomia construída com facetas

Em seu trabalho, Uddin e Janecek (2007) propõem um framework para integrar SCFs com sistemas de gerenciamento de conteúdo (SGC), em inglês *Content Management Systems*

(CMS), para aumentar a acessibilidade, a organização, a visualização e a navegação de conteúdos Web. O sistema é baseado no princípio de que taxonomias facetadas permitem ao usuário uma navegação mais eficiente do que uma taxonomia simples (estrutura de árvore) pode oferecer. A idéia dos autores é a construção de um sistema, baseado no framework, que desenvolva toda a parte teórica de SCFs, além de focar em tecnologias de Web Semântica, utilizando-se de ontologias e XML para armazenar o modelo e aplicando a taxonomia facetada para prover uma estrutura dinâmica de classificação para navegação sobre o conteúdo de sítio da Web.

A Figura 9 apresenta o framework proposto. Este possui três camadas: Base de conhecimento, Sistema de Gerência de Conteúdo (SGC) e Comunicador. A primeira camada possui como função armazenar a taxonomia criada e os índices para os documentos da coleção, que podem ser internos ou externos à aplicação sendo construída. A camada de SGC é a plataforma Web utilizada para gerenciar e publicar as informações disponibilizadas. A terceira camada é responsável pela interação com o usuário, a interface, apresentando ao mesmo a taxonomia de maneira mais adequada. Podemos notar que este framework segue o bastante difundido padrão de projeto MVC (*Model-View-Controller*) para desenvolvimento de sistemas em geral.

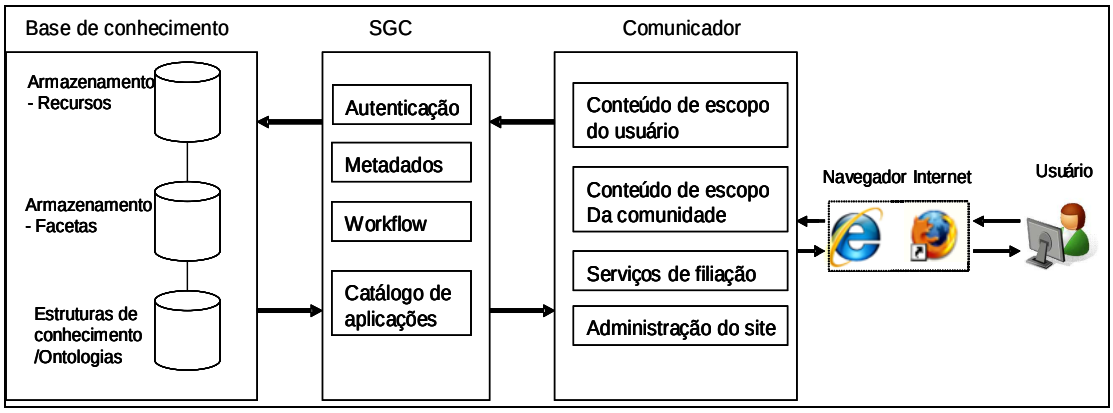

Figura 9. Framework proposto em (UDDIN, JANECEK, 2007)

3.2 Uma representação XML para taxonomias facetadas

Taxonomias facetadas podem ser facilmente representadas em arquivos XML, pois estes permitem uma organização hierárquica de estruturas e conteúdos que se adéquam ao enfoque das facetas. Entretanto, um arquivo XML oferece um grau de liberdade tão grande que, não controlado, se torna inadequado quando pensamos em intercâmbio e aproveitamento de informações por mais de um usuário, seja este humano ou um sistema.

Para solucionar este problema, podemos criar e associar ao XML um arquivo que será responsável por fornecer ao seu consumidor informações sobre seu conteúdo e como acessá-lo. Como exemplos, temos os padrões DTD [16] e o XML Schema[17], utilizados para a criação destes arquivos.

Entretanto, se cada sistema resolver criar suas próprias regras, o intercâmbio irá se tornar uma tarefa muito complicada. Torna-se necessário então o estabelecimento de um padrão, um mecanismo formal que defina como estes documentos XML serão trocados de maneira que as facetas, junto com suas estruturas hierárquicas, possam ser trocadas entre os usuários.

O XFML (DIJCK, 2009) foi desenvolvido com este propósito: possibilitar a representação e a troca de hierarquias (taxonomias) facetadas. Apesar de alguns outros padrões, como o RDF[18], por exemplo, possuírem um mecanismo de descrição mais rico, o XFML apresenta uma maior facilidade de criação e manipulação em relação a estes, tendo como vantagem o fato de ter sido construído especificamente para comportar taxonomias facetadas. Arquivos XFML podem ser processados por mecanismos padrão de *parser* XML.

A taxonomia apresentada na Figura 8 seria representada conforme a Figura 10 em um arquivo XFML:

---

[16] http://www.w3.org/TR/xhtml1/dtds.html

[17] http://www.w3.org/XML/Schema

[18] http://www.w3.org/RDF/

```
<?xml version="1.0" ?>
<xfml version="1.0" url="http://domain.com/xfml/map1.xml" language="en-us">
      <facet id="cidade">Cidade</facet>
      <facet id="local">Local</facet>
      <facet id="esp">Especialidade</facet>
      <topic id="rj" facetid="cidade">
       <name>RJ</name>
      </topic>
      <topic id="sp" facetid="cidade">
       <name>SP</name>
      </topic>
      <topic id="hosp" facetid="local">
       <name>Hospital</name>
      </topic>
      <topic id="posto" facetid="local">
       <name>Posto de Saude</name>
      </topic>
      <topic id="card" facetid="esp">
       <name>Cardiologia</name>
      </topic>
      <topic id="ped" facetid="esp">
       <name>Pediatria</name>
      </topic>
      <topic id="gineco" facetid="esp">
       <name>Ginecologia</name>
      </topic>
</xfml>
```

Figura 10.Representação da taxonomia apresentada pela Figura 8 em XFML

As facetas estão representadas pela marca "facet", enquanto as categorias pela marca "topic". Poderíamos ter vários níveis dentro de uma marca de tópico, bastando adicionar o atributo "parentTopicid". Supondo que existisse uma categoria "Cirurgia" e dentro desta existisse uma sub-categoria "Cirurgia Vascular", a marca de representação desta última seria a representada pela Figura 11.

```
      <topic id="cirur_vasc" facetid="esp" parentTopicid="cirurgia">
       <name>Cirurgia Vascular</name>
      </topic>
```

Figura 11. Exemplo de relação “pai-filho” dentro da marca "topic" em um arquivo XFML

3.3 Taxonomias Facetadas no auxílio à exploração

Ao disponibilizar um mecanismo de busca para o usuário, podemos fazê-lo de duas maneiras: direta e indireta. Na busca direta, o usuário escreve o termo ou frase desejada e o mecanismo retorna os documentos que satisfazem a estes. Na busca indireta, o espaço de

informação é mapeado, resultando na construção de uma taxonomia que melhor represente o mesmo. O usuário vai então navegar pelas categorias criadas até chegar aos documentos que satisfazem à necessidade de informação do usuário.

Enquanto o primeiro apresenta como vantagem a simplicidade, muitas vezes os resultados exibidos não satisfazem à necessidade de informação do usuário, não oferecendo recursos necessários para que o mesmo refine sua busca. Isto se deve ao fato do mecanismo não apresentar ao usuário quais critérios foram utilizados pela busca, ou seja, sob qual ou quais óticas os termos de entrada foram analisados.

Contrastando com a busca direta, as categorias utilizadas nos mecanismos indiretos dão ao usuário a noção sobre qual perspectiva este está navegando para chegar aos resultados. Entretanto, este tipo leva a uma demora maior para a obtenção dos resultados desejados, principalmente para coleções muito grandes e coleções cujo domínio não possa ser perfeitamente enquadrado em uma taxonomia.

Para combinar os benefícios das duas abordagens, mecanismos de busca apoiados por taxonomias facetadas vêm sendo alvo de estudos e aplicados em algumas situações, conforme podemos observar em (LI, BELKIN, 2008). A Busca Facetada começa com o usuário digitando os termos desejados em uma caixa de texto (assim como em uma busca direta). O mecanismo vai então realizar a busca dos termos nos documentos e, além de apresentar quais foram selecionados, disponibilizará ao usuário as categorias que agrupam os documentos retornados. Um ponto importante da Busca Facetada é que esta assume que um documento pode, e muitas vezes é, enquadrado em múltiplas e independentes facetas, ao invés de simplesmente enquadrá-lo em uma.

Outra alternativa para esta combinação de busca já está sendo apoiada por alguns mecanismos na Web, como o Open Directory[19], por exemplo. Estes disponibilizam uma navegação entre as categorias e, após o usuário selecionar a desejada, indicando sobre qual contexto sua consulta deve ser realizada, o sistema permite o envio de termos, que serão pesquisados somente junto aos documentos que estão classificados na categoria em questão. Como exemplo desta situação, podemos pensar em um usuário submetendo uma pesquisa

[19] http://www.dmoz.org/

com o termo "vírus". Utilizando um mecanismo de busca padrão, como o Google por exemplo, sem a prévia definição de contexto, simbolizada pela navegação entre as categorias, um número muito grande de documentos são retornados, sobre assuntos diversos como biologia, medicina, etc. Ao utilizar o Open Directory, o usuário pode navegar pelo caminho "Computers/Security/Malicious_Software" e pesquisar o termo sobre este contexto, ou seja, somente documentos relativos a programas maléficos ao computador seriam retornados.

Para exemplificar as situações apresentadas, podemos imaginar uma busca por "Sony LCD" em um sítio que disponibilize manuais de operação. Em um mecanismo de busca direta todos os documentos da coleção, nos quais os termos ou a frase inteira (conforme especificado na busca) estão presentes, seriam retornados. Em uma busca indireta, a taxonomia de representação do domínio seria apresentada e o usuário navegaria pelas categorias "Eletrônicos -> TVs -> LCD -> Sony" até encontrar os manuais referentes à marca Sony. Na Busca Facetada o usuário iria digitar os mesmos termos da busca direta, os mesmos documentos seriam retornados, porém desta vez associados às suas categorias.

3.4 Aspectos críticos na utilização de taxonomias facetadas

Apesar de ser um mecanismo eficiente e que acrescenta bastante ao processo de busca, a Busca Facetada ainda está sujeita a um problema inerente a todos os sistemas de classificação: o universo de informações nem sempre é visualizado da mesma maneira entre um usuário e o criador da classificação. Além disto, esta ótica pode variar de usuário a usuário. Por se tratar de um processo subjetivo, podemos até ter diferentes taxonomias criadas, sobre um mesmo domínio, caso diferentes pessoas as façam. Este problema foi apresentado por Furnas (1987), sendo denominado "Problema do Vocabulário".

Uma das iniciativas criadas para tentar resolver este problema é denominada *Dynamic Category Sets* (DCS) (TUNKELANG, 2006). Esta tem como principal foco permitir a utilização de categorias na busca sem que o usuário tenha necessariamente o mesmo modelo mental do universo da informação que o criador da taxonomia. Enquanto a busca através de uma categoria retorna uma única faceta como resultado, o DCS irá retornar um conjunto de facetas possíveis para a consulta realizada. A Figura 12 (TUNKELANG, 2006) ilustra o

resultado de uma busca por "viewsonic lcd". Os documentos encontrados são retornados de forma agrupada, através das categorias encontradas para os termos.

Figura 12. Resultado de uma busca por "viewsonic lcd" (TUNKELANG, 2006)

A solução possui como característica a orientação aos dados que estão sendo requisitados. Esta irá se basear no fato de que, apesar de existirem inúmeras possibilidades de combinações de facetas que satisfaçam a uma consulta, somente serão apresentadas aquelas cuja combinação faça sentido. O exemplo dado por Tunkelang (2006) ajuda a melhorar o entendimento deste princípio. Caso o usuário faça uma busca com os termos "Século 20 Bach", apesar de fazer parte de uma das respostas possíveis, o mecanismo não irá retornar como resultado as facetas de "período" e "compositor", pois o compositor alemão Johann Sebastian Bach viveu somente até 1750, sendo retornados somente as facetas de "período" e "autor",esta última em referência ao escritor Richard Bach.

Também são utilizados na solução recursos de processamento de texto na consulta entrada pelo usuário. Alguns mecanismos de recuperação da informação como eliminação de *stop words* e radicalização (BAEZA-YATES, RIBEIRO-NETO, 1999) são aplicados visando maior eficiência e uma maior cobertura dos resultados. Os conjuntos de resultados apresentados ao usuário são os de menor número de facetas possíveis que tenham significado para o mesmo.

Outro problema encontrado em taxonomias facetadas, quando utilizadas como mecanismos auxiliares em processos de busca, reside no fato de que nem todas as combinações dos termos das facetas criadas vão possuir documentos que possam ser indexados por esta combinação. Podemos imaginar uma taxonomia facetada criada para

auxiliar um sistema de busca por receituários pediátricos. A mesma seria composta por três facetas: Sexo, Faixa Etária e Curado. Cada faceta será organizada de acordo com a Figura 13.

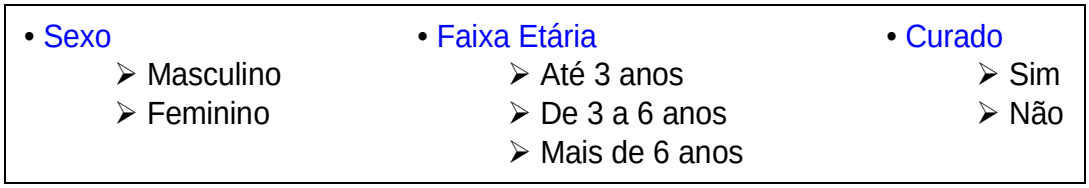

Figura 13. Taxonomia facetada construída para auxílio na busca por receituários pediátricos

Cada termo da taxonomia pode ser combinado, gerando 35 possibilidades. Supondo que o hospital em questão está beirando a perfeição e não exista nenhuma criança do sexo masculino, com mais de 6 anos, que não esteja curada, o usuário poderá ser levado por um caminho de navegação que irá resultar em uma busca vazia, uma vez que o mecanismo de navegação irá permitir que ele faça a seleção: Sexo=M, Faixa=6+, Curado=Não. No exemplo utilizado isso não seria problema, uma vez que a taxonomia criada é muito pequena e o usuário poderia voltar a procurar por outras categorias. Agora, imaginemos uma taxonomia com 10 facetas, cada uma com 10 termos. Seriam criados 10^10 combinações possíveis, muitas destas sem nenhum documento cujo conteúdo possa ser indexado pela combinação.

Percebemos então dois problemas quando utilizamos facetas nas taxonomias: o número de índices possíveis para os documentos e, conseqüentemente, o número de combinações geradas sem correspondência na base de documentos em questão. No trabalho apresentado por Tzitzikas e Analyti (2007), podemos observar a preocupação com este problema sendo ressaltado com o seguinte trecho:

> "A existência de combinações de categorias que não possuem significância para o domínio e o esforço preciso para especificar os conjuntos válidos são um problema prático, identificados desde Ranganathan há aproximadamente 80 anos atrás".

A Figura 14 (TZITZIKAS, ANALYTI, 2007) apresenta uma taxonomia facetada que serve como base para navegação do usuário em um sítio de hotéis, indexando "Hotel 1" e "Hotel 2" nas categorias disponibilizadas. Muitas combinações de categorias desta taxonomia não possuem relevância, pois não podem ser aplicadas a nenhum objeto do domínio. Como

exemplo, nunca teremos um objeto categorizado ao mesmo tempo em uma ilha grega (no exemplo “Creta” e “Kefalonia”) e em um esporte de inverno devido às características climáticas da região. Nunca teríamos também esportes marítimos em “Olimpo”, pois se trata de uma região montanhosa.

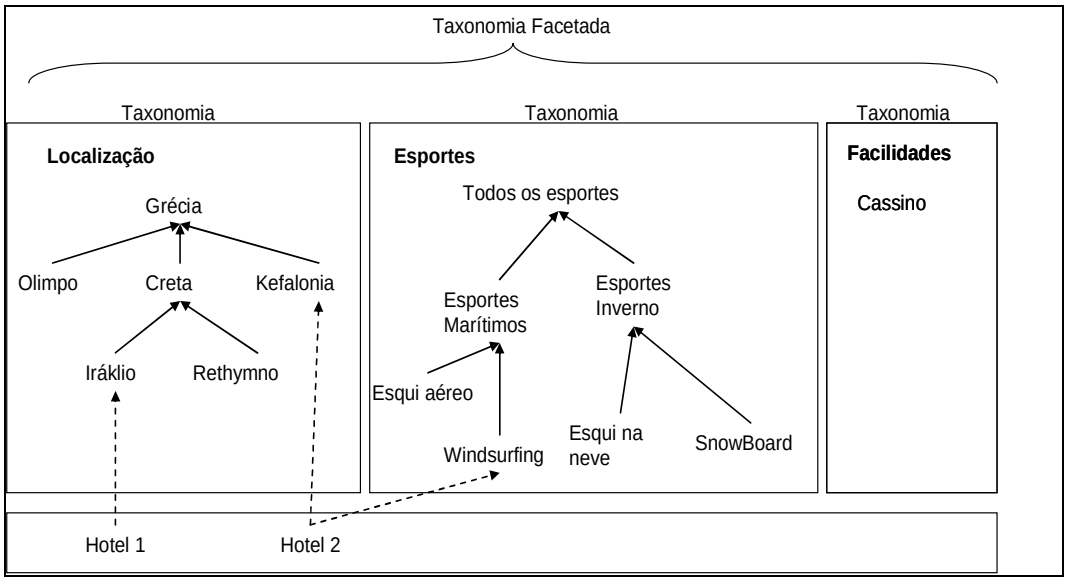

Figura 14. Taxonomia criada para sítios de busca de hotéis (TZITZIKAS, ANALYTI, 2007)

De maneira a prevenir este problema, além de enriquecer o mecanismo de taxonomias facetadas, Tzitzikas e Analyti (2007) apresentam uma solução baseada em operações algébricas, visando a definição de conjuntos de termos válidos em cima de uma taxonomia facetada, seguindo diretrizes de flexibilidade e eficiência. O algoritmo foi concebido visando a obtenção de expressões algébricas que irão determinar estes conjuntos de acordo com a taxonomia criada e o universo de objetos a ser classificado. A obtenção das combinações válidas pode ser realizada de duas maneiras: definindo-se as expressões algébricas que irão retornar os conjuntos, ou executando o algoritmo sobre uma taxonomia pré-definida, com este retornando os conjuntos. Atualizações na taxonomia, como inserções, alterações e exclusões, seja em categorias ou até mesmo na estrutura da mesma, estão previstas e são tratadas pelo mecanismo. Na taxonomia representada na Figura 14, teríamos, por exemplo, os seguintes conjuntos retornados pelo algoritmo:

Válidos: [{Kefalonia, Esqui aéreo, Cassino}, {Kefalonia, Windsurfing, Cassino}]
Não-válidos: [{Creta, Esportes inverno}, {Kefalonia, Esportes inverno}]

A utilização destes mecanismos evita possíveis erros de indexação assim como oferece uma melhor navegação para o usuário. Podemos observar que após a descoberta das combinações válidas para o domínio, pode-se construir um mecanismo de navegação inteligente, que levará o usuário a caminhos que contenham representação no domínio, ou seja, somente as combinações válidas seriam apresentadas durante a navegação. Ao escolher uma categoria de uma faceta, as demais categorias das facetas subseqüentes, que não fizessem parte dos conjuntos válidos, seriam desabilitadas para a navegação, tornando a busca pela informação muito mais eficiente e agradável para o usuário.

3.5 Considerações finais

Observando as definições e utilizações apresentadas, podemos observar que taxonomias facetadas são um poderoso mecanismo para a exploração de informações, principalmente as presentes em um universo não-estruturado. Através da construção de uma taxonomia específica para o domínio em questão, com a correta definição de suas facetas e hierarquias, o usuário tem o acesso às informações disponíveis facilitado, melhorando o desempenho do processo exploratório.

Vimos, no capítulo 2, que as dimensões desempenham este papel exploratório em um Data Warehouse. Como o objetivo deste trabalho consiste na construção de uma abordagem que permita a exploração conjunta de dados estruturados e não-estruturados, um caminho natural de solução se desenha na exploração das semelhanças existentes entre as dimensões de um DW e uma taxonomia facetada construída para o domínio. A abordagem proposta, apresentada nos capítulos seguintes, terá como base a exploração destas semelhanças.

4. Exploração conjunta de dados estruturados e não-estruturados

Os ambientes OLAP tradicionais proporcionam ao usuário uma navegação sobre diferentes perspectivas de um fato. Este papel é desempenhado pelas dimensões de um esquema dimensional, conforme observamos nos capítulos anteriores. Cada dimensão representa uma ou mais características da informação disponibilizada, como por exemplo, a data, o local, o cliente. Além de possuir diversos atributos que as qualificam, podemos ter uma ou mais hierarquias associadas a cada dimensão, servindo como mecanismo classificatório da mesma. É através desta que a análise pode ser detalhada ou expandida, conforme a direção escolhida na navegação do usuário, ou seja, as dimensões de um modelo dimensional representam um papel fundamental nas análises, sendo diretamente responsáveis pela exploração do domínio da informação.

Quando estamos falando sobre buscas em coleções de documentos, vimos que um dos mecanismos mais apropriados para a exploração desde universo não-estruturado é a construção de uma taxonomia facetada que represente o domínio da informação e consiga indexar os documentos presentes na coleção. Através da taxonomia, o usuário vai navegar nas diferentes classificações da informação, tendo este a capacidade de detalhar ou generalizar o grau de análise. Além disto, através das facetas, diferentes perspectivas podem ser agregadas no processo de exploração, permitindo ao mesmo um enriquecimento semântico da busca sendo realizada.

Em nossa abordagem, iremos utilizar uma taxonomia facetada como mecanismo de apoio à exploração integrada dos universos estruturado e não-estruturado. Esta é construída especificamente para o domínio da informação que está sendo estudado, estando diretamente associada às dimensões do modelo dimensional do DW, que servirão de ponto de partida para sua construção. Nossos esforços foram todos baseados na idéia de que esta exploração integrada pode ser fortemente apoiada nesta associação (taxonomia facetada e dimensões).

Quando nos referimos a uma exploração integrada, estamos definindo a capacidade de se analisar os dados dos dois universos em um mesmo contexto, ou seja, dada uma consulta

tradicional em um sistema OLAP, o usuário deve ser capaz de chegar aos documentos de uma coleção não-estruturada que dizem respeito a esta análise. Da mesma maneira, ao realizar uma busca por termos em uma coleção, o usuário deve ser capaz de, a partir dos documentos encontrados, analisar os fatos correspondentes a estes, que estão presentes no DW. O principal foco da abordagem é o enriquecimento das análises realizadas em um ambiente de suporte à decisão, complementando-o com informações de um universo não-estruturado, fornecendo um enriquecimento das consultas e análises, possibilitando uma melhor compreensão dos fatos explorados**.**

Na seção 4.1 apresentamos como obter a taxonomia facetada a partir de uma análise do modelo dimensional. Entretanto, ao restringir a formação da taxonomia a análises por sobre o modelo dimensional, podemos estar ignorando algumas informações que, por determinado motivo, não foram enquadradas na modelagem. Ao deixar estas informações de fora da taxonomia, estaríamos empobrecendo o poder analítico. A seção 4.2 visa a explicar como podemos enriquecer a taxonomia facetada através da utilização de recursos específicos do domínio da informação. Na seção 4.3 veremos como funciona o caminho de “mão-dupla” na exploração das informações, tanto para análises que comecem no universo não-estruturado de informações (seção 4.3.1), quanto para as análises que comecem no universo estruturado (seção 4.3.2).

4.1 Construindo uma taxonomia facetada a partir do modelo dimensional

Para permitir a exploração das semelhanças existentes entre uma taxonomia facetada e o conjunto de dimensões de um modelo dimensional, construindo uma ponte entre os universos não-estruturado e estruturado, o primeiro passo é a criação da taxonomia facetada na qual os usuários irão navegar para encontrar os documentos presentes no universo não-estruturado. Além de servir como mecanismo exploratório, a taxonomia terá um papel fundamental no processo de indexação dos documentos da coleção.

Conforme já colocado, a solução tem como um de seus objetivos principais a exploração da semelhança, tanto em relação à composição estrutural, quanto em relação à função desempenhada em cada universo (estruturado e não-estruturado) entre taxonomias

facetadas e dimensões de um modelo dimensional. Para construir a taxonomia facetada temos então que estabelecer um processo de obtenção da mesma a partir do modelo dimensional, um mapeamento que torne esta ligação possível. Este procedimento poderá construir a taxonomia por completo ou servir de base, de ponto de partida, para a construção de uma taxonomia mais completa. Um processo inverso, de obtenção do modelo dimensional a partir da uma taxonomia facetada também seria possível, mas não será abordado neste trabalho por não pertencer totalmente ao escopo planejado para a solução.

4.1.1 Conceitos e seus relacionamentos

Primeiramente devemos estabelecer os conceitos de facetas, categorias e atributos que, junto com os termos, irão compor a taxonomia facetada que servirá de apoio para a exploração proposta neste trabalho. Uma vez que estamos dividindo a taxonomia em diversas perspectivas macro, cada uma destas será uma faceta. Cada faceta irá então possuir uma ou mais categorias, podendo ser encaradas como as características da faceta. Estas características possuem qualificações, atributos que as descrevem. Associados a estes atributos estarão então os termos, as “folhas da árvore”. É através deste último nível que o usuário irá disparar o processo de busca. Entretanto, os atributos são agentes fundamentais no mecanismo de indexação dos documentos da coleção não-estruturada, pois é através destes que a ligação com as dimensões será realizada. Mesmo que uma categoria possua somente uma possível qualificação, esta terá obrigatoriamente como filho um atributo, pois este será o responsável por associá-la a uma dimensão do modelo dimensional.

Atributos de uma mesma dimensão podem possuir relações de “pai e filho” entre eles, ou seja, podem existir hierarquias nas quais o elemento “pai” possui a função de agregar informações de seus respectivos filhos. Operações muito comuns em ambientes analíticos, o detalhamento (*drill-down*) e a agregação (*roll-up*) podem ser obtidos através da navegação sobre uma taxonomia que consiga representar estas relações. A Figura 15 ilustra as relações entre facetas, categorias e atributos idealizadas para este trabalho.

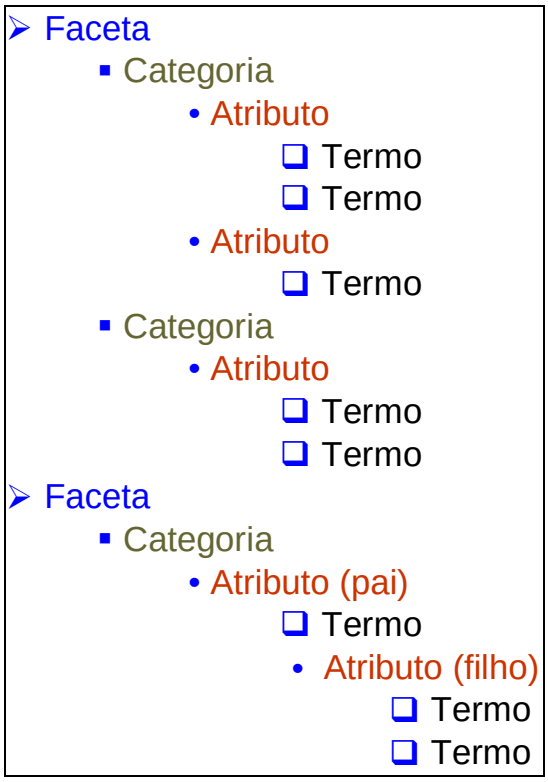

Figura 15. Conceitos utilizados para taxonomia facetada

Tendo definido os elementos básicos que irão compor a taxonomia facetada, que será utilizada neste trabalho, devemos olhar para o modelo dimensional para extrair os padrões, as semelhanças deste de maneira a criar um mapeamento entre os conceitos descritos e a estrutura das dimensões.

Uma dimensão representa uma diferente perspectiva que um determinado fato pode possuir. Uma tabela dimensional irá encapsular informações sobre esta perspectiva de maneira a fornecer o contexto necessário no momento da execução das consultas. Informações sobre o local onde o fato ocorreu seriam então armazenadas em uma dimensão denominada “Localidade”. Esta irá possuir, em suas colunas, a descrição completa do local, podendo apresentar diversos níveis do mesmo, de maneira a fornecer ao usuário possibilidades de detalhamento ou expansão de determinada ocorrência, ou seja, fornecer ao usuário a capacidade de caracterizar o fato buscado em determinada ótica. Esta caracterização não se aplica somente a hierarquias de informações, mas também a agrupamentos que estas possam eventualmente possuir. Uma dimensão típica de localidade poderia ser modelada conforma a Figura 16, que irá conter informações sobre a cidade, o estado e o país da ocorrência.

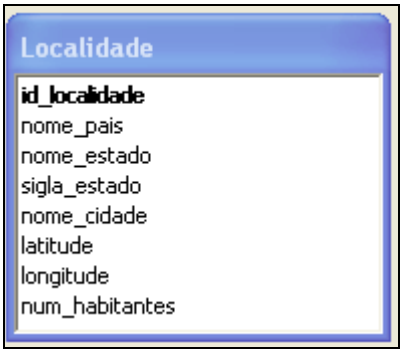

Figura 16. Dimensão Localidade

Observamos então que a dimensão localidade possui algumas características (país, estado e cidade), que são compostas por informações, atributos, que as descrevem. Notamos aqui claramente a ligação entre conceitos dos dois mecanismos (dimensões e taxonomias): uma dimensão possui categorias, que por sua vez possui atributos que as compõem. Podemos observar também que existe uma relação hierárquica entre estes atributos. Os termos são, naturalmente, o conteúdo destes atributos. Se estivéssemos falando somente de taxonomias, poderíamos parar por aqui. Entretanto, escolhemos como mecanismo para a solução proposta uma taxonomia facetada, cujos benefícios já foram explorados e apresentados ao longo deste trabalho. Fica então faltando o mapeamento entre estas dimensões e as facetas.

Uma faceta, assim como uma dimensão, representa uma perspectiva do domínio da informação que está sendo analisada. Entretanto, podemos considerar a faceta como sendo de um nível mais abrangente que a dimensão, ou seja, uma faceta poderá abrigar mais de uma dimensão que ilustre a mesma ótica desta perspectiva macro. Em um modelo que atenda a um hospital, a faceta “Quem” poderá agrupar as dimensões “Paciente” e “Médico” para ilustrar o atendimento a um paciente por um médico.

Temos então como modelo conceitual o apresentado na Figura 17, no qual vemos que uma dimensão pode aparecer em uma ou mais facetas. Apesar de possuir uma natureza incomum em uma primeira vista, esta situação poderá ocorrer em alguns modelos multidimensionais específicos. Podemos observar um exemplo desta quando nos deparamos com dimensões que se encaixam no conceito de “dimensões sucatas”. Nestas, atributos que não possuem correlação são armazenados em uma mesma dimensão. Veremos com mais

detalhes na seção 4.1.4 o conceito e uma proposta de tratamento para este caso, assim como alguns outros específicos.

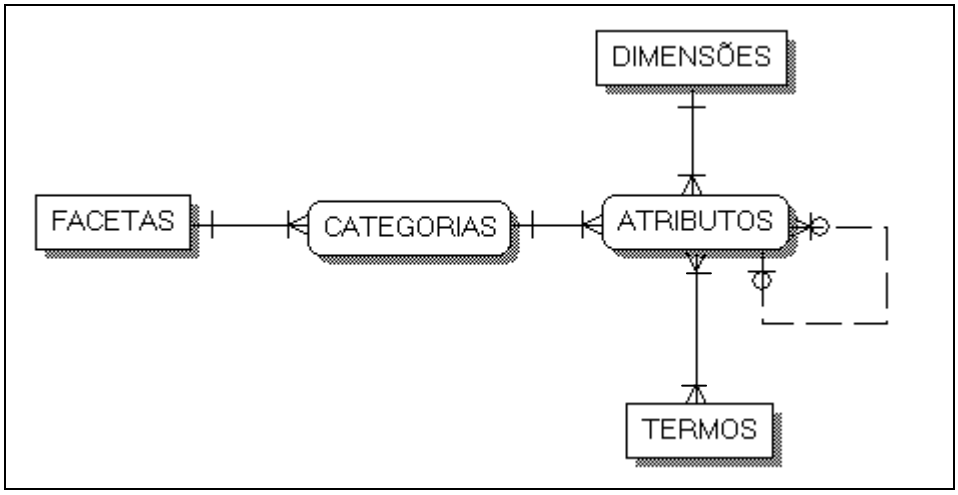

Figura 17. Modelo conceitual para mapeamento

4.1.2 Etapas para obtenção da taxonomia facetada

A criação da taxonomia facetada começa com a definição de quais facetas farão parte desta. Esta definição é baseada no assunto que está sendo tratado no modelo dimensional e nas dimensões que o compõem. Nem todas as dimensões presentes no modelo serão úteis para o processo de análise conjunta dos dois universos. Devemos então selecionar quais dimensões devem fazer parte deste processo. Tendo selecionado estas, devemos enquadrar cada uma em sua respectiva faceta. Após este enquadramento, devemos estabelecer as categorias de cada faceta. Para tal, devemos analisar todas as dimensões selecionadas, encontrando as diferentes categorias em cada uma. Definidas as categorias (características), o último passo consiste na definição de quais informações serão utilizadas sobre estas, ou seja, quais serão os atributos que serão trabalhados. Nesta etapa são identificadas eventuais hierarquias existentes entre os atributos escolhidos. A Figura 18 ilustra o fluxo proposto para a obtenção da taxonomia facetada a partir do modelo dimensional.

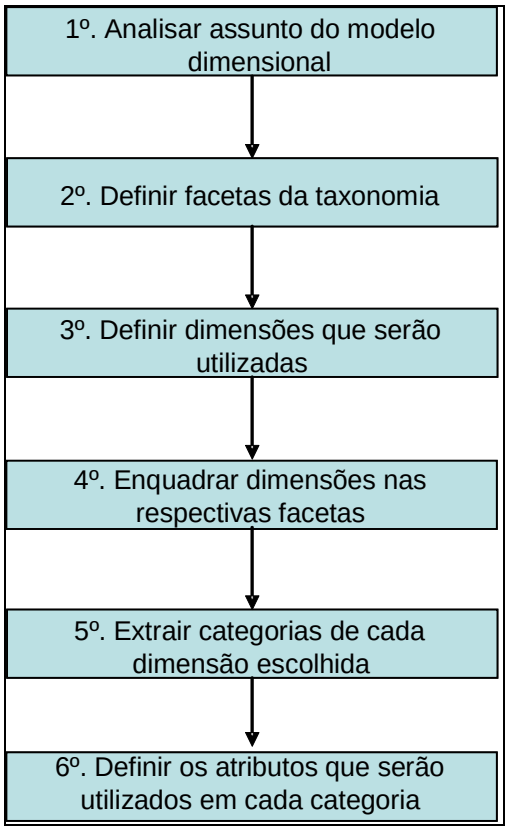

Figura 18. Etapas para obtenção da taxonomia facetada

De acordo com o modelo que está sendo estudado, uma diferente taxonomia facetada será criada para apoiar a exploração das informações. Após a aplicação destes passos, a taxonomia obtida pode satisfazer totalmente a exploração ou poderá necessitar de aprimoramentos. Estudaremos esta situação em maiores detalhes na seção 4.2.

4.1.3 Formalização da taxonomia facetada

Para construir a taxonomia, devemos estabelecer um mecanismo que expresse, principalmente, a relação entre os integrantes da mesma e os objetos presentes no banco de dados do modelo dimensional. Vimos, na seção 3.2, que podemos expressar estas relações através da construção de um arquivo XML, que irá conter as informações necessárias para a implementação da solução. O padrão XFML se mostrou muito útil no caso de intercâmbio de taxonomias entre sistemas. Entretanto, não foi adotado pela abordagem proposta neste trabalho. Apesar de não possuir grande complexidade em sua estrutura, quando comparamos com os demais padrões, optamos por uma estrutura mais simples e de mais fácil

entendimento, uma vez que o intercâmbio da taxonomia facetada, em um primeiro momento, não será necessário para a solução desenvolvida.

Podemos visualizar um exemplo de arquivo, criado de acordo com os padrões estabelecidos, na Figura 19. O documento possui como raiz a marca “dimensoes”, que irá abrigar todas as dimensões selecionadas no passo três do processo exposto na Figura 18. Cada dimensão será representada pela marca “dimensao”, que irá possuir um nome (descrito na marca “nomedim”), uma tabela associada (descrita na marca “tabela”) e um conjunto de categorias que a caracteriza. Cada marca de “dimensão” possui ainda uma marca denominada “origem”. Conforme veremos na seção 4.2, será possível utilizar fontes de informação externas ao modelo dimensional para enriquecer a taxonomia criada. A marca “origem” expressa se a dimensão é interna ao DW (“I”) ou foi obtida de uma fonte externa ao mesmo (“E”). O conjunto de categorias de uma dimensão tem como “raiz” a marca “categorias”, sendo cada uma expressa por um nome (marca “nomecat”) e uma lista de atributos (marca “atributos”). Esta última contém todas as informações, selecionadas no passo seis do processo apresentado na Figura 18, relacionadas à parte mais técnica do mapeamento. Em cada marca “atributo”, é indicado o nome (marca “nomeatrib”), a coluna sobre a qual deve ser realizado o mapeamento no banco de dados (marca “coluna”) e o tipo (marca “tipo”). O tipo do atributo é muito importante para o mecanismo que realiza a conexão do mundo não-estruturado para o estruturado, uma vez que irá influenciar na maneira como a consulta a este universo será gerada. Para este trabalho foram considerados dois tipos: inteiro e texto. Os valores 0 e 1 são atribuídos, respectivamente, para cada caso. Cada marca atributo possui ainda um identificador (“id”) e, caso exista uma relação hierárquica entre atributos, o identificador de seu pai (“pai”).

```
<dimensoes>
   <dimensao>
      <nomedim>paciente</nomedim>
      <origem>I</origem>
      <tabela>dimPaciente</tabela>
      <categoria>
         <nomecat>Paciente</nomecat>
         <atributo id='nome'>
            <nomeatrib>Nome</nomeatrib>
            <coluna>nome</coluna>
            <tipo>1</tipo>
         </atributo>
         <atributo id='profissao'>
            <nomeatrib>Profissao</nomeatrib>
            <coluna>profissao</coluna>
            <tipo>1</tipo>
         </atributo>
      </categoria>
   </dimensao>
</dimensoes>
```

Figura 19. Exemplo de mapeamento no arquivo de configuração para dimensões

Logicamente, informações sobre a estrutura da taxonomia em si são de vital importância. Para representar tal estrutura, criamos dois arquivos em XML. O primeiro, conforme exemplo visualizado na Figura 20, irá conter a relação das facetas utilizadas para o domínio em questão. Um segundo arquivo servirá para estabelecer quais categorias farão parte de cada faceta. Devemos observar que as categorias presentes neste, conforme exemplo na Figura 21, são as mesmas apresentadas pelo arquivo de exemplo apresentado na Figura 19. Poderíamos agrupar as informações dos dois arquivos em um único documento. Foi adotada, na arquitetura exposta pelo capítulo 5, a separação das informações em dois arquivos unicamente para facilitar o processamento da lógica desenvolvida. Futuros estudos podem ser realizados com esta junção.

```
<facetas>
   <faceta>
      <nome>Quando</nome>
   </faceta>
   <faceta>
      <nome>Quem</nome>
   </faceta>
</facetas>
```

Figura 20. Exemplo de criação de facetas no arquivo de configuração de facetas

```
<categorias>
   <categoria>
      <nome>Paciente</nome>
      <faceta>Quem</faceta>
   </categoria>
   <categoria>
      <nome>Medico</nome>
      <faceta>Quem</faceta>
   </categoria>
   <categoria>
      <nome>Ano</nome>
      <faceta>Quando</faceta>
   </categoria>
</categorias>
```

Figura 21. Exemplo de criação de categorias no arquivo de configuração de categorias

Adotando o procedimento descrito na Figura 18 para a dimensão apresentada na Figura 16, teríamos como resultado o mapeamento apresentado pelo Anexo 1.

Analisando o Anexo 1, verificamos, na dimensão, que existem informações adicionais sobre uma cidade que não foram mapeadas para a construção da taxonomia. De acordo com o último passo definido no procedimento da Figura 18, o usuário responsável pelo mapeamento pode julgar que as informações de latitude e longitude não são necessárias para o enriquecimento das análises. A seleção de informações que realmente irão agregar no processo decisório é de fundamental importância para o sucesso da solução. Todas as informações escolhidas serão mapeadas em atributos e irão compor a taxonomia de navegação, assim como a estrutura de índices. Informações excessivas podem tornar a navegação na taxonomia uma tarefa desagradável, assim como onerar o processo de indexação de documentos desnecessariamente, uma vez que estes não serão potencialmente utilizados pelos usuários.

4.1.4 Aplicando o mapeamento para diferentes modelos dimensionais

Algumas questões de modelagem (KIMBALL, ROSS, 2002) podem confundir um pouco o modelo de mapeamento sugerido. Embora existam formas de tratamento para cada questão, nosso foco será sugerir uma maneira de realizar o mapeamento destas para a taxonomia facetada. O mapeamento proposto seria então utilizado da maneira apresentada nas subseções a seguir, de acordo com o problema encontrado.

4.1.4.1 Dimensões "Floco de Neve" **(*snowflake*)**

Uma modelagem dimensional pode apresentar algumas variações em relação à construção das dimensões e suas ligações com as tabelas de fatos. A diferença clássica está nos modelos estrela versus os modelos floco de neve (KIMBALL, ROSS, 2002). Se, no primeiro caso, as dimensões não são normalizadas, no segundo existe uma preocupação em normalizar o conteúdo dimensional. Vale ressaltar que, por ainda se tratar de um modelo voltado para o mundo analítico, esta normalização é geralmente encontrada de maneira mais flexível do que em esquemas voltados a sistemas transacionais.

Conforme mencionado, um modelo "floco de neve" irá apresentar suas dimensões normalizadas, possuindo uma tabela para cada conjunto de informações presente na mesma. A tabela com menor nível de detalhe irá possuir a chave primária para a ligação à tabela de fatos (existem variações de modelagem que não seguem essa regra, como em casos de modelos de agregação, por exemplo). A dimensão de localidade da Figura 16 seria desmembrada em tabelas de "país", "estado" e "cidade", conforme a Figura 22.

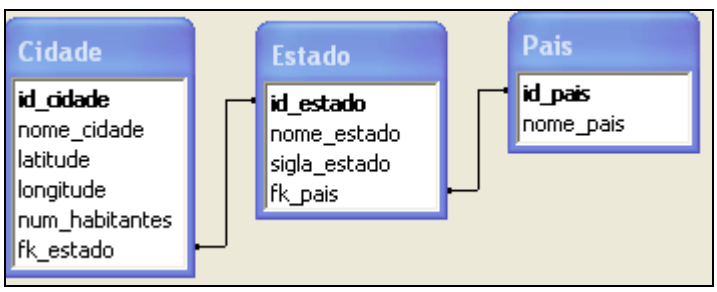

Figura 22. Dimensão Localidade: Floco de Neve

O mapeamento a ser realizado deve levar em consideração esta separação da informação em mais de uma tabela, sendo para isto necessária a criação de uma marca “dimensão” para cada nível hierárquico apresentado em uma tabela distinta. Esta limitação influi na representação da hierarquia, que na situação atual não estaria representada. Seria possível, através de uma mudança no modelo proposto, incluir todas as tabelas em uma mesma dimensão, ampliando assim este conceito. Esta modificação, entretanto, não foi implementada neste trabalho, ficando como sugestão de trabalho futuro com o objetivo de comparar os mapeamentos derivados e a taxonomia gerada. Na Figura 23 observamos um fragmento de como ficaria o mapeamento para a estrutura apresentada na Figura 22. O mapeamento completo pode ser visualizado no Anexo 2.

```
...
<dimensao>
   <nomedim>Localidade-Pais</nomedim>
   <origem>I</origem>
   <tabela>Pais</tabela>
   <categoria>
      <nomecat>Pais</nomecat>
      <atributo id='pais'>
         <nomeatrib>Pais</nomeatrib>
         <coluna>nome_pais</coluna>
         <tipo>1</tipo>
      </atributo>
   </categoria>
</dimensao>
<dimensao>
   <nomedim>Localidade-Estado</nomedim>
   <origem>I</origem>
   <tabela>Estado</tabela>
   <categoria>
      <nomecat>Estado</nomecat>
      <atributo id=’uf’>
         <nomeatrib>Nome</nomeatrib>
         <coluna>nome_estado</coluna>
         <tipo>1</tipo>
      </atributo>
...
```

Figura 23. Mapeamento para dimensões "Floco de Neve"

Para melhor exemplificar as demais questões a serem abordadas, tomaremos por base o modelo da Figura 24. Neste, representamos um simples esquema para análises de vendas realizadas pela Internet. O modelo foi criado somente para ilustrar as situações a serem descritas, não podendo servir como padrão de modelagem para este tipo de necessidade. Tais questões serão discutidas nas próximas subseções.

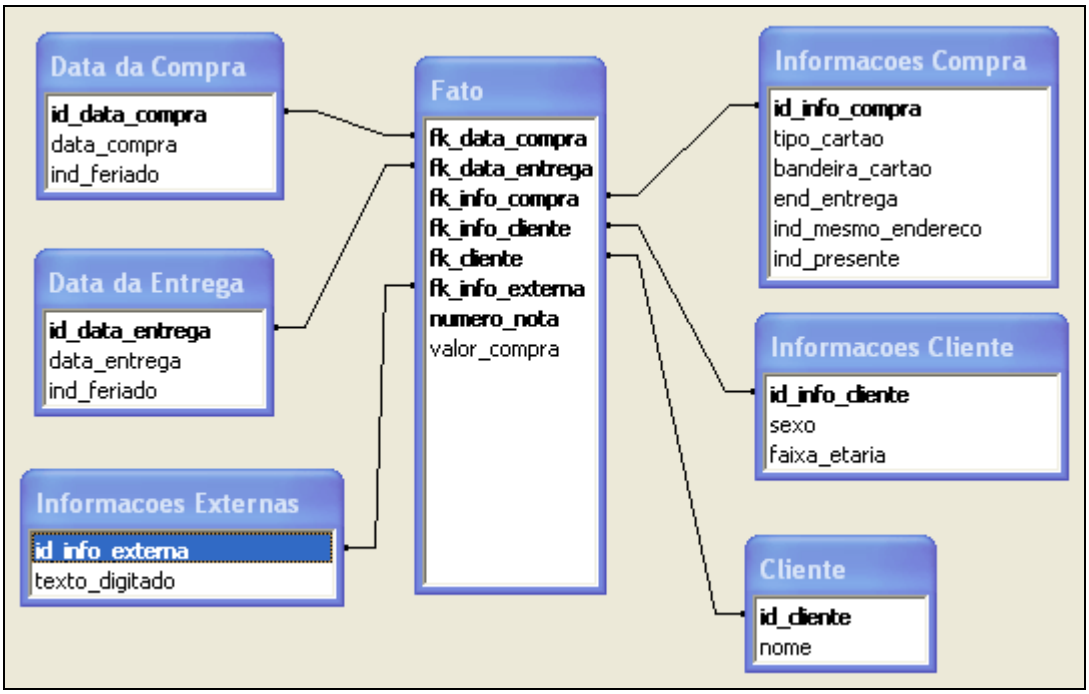

Figura 24. Modelo de vendas pela Internet

4.1.4.2 **Dimensões com mais de um papel (*role playing dimensions)***

É muito comum encontrarmos dimensões que possuem mais de um papel em um modelo dimensional (KIMBALL, ROSS, 2002). No modelo apresentado pela Figura 24, a dimensão de data possui um duplo papel: a data da compra e a data da entrega. Existem algumas soluções possíveis para esta característica de modelagem como, por exemplo, a criação de visões ou até mesmo a criação de tabelas distintas de conteúdo igual. Independente da solução adotada, cada papel deve ser mapeado como se fosse realmente uma dimensão à parte. No caso da utilização de visões, as mesmas deverão estar expressas na marca de "tabela". A Figura 25 ilustra o mapeamento no caso da Figura 24.

```
<dimensoes>
   <dimensao>
      <nomedim>Data da Compra</nomedim>
      <origem>I</origem>
      <tabela>Data_Compra</tabela>
      <categoria>
         <nomecat>Data da Compra</nomecat>
         <atributo id='dtcompra'>
            <nomeatrib>Data</nomeatrib>
            <coluna>data_compra</coluna>
            <tipo>1</tipo>         </atributo>      </categoria>
   </dimensao>
   <dimensao>
      <nomedim>Data da Entrega</nomedim>
      <origem>I</origem>
      <tabela>Data_Entrega</tabela>
      <categoria>
         <nomecat>Data da Entrega</nomecat>
         <atributo id='dtentrega'>
            <nomeatrib>Data</nomeatrib>
            <coluna>data_entrega</coluna>
            <tipo>1</tipo>
         </atributo>
      </categoria>
   </dimensao>
</dimensoes>
```

Figura 25. Mapeamento para dimensões com mais de um papel

4.1.4.3 **Dimensões demográficas**

Existem algumas dimensões cujo conteúdo é muito extenso, como a de clientes, por exemplo. Estas possuem atributos cujo domínio pode ser controlado e são de freqüente uso por parte do usuário de relatórios analíticos. Neste caso, é comum que o modelador dimensional crie uma dimensão à parte com estes atributos (dimensão demográfica) (KIMBALL, ROSS, 2002). O modelo da Figura 24 possui a dimensão “Cliente” e a dimensão “Informações do Cliente”, cujo conteúdo contempla a faixa etária do mesmo e seu sexo. Seria natural enquadrar os atributos da dimensão demográfica na mesma categoria da qual estes foram “desmembrados” no modelo dimensional. Entretanto, a concepção atual da solução não permite tal configuração, uma vez que atributos de uma mesma categoria obrigatoriamente estão em uma mesma tabela. Com o mecanismo atual, as dimensões demográficas devem ser tratadas como outra qualquer, sem ligação explícita na taxonomia com a dimensão da qual foi desmembrada. Um caminho natural para diminuir o impacto deste problema é agrupar as duas dimensões em uma mesma faceta. A Figura 26 exemplifica o mapeamento para o modelo da Figura 24.

```
<dimensoes>
   <dimensao>
      <nomedim>Informações_Cliente</nomedim>
      <origem>I</origem>
      <tabela>Informacoes_Cliente</tabela>
         <categoria>
            <nomecat>
               Informações Adicionais do Cliente
            </nomecat>
            <atributo id='sexo'>
               <nomeatrib>Sexo</nomeatrib>
               <coluna>sexo</coluna>
               <tipo>1</tipo>
            </atributo>
            <atributo id='faixa'>
               <nomeatrib>Faixa Etária</nomeatrib>
               <coluna>faixa_etaria</coluna>
               <tipo>1</tipo>
            </atributo>
         </categoria>
   </dimensao>
</dimensoes>
```

Figura 26. Mapeamento para dimensões demográficas

4.1.4.4 **Dimensões degeneradas (descaracterizadas)**

Dimensões degeneradas são dimensões que possuem somente um atributo, cuja granularidade é a mesma da tabela de fatos. Neste caso, o atributo é armazenado diretamente na tabela de fatos, sem a necessidade de se criar uma tabela para representar a dimensão (KIMBALL, ROSS, 2002). Na Figura 24, a nota fiscal foi modelada como uma dimensão degenerada (coluna “numero_nota” da tabela fato). O mapeamento será realizado de maneira natural, criando-se uma dimensão na faceta correspondente. Esta dimensão terá uma única característica, com um único atributo, que irá apontar diretamente para a tabela de fatos, ao invés de uma outra tabela dimensional, conforme a Figura 27. Deve-se, entretanto, tomar muito cuidado com a utilização destes casos, uma vez que o mecanismo de indexação irá varrer toda a tabela de fatos no momento de criação dos índices. Como as tabelas de fatos são muito grandes, a utilização de dimensões degeneradas pode causar problemas graves de desempenho no mecanismo da solução e, em casos extremos, com muitas destas situações ocorrendo em várias tabelas de fatos gigantescas, até mesmo inviabilizá-lo.

```
<dimensoes>
   <dimensao>
      <nomedim>Nota Fiscal</nomedim>
      <origem>I</origem>
      <tabela>Fato</tabela>
      <categoria>
         <nomecat>Nota Fiscal</nomecat>
         <atributo id='numnota'>
            <nomeatrib>Nº da Nota</nomeatrib>
            <coluna>numero_nota</coluna>
            <tipo>0</tipo>
         </atributo>
      </categoria>
   </dimensao>
</dimensoes>
```

Figura 27. Mapeamento para dimensões degeneradas

4.1.4.5 **Dimensões sujas (*dirty dimensions*)**

Dimensões sujas (KIMBALL, ROSS, 2002) ocorrem quando o conteúdo dimensional não é confiável . Isso acontece normalmente quando o conteúdo é proveniente de entradas diretas do usuário. No modelo da Figura 24, a dimensão “Informações Entradas” contém textos provenientes de um campo de livre digitação na página de finalização de compras, onde o usuário pode escrever sobre sua experiência no sítio, problemas encontrados, fazendo críticas ou sugestões. Neste caso, sua utilização não é recomendada, uma vez que o atributo a ser usado para navegação e indexação não irá obter um resultado satisfatório. Caso seja extremamente necessário, mecanismos de tratamento devem ser aplicados na dimensão na tentativa de agregar um maior grau de confiabilidade à informação. A Figura 28 ilustra como seria o mapeamento para o modelo da Figura 24.

```
<dimensoes>
   <dimensao>
      <nomedim>Informações Externas</nomedim>
      <origem>I</origem>
      <tabela>Informacoes_Externas</tabela>
      <categoria>
         <nomecat>Observações do usuário</nomecat>
         <atributo id='texto'>
            <nomeatrib>
               Texto Inserido pelo Usuário
            </nomeatrib>
            <coluna> texto_digitado </coluna>
            <tipo>1</tipo>
         </atributo>
      </categoria>
   </dimensao>
</dimensoes>
```

Figura 28. Mapeamento para dimensões "sujas"

4.1.4.6 **Dimensões sucata (*junk dimensions*)**

Quando agrupamos em uma única tabela dimensional atributos diversos que não se correlacionam, chamamos esta de dimensão sucata(KIMBALL, ROSS, 2002). No modelo da Figura 24, informações extras sobre a compra, como o cartão utilizado, se esta incluiu embalagem para presente e se o endereço de entrega era o mesmo do cliente, foram agrupadas na dimensão “Informações Adicionais”. Apesar de estar em uma única tabela, o mapeamento deve ser realizado como se fossem dimensões distintas, cada uma pertencendo à faceta pertinente, de acordo com a categoria associada, possuindo seus respectivos atributos. Entretanto, as diversas dimensões encontradas irão apontar para a mesma tabela do banco de dados no arquivo de configuração. Podemos observar um fragmento deste mapeamento, para o modelo da Figura 24, na Figura 29. Um exemplo completo é apresentado no Anexo 3.

```
...
   <dimensao>
      <nomedim>Cartao de Credito</nomedim>
      <origem>I</origem>
      <tabela>Informacoes_Compra</tabela>
      <categoria>
         <nomecat>Cartão de Crédito</nomecat>
         <atributo id='tp'>
            <nomeatrib>
               Débito ou Crédito
            </nomeatrib>
            <coluna>tipo_cartao</coluna>
            <tipo>1</tipo>
         </atributo>
         ...
   </dimensao>
   <dimensao>
      <nomedim>Endereco de Entrega</nomedim>
      <origem>I</origem>
      <tabela>Informacoes_Compra</tabela>
      <categoria>
         <nomecat>Endereço de Entrega</nomecat>
         ...
```

Figura 29. Mapeamento para dimensões "sucata"

4.2 Aprimorando a taxonomia com recursos específicos de domínio

Na seção anterior vimos como obter uma taxonomia facetada a partir de uma análise sobre o modelo dimensional do DW. Explorando as semelhanças existentes entre uma taxonomia e as dimensões do DW, conseguimos obter uma taxonomia que servirá de base para a exploração conjunta dos dados presentes no universo estruturado e no universo não-estruturado.

Em um universo ideal, no qual todas as informações necessárias ao negócio foram previstas, modeladas e disponibilizadas no DW, a taxonomia criada permitirá uma total exploração dos dois universos, uma vez que esta irá conter todos os termos relevantes para o assunto, já organizados de maneira pertinente pelo processo de modelagem analítica. Entretanto, caso alguma informação não exista no modelo dimensional, mas represente valor na coleção de documentos, esta ficaria de fora do processo exploratório, uma vez que a taxonomia criada estaria restrita aos mecanismos aplicados até o momento.

Existem diversos recursos específicos de domínio, como Glossários, Taxonomias e Ontologias que podem ser utilizados para enriquecer a taxonomia facetada criada,

aumentando a capacidade do processo exploratório. Em um universo corporativo, podem existir até mesmo tabelas, em outros sistemas que não o DW, que auxiliem no processo de identificação e classificação de informações do negócio.

O desafio consiste em permitir a utilização destes recursos no processo de construção da taxonomia facetada, não só pela adição de termos relevantes, mas no enquadramento destes na estrutura hierárquica da base criada. Para cada tipo de fonte, teremos que adotar um procedimento que obtenha o termo e a classificação facetada deste, de maneira a enquadrá-lo no mecanismo exploratório obtido. O universo de recursos específicos de domínio é muito vasto e possui inúmeras peculiaridades. Neste trabalho procuramos identificar algumas dessas possibilidades e sugerir uma técnica de absorção destas à taxonomia facetada. Veremos, no capítulo 6, que na aplicação do protótipo desenvolvido em um domínio médico foi utilizada uma tabela de domínio como recurso para enriquecer o mecanismo exploratório.

Na Figura 30, temos a ilustração de como se configura o universo de exploração disponível para o usuário de acordo com a cobertura realizada pelo processo de identificação e mapeamento da taxonomia facetada que irá servir de mecanismo exploratório. Considerando que as linhas tracejadas em diagonal representam as informações mapeadas no processo, temos na Figura 30 (A) uma situação na qual o modelo dimensional foi construído de maneira a comportar todas as informações necessárias ao processo analítico. Na Figura 30 (B) a área tracejada, demarcada pelo círculo, representa as informações disponíveis no modelo dimensional do DW, enquanto a área total do retângulo representa o universo de informações importantes ao processo decisório. Podemos observar que algumas informações, representadas pelas áreas brancas, estarão fora da exploração conjunta, disponível através da taxonomia facetada criada.

Ao utilizar um recurso que contenha um conhecimento prévio do domínio em questão (chamaremos neste trabalho de “recurso de domínio”), estamos possibilitando a utilização deste conhecimento no processo de exploração conjunta, permitindo que as informações presentes sejam exibidas ao usuário durante sua navegação sobre os universos. Na Figura 30 (C) representamos este conhecimento prévio pela área preenchida no canto superior esquerdo. A Figura 30 (D) ilustra a absorção destas informações. Seguindo a idéia apresentada pela Figura 30, as demais áreas brancas podem ser “preenchidas” através da utilização de outros

recursos de domínio, sendo perfeitamente possível que um recurso tenha a abrangência de mais de uma área de informação.

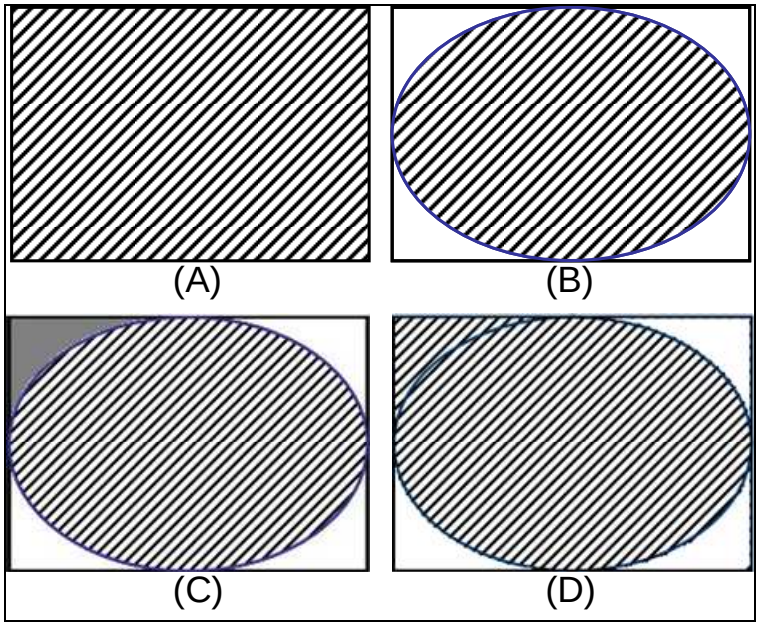

Figura 30. Configurações do universo de exploração disponível

Para utilizar o recurso específico de domínio, devemos primeiro analisá-lo e decidir se o mesmo resolve o problema causado pela ausência de determinada informação no modelo dimensional. Em positivo, devemos escolher quais informações presentes no recurso serão utilizadas para o enriquecimento do processo, uma vez que nem todas as informações presentes podem ser úteis para a exploração do universo em questão. Uma vez identificadas, o próximo passo, de fundamental importância para o processo, é o enquadramento destas informações na taxonomia facetada.

Cada conjunto de informações do recurso será representado através de um atributo na taxonomia utilizada para a exploração, tendo este que ser adicionado à taxonomia de base criada, conforme explicado na seção 4.1, ou seja, cada assunto coberto por um conjunto de informações será tratado como um atributo do mecanismo exploratório. Neste ponto algumas situações são possíveis, tais como:

- **Classificação em uma nova faceta**: Caso o conjunto de informações criado não possa ser mapeado em nenhuma faceta existente, devemos criar uma nova para adicioná-lo à mesma. Para tal, devemos criar uma nova marcação de “faceta” no arquivo de mapeamento de facetas. Conforme vimos na seção 4.1, uma faceta

sempre possui pelo menos uma categoria associada, o que significa que nesta situação devemos criar uma nova categoria no arquivo de mapeamento de categorias, associando a mesma com a faceta criada.

- **Classificação em uma nova categoria:** Caso o atributo (conjunto de informações) possa ser classificado em uma faceta existente, mas nenhuma categoria criada possa representá-lo de maneira adequada, devemos criar uma nova categoria no arquivo de mapeamento de categorias, associando a mesma com uma faceta existente no arquivo de mapeamento de facetas.
- **Classificação em uma categoria existente:** Podendo o atributo fazer parte de uma categoria existente, devemos simplesmente associá-lo à mesma. Esta associação será explicada a seguir.

Uma vez que a classificação dos atributos foi definida e descrita nos arquivos responsáveis, devemos realizar o mapeamento dos mesmos para torná-los parte do processo de navegação e indexação das informações Realizamos este mapeamento da mesma maneira como fazemos para atributos vindos do modelo dimensional do DW. Podemos agrupar os atributos obtidos em uma ou mais dimensões, de acordo com a representatividade dos mesmos. Conforme observamos na Figura 19, uma dimensão está associada a uma tabela onde os termos que as representam estarão armazenados. Nos casos em que a informação é proveniente de recursos externos, devemos então criar uma tabela contendo estas informações, para que os mecanismos responsáveis possam realizar seu processamento. Podemos denominar esta tabela, de “tabela de domínio”.

Neste ponto o tipo de recurso irá influenciar o tratamento a ser realizado. Alguns recursos são de simples incorporação. Tabelas de domínio podem ser diretamente mapeadas e “arquivos texto” podem, por um processo simples de carga, ser inseridos em uma tabela criada para auxiliar o processo. Para mecanismos mais complexos, como Taxonomias e Ontologias, a tática será a mesma, ou seja, criaremos uma tabela para armazenar as informações escolhidas. Entretanto, é necessário o desenvolvimento de um programa com uma lógica um pouco mais complexa para tal.

No arquivo de mapeamento, como a fonte de informação é externa, devemos atribuir o valor “E”, abreviação de “Externa”, na marca “origem” da marcação de dimensão, indicando que os termos presentes neste atributo não possuem origem no DW. Esta marcação irá

influenciar, conforme veremos nos capítulos 5 e 6, na indexação e na disponibilização das informações. Estando a tabela carregada e apontada corretamente no arquivo de configuração, as informações serão incorporadas ao processo de exploração através do acionamento dos mecanismos desenvolvidos, que serão apresentados no capítulo 5.

A Figura 31 apresenta o fluxo de atividades necessárias para a utilização de recursos de domínio no enriquecimento da taxonomia facetada base, criada através da análise do modelo dimensional. O processo se inicia com a escolha do recurso e termina com a execução dos mecanismos necessários.

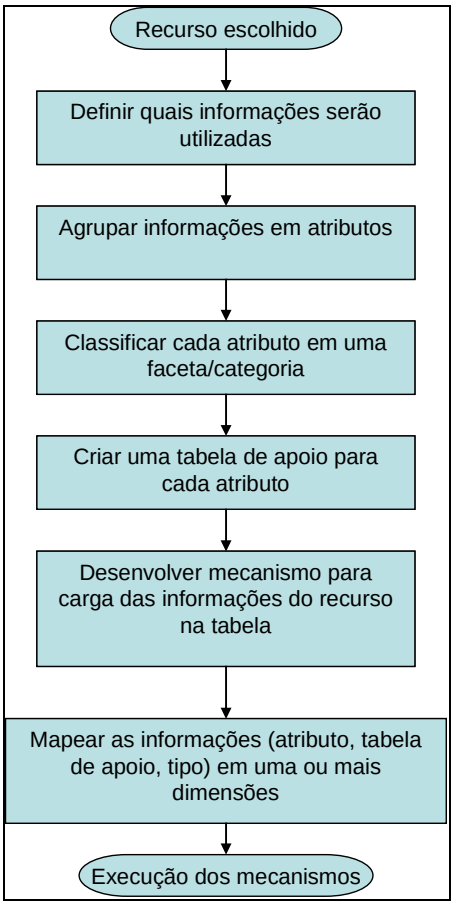

Figura 31. Processo de enriquecimento da taxonomia base com a utilização de recursos específicos de domínio

4.3 Explorando conjuntamente os dois universos

O enriquecimento da análise das informações só será possível através da construção de um mecanismo eficiente, que permita a realização da exploração conjunta entre os dois universos. No capítulo 5 vamos apresentar a arquitetura da solução elaborada para permitir tal exploração. Esta irá então possibilitar ao usuário dois caminhos a seguir, cada um representando uma via de acesso para o universo desejado. Veremos a seguir nossa idéia de exploração conjunta, na qual o usuário, a partir de um universo, pode enriquecer sua análise com informações do outro universo. A seção 4.3.1 irá descrever como o usuário, a partir de uma exploração na coleção de documentos, pode analisar dados presentes no universo estruturado (DW). A seção 4.3.2 descreve o "caminho contrário", no qual o usuário parte de um relatório, com informações provenientes do mundo estruturado (DW) e obtém informações relacionadas a estas do mundo não-estruturado (coleção de documentos).

4.3.1 De documentos não-estruturados para DW

Este caminho tem como objetivo enriquecer as informações retornadas pela exploração do universo não estruturado com informações do DW. Tendo a taxonomia facetada criada à sua disposição, o usuário vai realizar sua exploração sobre a estrutura criada, onde serão exibidas as categorias, atributos e termos pertinentes a cada faceta. A busca pela informação se assemelha a uma busca em um mecanismo OLAP tradicional, na qual o usuário pode expandir ou restringir o universo de informação a seu dispor, assim como aplicar filtros para limitar este de maneira a selecionar somente as fatias de dados que lhe interesse. Um aspecto importante desta exploração está na representação do fato. No momento da navegação, o fato se caracteriza pelo documento, ou seja, o resultado final da consulta será um conjunto de documentos, presentes na coleção, que satisfaçam à seleção realizada pelo usuário.

Entretanto, apresentar somente os documentos não permite ao usuário uma maior exploração dos dados disponíveis, necessitando ainda de uma ligação envolvendo os dados no DW estruturado. Esta ligação acontece através da exibição dos termos encontrados, para cada documento retornado, que possuem entrada na estrutura de índices criada, ou seja, termos que,

através de seus atributos, foram selecionados para fazer parte da exploração analítica do universo de informação.

Além dos termos, é importante ressaltar as respectivas relevâncias que estes possuem em cada documento. Os termos são apresentados separados pelas categorias e atributos que os englobam, contendo um link para o arquivo do documento, de maneira a propiciar a visualização do mesmo caso seja necessária ao usuário. Um importante aspecto deve ser ressaltado: os termos apresentados são representantes do domínio, podendo ou não possuir correspondência nas dimensões do DW estruturado. No capítulo 5 vamos observar o tipo de tratamento proposto para cada um dos casos. O usuário então pode selecionar os termos encontrados em um ou mais documentos, e estes serão usados para a construção da consulta que será enviada ao DW para a construção do relatório.

Um relatório OLAP, que será o resultado da consulta enviada ao sistema gerenciador de banco de dados (SGBD) do DW, é composto por atributos e métricas. Ao selecionar os termos nos documentos retornados pela busca na coleção não-estruturada, estamos definindo os atributos, e seus respectivos filtros, que farão parte do relatório. Não estaríamos, entretanto, definindo quais métricas seriam analisadas sob estas perspectivas. Conforme mencionado, os fatos retornados pela solução, neste momento, são os próprios documentos, que não representam as métricas do DW.

A solução para esta situação está na apresentação das métricas disponíveis no DW para que o usuário realize sua escolha. Ou seja, uma vez escolhidos os termos, o usuário deve escolher uma ou mais métricas, dentre as já existentes no DW, e adicioná-las à consulta ao mesmo.

Considerando um modelo analítico completo de um DW de uma grande (ou até mesmo média) empresa, apresentar todas as métricas disponíveis nas tabelas de fatos seria uma tarefa muito custosa, em termos de processamento, e ao mesmo tempo muito pouco eficiente, uma vez que uma métrica pode não possuir análise sobre a ótica de determinado atributo. Além disto, o usuário seria apresentado a uma lista com inúmeras possibilidades, tornando a tarefa de escolha da métrica muito complicada para o mesmo. Para solucionar este problema, a arquitetura concebida fará uso de um arquivo de mapeamento do modelo

dimensional, a ser descrito no capítulo 5. Através da utilização deste, somente métricas que podem ser analisadas com os atributos escolhidos são exibidas.

Podemos então imaginar o seguinte cenário para ilustrar uma exploração iniciando no universo não-estruturado. Em uma coleção de documentos pertinentes à área tributária, contendo notícias sobre variações da carga de impostos aplicada no Brasil, poderíamos navegar na taxonomia disponibilizada, encontrando diversos documentos se referindo a Brasília. Selecionaríamos então o termo “DRF-Brasília” e o submeteríamos ao DW, obtendo como retorno o relatório apresentado na Figura 32. Neste, o usuário pode analisar as informações e voltar a realizar a consulta à base de documentos.

CPF do Declarante	DRF Jurisdição do Declarante		Bairro/Distrito do Declarante	Ocupação Principal do Declarante	Métrica	Valor Imposto a Ser Restituído Pág. Resumo
00000031018	01.1.01.00	DRF - Brasília	GUARA I	Não informado		8,353.21
00000059055	01.1.01.00	DRF - Brasília	SOBRADINHO	Economista, administrador, contador, auditor e afins		30,131.98
00000309778	01.1.01.00	DRF - Brasília	LAGO SUL	Outras ocupações não especificadas anteriormente		5,667.36
00002216978	01.1.01.00	DRF - Brasília	GUARA II	Não informado		4,361.23
00002350313	01.1.01.00	DRF - Brasília	LAGO SUL	Médico		36,625.17

Figura 32. Seleção de termos baseada no relatório retornado

Teríamos então, neste cenário, uma análise conjunta entre uma coleção de documentos com notícias relacionadas a tributos e um DW com informações sobre pagamentos, recebimentos e declarações que expressam o relacionamento do contribuinte com a União. Através desta abordagem, por exemplo, análises sobre quedas de arrecadação poderiam ser enriquecidas por documentos que ilustrariam o cenário da época, o que, nas mãos de um especialista, pode resultar em uma descoberta.

Os passos de uma consulta que começa no universo não-estruturado e realiza a análise no universo do DW estruturado seriam os seguintes:

1. O usuário navega na taxonomia criada, escolhendo os atributos das categorias desejadas.
2. Através da expansão dos atributos, os termos mapeados para estes são exibidos.

3. O usuário escolhe os termos para os quais deseja realizar sua busca e o tipo de visualização (maiores detalhes na seção 5.5).
4. Os documentos nos quais os termos foram encontrados são apresentados para o usuário, com um link de acesso rápido para o mesmo. O número de vezes que o termo ocorre neste documento também é exibido.
5. O usuário escolhe os termos desejados e solicita a geração da consulta.
6. Será exibida uma lista das métricas presentes no DW estruturado, de acordo com a categoria dos termos escolhidos, para que este possa selecionar uma ou mais para análise.
7. Uma consulta é montada e enviada ao DW estruturado.
8. A consulta é executada no DW estruturado e os resultados são retornados ao usuário em formato de um relatório OLAP.

4.3.2 Do DW para documentos não-estruturados

Dado um relatório retornado pelo DW, poder-se-á encontrar os documentos da coleção que ampliem a análise do resultado. Esta exploração vai possibilitar ao usuário navegar de um relatório OLAP até os documentos de uma coleção. A solução encontrada foi a criação uma nova operação OLAP, denominada "*drill* para documento", na qual o usuário escolhe um ou mais termos, disponibilizados nos relatórios, e os utiliza como entrada de uma busca.

A maioria das ferramentas OLAP encontradas no mercado disponibilizam uma operação semelhante, na qual o usuário pode realizar o detalhamento de uma informação, de acordo com novas perspectivas, clicando na informação desejada e selecionando um novo modelo de exibição do mesmo. Ao estar em um relatório com dados referentes a notas fiscais, o usuário poderia, por exemplo, selecionar os códigos das notas desejadas e solicitar a análise dos dados pertinentes aos itens das mesmas. Esta análise pode ser realizada através de um novo modelo, construído de maneira a adequar as informações à nova granularidade. A operação sugerida segue o mesmo conceito, tendo como principal diferença o destino dos dados, que seria a coleção dos documentos.

As informações selecionadas nos relatórios funcionam como filtro para o mecanismo de busca de documentos da coleção. Assim, ao navegar por este caminho, os passos realizados pelo usuário em uma consulta seriam os seguintes:

1. O usuário navega em uma ferramenta OLAP, escolhendo os atributos e as métricas que irão formar o relatório.
2. O relatório é executado e os dados presentes no DW são recuperados e exibidos para o usuário.
3. O usuário analisa os resultados e escolhe os termos das linhas retornadas que sejam relevantes e necessitem de uma análise mais ampla.
4. Os termos presentes na seleção, provenientes dos domínios encontrados para os atributos, são então submetidos à coleção de documentos.
5. O mecanismo de busca será acionado, tratando os termos como filtros.
6. Os documentos contendo os termos desejados serão exibidos, juntamente com a freqüência de cada termo para o documento em questão.

A Figura 32 apresentou um relatório, submetido a um DW estruturado, contendo informações de valores a restituir de uma declaração de imposto de renda para determinados CPFs. Baseado no resultado obtido, o usuário poderia então selecionar os termos relevantes a serem pesquisados na coleção de documentos. No caso da Figura 32, os termos que seriam utilizados no critério de busca seriam “Sobradinho”, “Lago Sul” e “Médico”. Estes foram escolhidos devido ao grande valor apresentado no resultado da consulta ao modelo dimensional.

Podemos observar que, através dos caminhos apresentados, a análise pode ser feita em qualquer direção a qualquer momento, ou seja, o usuário pode partir do mundo estruturado, realizar a análise dos documentos de acordo com atributos escolhidos e voltar para o DW até mesmo com outras informações que este venha a encontrar na coleção. A Figura 33 exemplifica como seria realizada uma análise utilizando-se as duas vias. As setas pontilhadas indicam o fluxo de informações dentro de um mesmo ambiente (estruturado ou não-estruturado). A seta de número 1 representa uma análise que começa na coleção de documentos, o universo não-estruturado, e termina no ambiente estruturado, no DW. A seta 2 ilustra o caminho oposto. Entretanto, podemos considerar as duas como parte de um único processo, com o usuário indo e vindo de ambos os universos através dos mecanismos

apresentados na solução, possibilitando ao mesmo um grande enriquecimento de suas análises.

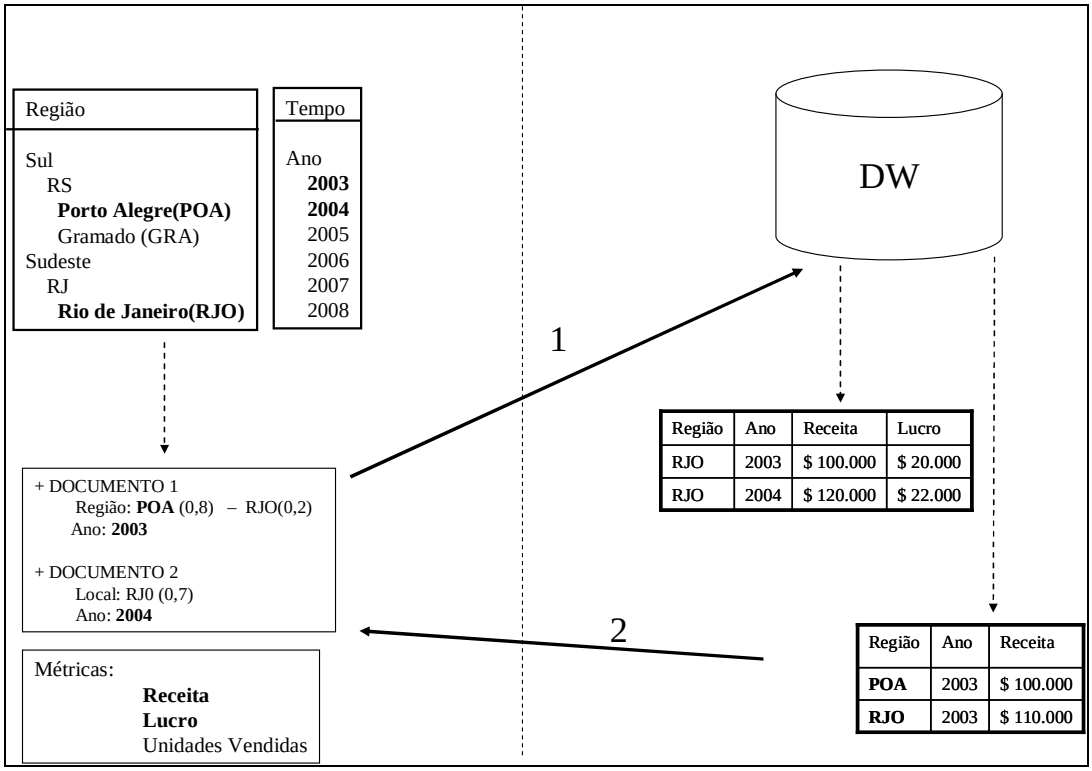

Figura 33. Exemplo de análise em "mão-dupla"

O sucesso desta abordagem depende em muito do sucesso de dois mecanismos: a criação de uma taxonomia que represente bem o domínio e a indexação dos documentos de acordo com os termos relevantes para o assunto. A importância do primeiro já foi discutida ao longo deste trabalho, porém vale ressaltar a importância da visão conjunta, em termos de classificação, das categorias a serem criadas e das dimensões apresentadas pelo modelo de dados dimensional.

O mecanismo de indexação deve ser capaz de reproduzir de maneira mais fidedigna o conteúdo dos mesmos em relação ao conteúdo das dimensões. É através deste conteúdo que a tabela de índices, utilizada pela busca, é criada e seu preenchimento influencia diretamente na precisão dos resultados. Maiores detalhes sobre a geração do índice serão apresentados no capítulo 5.

4.4 Considerações Finais

Vimos neste capítulo que através de uma taxonomia facetada podemos permitir uma exploração conjunta entre informações provenientes de um universo estruturado com informações provenientes de um universo não-estruturado. A construção da mesma foi apresentada, sendo o modelo dimensional o ponto de partida. Através de seu enriquecimento com recursos específicos de domínios, podemos cobrir um universo maior de informações passíveis de análises.

Construir somente a taxonomia facetada não é suficiente para fornecer a exploração desejada. Uma arquitetura que a utilize como mecanismo exploratório e ofereça o suporte necessário para a exploração é necessária. Idealizamos esta arquitetura e apresentaremos a mesma no capítulo 5. Após a elaboração desta arquitetura, para evidenciar que as propostas são capazes de oferecer a capacidade analítica desejada, ou seja, oferecer a exploração conjunta dos universos estruturado e não-estruturado, desenvolvemos um protótipo, que será apresentado na seção 5.4.

5. Arquitetura e Protótipo da solução

O principal objetivo da arquitetura da solução é permitir a exploração conjunta de dados presentes em um universo estruturado e dados presentes em um universo não-estruturado. Nesta solução, a associação discutida ao longo deste trabalho entre as dimensões de um DW e uma taxonomia facetada é fortemente utilizada, sendo esta construída especificamente para representar o domínio da informação sendo explorada. Para possibilitar uma análise, por parte do usuário, nos dois ambientes (estruturado e não-estruturado), foram desenvolvidos mecanismos que tentam preencher a lacuna existente entre estes universos, tornando possível uma análise de "mão-dupla" sobre a informação.

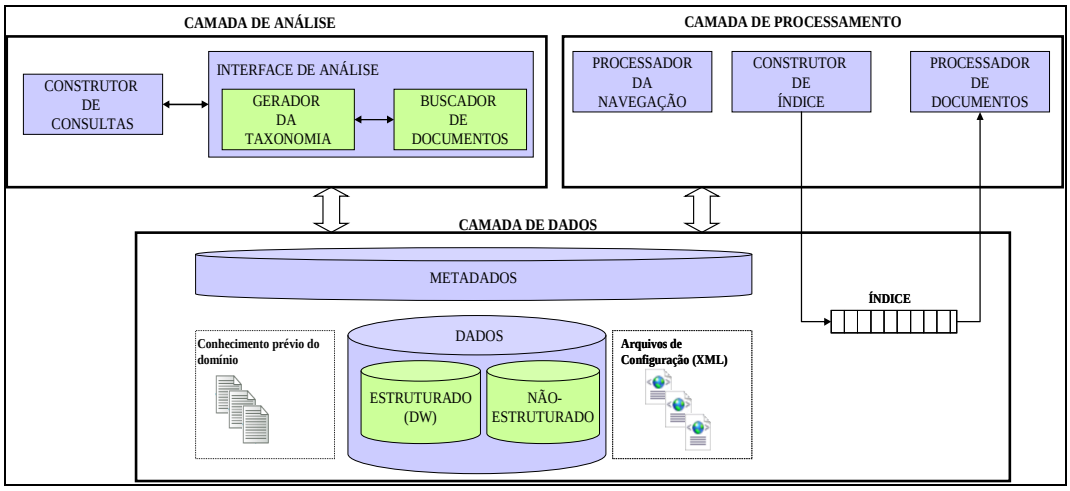

Figura 34. Arquitetura da solução

A Figura 34 ilustra a arquitetura proposta. Esta é composta por três camadas: Processamento, Dados e Análise. A camada de processamento é responsável pela indexação dos documentos do repositório não-estruturado e pela construção da taxonomia facetada que apóia o processo exploratório. Podemos resumi-la como sendo a infra-estrutura necessária para a solução. Como entradas da camada, temos os documentos, de natureza não-estruturada, presentes na coleção, os arquivos de configuração, apresentados no capítulo 4, e os representantes do domínio. Estes últimos são o conteúdo das dimensões do DW e os recursos específicos de domínio utilizados.

Na camada de dados estão presentes os metadados e os dados necessários para a realização da exploração. A camada de análise é responsável por fornecer a exploração

conjunta ao usuário, apresentando o mecanismo exploratório, no caso desta solução a taxonomia facetada, e oferecendo os mecanismos necessários para a visualização e a navegação sobre as informações disponíveis. Cada camada será explicada, em maiores detalhes, nas seções 5.1, 5.2 e 5.3.

De maneira a aplicar os conceitos apresentados por cada camada, desenvolvemos um protótipo, que será apresentado na seção 5.4. Como a visualização das informações possui fundamental importância em um ambiente analítico, foi criada uma seção específica, a 5.5, para abordar o tema. Finalizando o capítulo, a seção 5.6 apresenta algumas considerações sobre a arquitetura e o protótipo.

5.1 Camada de processamento

A camada de processamento pode ser dividida em três mecanismos principais: Processador da Navegação, Construtor de Índice e Processador de Documentos. O Processador da Navegação faz uso de três documentos de configuração, criados e armazenados em XML. Estes servem de base para a navegação e a conexão entre os elementos da arquitetura, tendo sido apresentados no capítulo 4. O primeiro representa as facetas (Figura 20); o segundo contém as categorias que as compõem (Figura 21); o relacionamento das categorias com as dimensões do DW é criado em um terceiro arquivo (Figura 19). Todos os três documentos de configuração são processados e armazenados no repositório de metadados, servindo de base para os demais mecanismos. A taxonomia facetada que apóia a exploração é armazenada, neste momento, nas tabelas do repositório de metadados, sendo posteriormente lida e apresentada ao usuário pelos mecanismos presentes na camada de análise (seção 5.3).

O Construtor de Índice é responsável, juntamente com o Processador de Documentos, pela ligação entre os dados do universo estruturado e do não-estruturado. As dimensões do DW são lidas e seus atributos são armazenados no repositório de metadados. Como as dimensões representam os diferentes aspectos sobre os quais um fato pode ser analisado, as informações presentes nos atributos se tornam então índices naturais para documentos cujo conteúdo faz parte do contexto sendo explorado. Baseado nas informações armazenadas no

repositório de metadados pelo mecanismo anterior, o índice é construído unindo termos, documentos e dimensões.

Além da leitura do conteúdo dimensional, os recursos específicos de domínio, também apontados pelos arquivos de configuração, são lidos e armazenados no repositório de metadados, fazendo parte do índice. Sua ligação com alguma dimensão do DW será determinada pela configuração do mesmo. O Construtor de Índice realiza a leitura desta informação e a grava no repositório de metadados. Veremos, na seção 5.3 (camada de análise) que este indicador fará com que a informação seja apresentada de maneira diferente para o usuário. Paralelamente ao preenchimento do repositório de metadados, um índice auxiliar é construído e armazenado na memória, tornando o processamento dos documentos mais ágil.

Os documentos armazenados em um determinado diretório são lidos e processados pelo Processador de Documentos. Este analisa o índice construído e, caso encontre algum termo no documento que pertença a este, uma referência para o documento é armazenada no repositório de metadados, assim como o número de vezes que o termo aparece no mesmo. É neste momento que é realizada a conexão entre o universo estruturado e o não-estruturado. Como sabemos, pelos passos anteriores, que o termo veio de uma determinada dimensão, o link entre estes pode ser criado e armazenado no repositório de metadados.

A Figura 35 representa o funcionamento da camada de processamento. As setas representam o fluxo da informação entre os componentes da mesma, possuindo uma numeração associada que indica a ordem em que este ocorre. As setas que têm como destino o repositório de metadados possuem a informação do conteúdo que é enviado para armazenamento no mesmo. O primeiro passo da camada é a leitura dos arquivos de configuração, tendo como entrada as informações dos arquivos e como saída as facetas, categorias, atributos e dimensões, sendo estas e seus relacionamentos armazenados no repositório de metadados (passo 2). Após a carga destas, será realizada a análise e leitura do DW (passo 3). Através do Construtor de Índice, as tabelas dimensionais e os recursos de domínio, se existirem, são lidos e, a partir do seu conteúdo, o índice é criado. Ao mesmo tempo em que os termos e suas respectivas relações com as dimensões, são armazenados no repositório de metadados (passo 4), uma lista com estes elementos é construída na memória (passo 4’), de maneira a funcionar como acelerador no processamento dos documentos.

Tendo as informações de facetas, categorias, atributos, termos e dimensões processadas e armazenadas, o próximo passo (passo 5) é a análise dos documentos, de conteúdo não-estruturado. Os documentos a serem analisados deverão estar agrupados em uma única pasta, pré-definida para o projeto. O Processador de Documentos realiza uma leitura seqüencial de cada documento presente nesta pasta e, com o auxílio do índice, identifica aqueles em que exista a presença de termos relevantes. Quando o termo é identificado, o processador realiza uma contagem de quantas vezes o mesmo ocorre no documento em questão, possibilitando ao usuário uma noção maior de relevância do termo no documento.

Após o processamento do documento, o Processador de Documentos armazena as informações no repositório de metadados (passo 6). Podemos observar na Figura 35 que este tem como saída termos e documentos. O documento e o relacionamento deste com o termo são as informações que faltavam no processamento da camada. Além de uma referência para o documento, são gravadas no repositório de metadados outras informações auxiliares sobre o mesmo, como o caminho para acesso e a freqüência do termo neste.

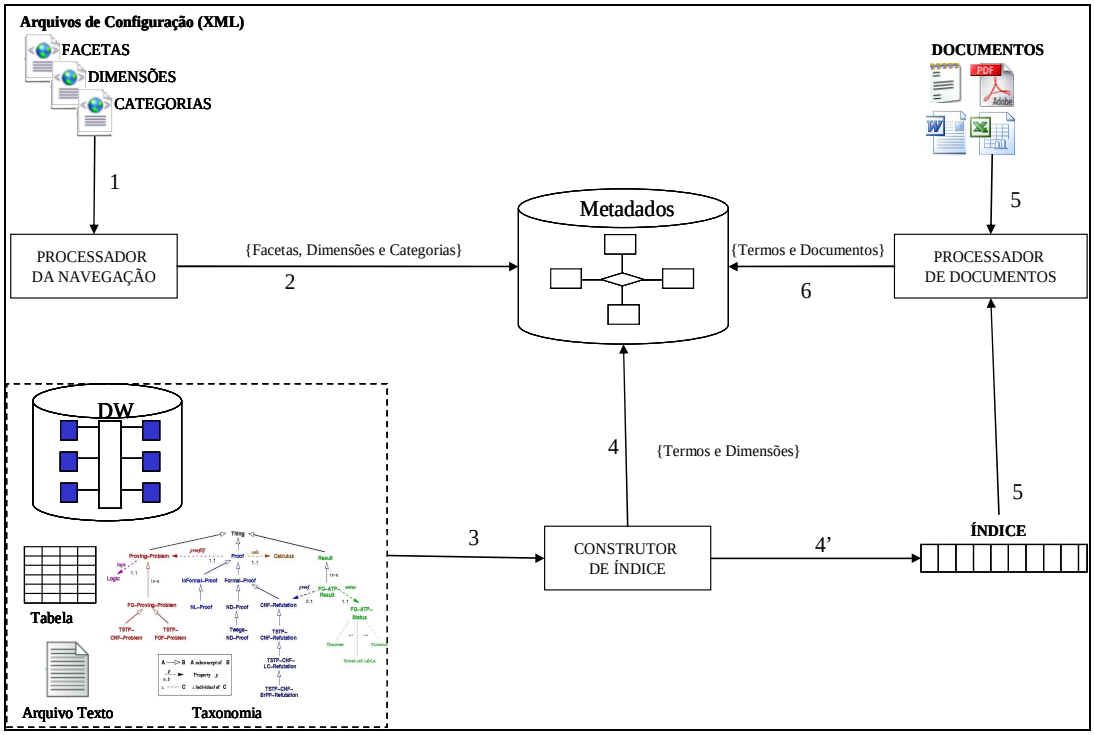

Figura 35. Funcionamento da camada de processamento

5.2 Camada de dados

A camada de dados é composta pelos repositórios de dados (estruturados e não-estruturados), pelo repositório de metadados, contendo o mapeamento criado entre os dois universos, e por recursos que expressem um conhecimento prévio do domínio sendo estudado. Os três juntos podem ser encarados como representantes do conhecimento da companhia. Além destes, os arquivos de configuração também são considerados integrantes desta camada. As operações OLAP são executadas na base estruturada (DW), cujo modelo é construído para oferecer o apoio necessário às operações analíticas. Na Figura 34, a representação dos dados de natureza não-estruturada pode dar a impressão que estes estão armazenados em uma base de dados. Entretanto, conforme mencionado na descrição da camada de processamento, este repositório é na verdade um diretório do sistema operacional, contendo todos os documentos relevantes para o processo exploratório do universo de informações desejado. Nenhum processo de estruturação é realizado em cima dos dados não-estruturados, ou seja, estes não são carregados em um banco de dados para que possam ser analisados.

O repositório de metadados corresponde a um conjunto de tabelas que se relacionam com o objetivo de possibilitar a aplicação das técnicas apresentadas pela abordagem. Na seção 4.1.1 foram definidos os conceitos que fazem parte do mecanismo de apoio à exploração. A Figura 36 apresenta o metamodelo criado de acordo com estes conceitos. Conforme vimos no capítulo 4, uma faceta possui uma ou mais categorias. Cada categoria, por sua vez, possui um ou mais atributos. Um termo pode estar presente em um ou mais documentos, assim como um documento pode possuir um ou mais termos. Um atributo pode estar presente em uma ou mais dimensões, assim como uma dimensão possui no mínimo um atributo. Os termos são associados a um atributo, podendo um termo estar presente em mais de um atributo. Naturalmente, um atributo possui um ou mais termos. A relação hierárquica entre atributos é expressa através de um auto-relacionamento da tabela correspondente. Alguns aspectos relativos às informações que serão armazenadas, assim com a ordem em que acontecem, foram já representados na Figura 35.

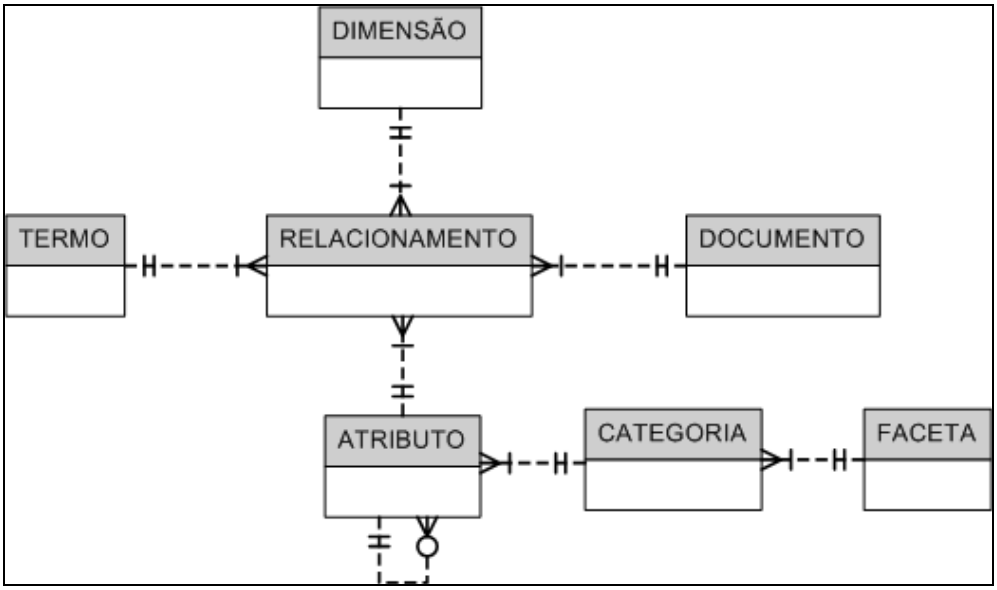

Figura 36. Metamodelo, utilizando o modelo de Martin (“Pé de Galinha”), de suporte à solução

5.3 Camada de análise

A camada de análise pode ser dividida entre a Interface de Análise e o Construtor de Consultas. Através da Interface de Análise é possível realizar a navegação na taxonomia facetada, disponibilizada pelo Gerador da Taxonomia, selecionando-se os termos desejados, obtendo como resposta os documentos que possuem relação com estes, assim como os demais termos presentes nos documentos que existam no domínio da informação. Estes últimos são aqueles presentes no índice construído com as informações dos atributos das dimensões. Estes termos podem ser enviados para o DW através de uma chamada ao construtor de consultas (explicado a seguir). No caso do usuário já estar trabalhando com um relatório estruturado, a interface possibilita a seleção dos atributos desejados no relatório para um posterior envio à coleção de documentos. O Construtor de Consultas é responsável pela construção da consulta a ser enviada para o DW. Baseado nos termos selecionados, o mesmo realiza uma leitura no repositório de metadados e constrói os comandos necessários para o acesso aos dados.

Na Figura 37 temos a ilustração de como funciona a camada de análise. Nesta figura, exemplifica-se o fluxo de análises que começam no universo não-estruturado e derivam para o universo estruturado (DW). As setas representam o fluxo da informação e sua numeração indica uma seqüência lógica de como esta irá trafegar durante a análise.

O usuário tem como ponto de acesso ao ambiente analítico um navegador de Internet padrão (Internet Explorer, Mozila Firefox, etc.). Ao acessar a URL do ambiente, o mecanismo Gerador da Taxonomia realiza uma leitura no repositório de metadados e monta a taxonomia que serve como mecanismo navegacional para o usuário. Após selecionar os termos dos atributos desejados, que são apresentados de maneira agrupada pelas categorias às quais pertencem, o usuário solicita a listagem de documentos que satisfazem à sua solicitação. Neste momento, o Buscador de Documentos é acionado, tendo este o seguinte funcionamento: uma análise no repositório de metadados é realizada e os termos pertinentes, assim como os documentos nos quais estes aparecem (com suas respectivas freqüências) são exibidos. Cada documento possui ainda um link para o arquivo, caso o usuário deseje visualizá-lo. Na seção 5.5 apresentamos as formas de visualização da informação desenvolvidas para a abordagem.

No momento da exibição das informações retornadas, a camada é responsável por indicar ao usuário se o termo encontrado possui ou não correspondência no modelo dimensional do DW. Esta não correspondência pode ocorrer devido à adição de termos vindos de recursos específicos de domínio, cujo conteúdo não possua relação direta com uma dimensão no DW. Tendo os termos à sua disposição, o usuário pode selecionar um ou mais e submetê-los a uma consulta no DW. Neste ponto atua o mecanismo denominado Construtor de Consulta, que é responsável por montar o comando de acesso à base (podendo este ser escrito em SQL, MDX, ou qualquer outra linguagem de acesso), enviando o mesmo para a base de dados.

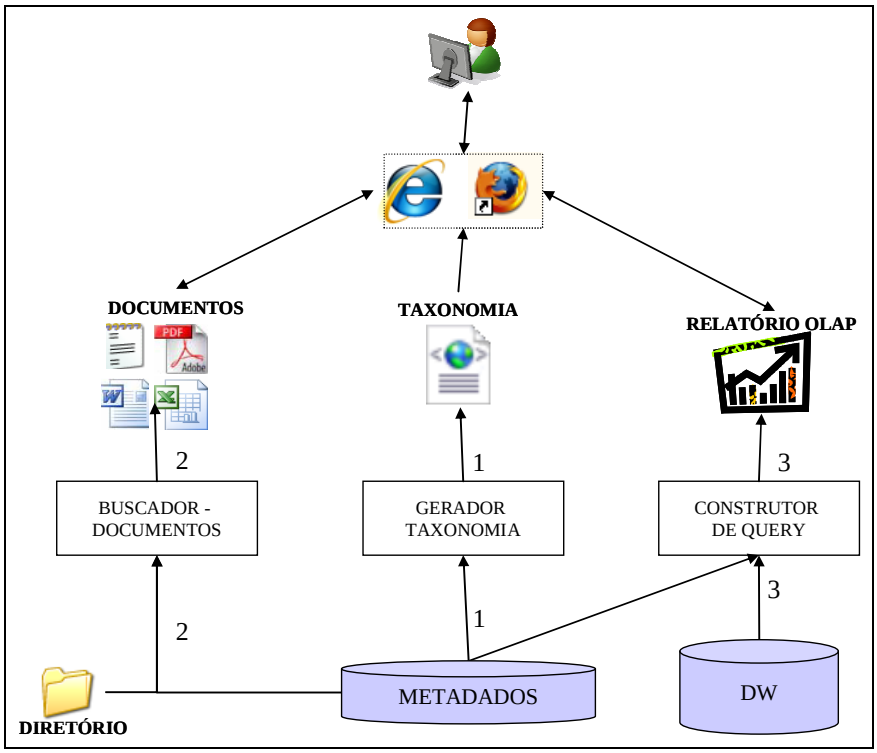

Figura 37. Camada de análise

Não temos, no repositório de metadados, informações sobre quais são as chaves primárias das dimensões, quais são as dimensões que compõem uma determinada chave da tabela de fatos, quais são as tabelas de fatos e quais fatos podem ser analisados nestas pelas diferentes combinações de dimensões. Todas estas informações devem ser utilizadas para a construção da consulta a ser enviada para o DW. Devemos então obtê-las de uma segunda fonte.

Teríamos algumas opções de como obter tais informações, como por exemplo, adicioná-las aos arquivos de mapeamento (descritos no capítulo 4). Esta solução traria alguns problemas como o excesso de informações a serem disponibilizadas nestes arquivos, além de tornar o processamento destes mais lento. Além disto, adicionar informações tais como chaves de tabelas ao arquivo iria desviar o mesmo de seu propósito, que é representar a ligação da taxonomia facetada com o modelo dimensional.

Outra alternativa seria a leitura dos metadados do SGBD para obter as informações. Esta solução teria como principal vantagem não necessitar de nenhuma configuração a mais na solução geral, bastando a adição de alguns métodos no mecanismo que implementa o

Gerador de Consulta. Um aspecto negativo a esta escolha seria a dependência do sistema gerenciador de banco de dados no qual o DW está hospedado. Diferentes SGBDs possuem diferentes formas de acesso a estas informações, o que tornaria o código desenvolvido dependente destes. Para contornar este problema, o código a ser desenvolvido deveria ser preparado para permitir uma gama de SGBDs conhecidos no mercado, o que tornaria a solução muito mais complexa. Outro problema identificado caso esta solução fosse adotada seria o fato da dependência do criador do banco de dados. Apesar de ser uma boa prática, quase obrigatória aos profissionais da área, nem sempre as chaves são criadas de maneira devida, ou seja, esta solução teria que assumir que o banco estaria criado corretamente.

A solução adotada para o problema na abordagem proposta é a criação de um arquivo XML que represente o modelo de dados dimensional. Nele estão contidas as descrições das tabelas de fatos que serão utilizadas, com as respectivas informações relevantes, e seus relacionamentos com as dimensões. Para exemplificar como seria criado este arquivo, tomemos como exemplo o modelo dimensional, obtido de (Moreira; Cordeiro; Campos, 2009), apresentado pela Figura 38. Este possui uma tabela de fatos cujas métricas “Valor da Fatura” e “Contador” podem ser analisadas sob a ótica das dimensões “Paciente”, “Tempo”, “Serviço” e “Convênio”. A Figura 39 mostra como seria o arquivo de configuração para o modelo da Figura 38.

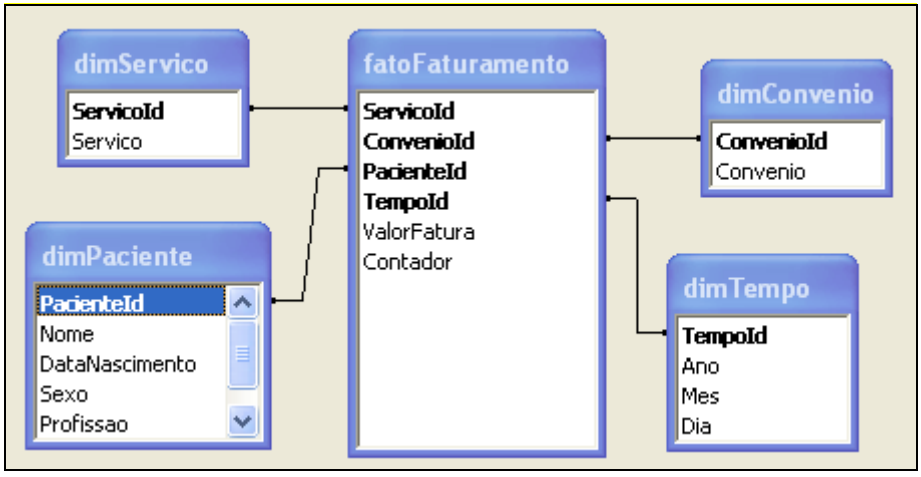

Figura 38. Modelo dimensional (Moreira; Cordeiro; Campos, 2009), descrito na Figura 39

```
<modelo>
 <fato nome="Faturamento" tabela="fatoFaturamento">
   <dimensoes>
      <dimensao>
         <tabela>dimPaciente</tabela>
         <pk>PacienteId</pk>
         <fk>PacienteId</fk>
      </dimensao>
      <dimensao>
         <tabela>dimTempo</tabela>
         <pk>TempoId</pk>
         <fk>TempoId</fk>
      </dimensao>
      <dimensao>
         <tabela>dimServico</tabela>
         <pk>ServicoId</pk>
         <fk>ServicoId</fk>
      </dimensao>
      <dimensao>
         <tabela>dimConvenio</tabela>
         <pk>ConvenioId</pk>
         <fk>ConvenioId</fk>
      </dimensao>   </dimensoes>
   <metricas>
      <metrica>
         <nome>Valor da Fatura</nome>
         <coluna>ValorFatura</coluna>
      </metrica>
      <metrica>
         <nome>Quantidade</nome>
         <coluna>Contador</coluna>
      </metrica>
   </metricas>
 </fato>
</modelo>
```

Figura 39. Arquivo que representa o modelo da Figura 38

Conforme observamos na Figura 39, cada elemento representando uma tabela de fatos possui uma lista de dimensões associadas, assim como uma lista de métricas disponíveis. Sabendo, através do repositório de metadados, quais tabelas estão relacionadas aos termos escolhidos, podemos obter quais tabelas de fatos estão inseridas no contexto da consulta. Analisando estas, percorremos a lista de métricas disponíveis e a apresentamos ao usuário. Este procedimento traz benefícios de desempenho e torna a interface com o usuário muito mais amigável.

5.4 O protótipo desenvolvido

De maneira a implementar a abordagem apresentada nas seções anteriores, desenvolvemos um protótipo, instanciando a arquitetura idealizada, tendo como objetivo principal evidenciar que os mecanismos propostos permitem ao usuário uma exploração conjunta de um universo estruturado (DW) e um não-estruturado, cujos domínios de informação sejam os mesmos ou estejam fortemente relacionados. Nas seções a seguir apresentamos, para cada camada, o que foi construído.

5.4.1 Camada de Processamento

O Processador da Navegação foi desenvolvido na linguagem JAVA[20], utilizando a API JDOM[21] para a leitura dos arquivos XML. Esta leitura é realizada através do SAX[22], cujos algoritmos são disponíveis na API utilizada. Para cada elemento apresentado na seção 4.1.1 foi criada uma classe JAVA, sendo sua persistência realizada no repositório de metadados.

Para realizar o processamento dos documentos, foi criada uma classe em JAVA. Recebendo como parâmetro o diretório que servirá de repositório, o código realiza uma leitura seqüencial nos documentos presentes neste. Ao encontrar uma seqüência de caracteres entre dois espaçamentos, a mesma verifica no índice construído na memória se esta seqüência corresponde a algum termo relevante. Caso exista a correspondência, a ligação é registrada no repositório de metadados. Foram testados, com sucesso, arquivos das seguintes extensões: “.txt”, “.doc”, “.html”, “.xls”, “.pdf”, “.ppt”.

Foram criadas, para desenvolver o Construtor de Índice, algumas classes na linguagem JAVA, de acordo com o especificado na seção 5.1. Através de conexões OBDC[23], o construtor de índice acessa os representantes do domínio (dimensões do DW e recursos específicos), gerando o índice que será acessado pelo Processador de Documentos.

---

[20] http://java.sun.com/

[21] http://www.jdom.org/

[22] http://www.saxproject.org/

[23] http://msdn.microsoft.com/en-us/library/ms710252(VS.85).aspx

Ao acessar o termo representante do domínio, o mecanismo desenvolvido verifica se o mesmo tem como origem uma dimensão com correspondência ou não nas dimensões presentes no DW. Esta informação é então armazenada no repositório de metadados, sendo utilizada pela camada de análise.

5.4.2 Camada de Dados

Na camada de dados, o único elemento que deve ser construído, segundo a abordagem proposta, é o repositório de metadados. Para tal, foi criado um banco de dados em Access[24]. O metamodelo apresentado na seção 5.2, foi implementado conforme especificado, podendo ser visualizado na Figura 40. Por limitações de tempo, não foi possível criar o auto-relacionamento existente para a tabela de atributos. Isto irá afetar, conforme veremos na seção 5.4.3 na representação da hierarquia entre atributos, não sendo esta explicitamente apresentada ao usuário. As tabelas criadas possuem a seguinte estrutura:

- Dimensões
  - Id_dim: identificador único da dimensão.
  - Nome: nome, significativo ao usuário, da dimensão.
  - Tabela: nome da tabela do banco de dados.
  - Tipo: indica se a dimensão é interna (“I”), possuindo correspondência no DW, ou externa (“E”), tendo origem em um recurso específico de domínio.
- Termos
  - Id_termo: identificador único do termo.
  - Termo: termo encontrado.
- Documentos
  - Id_doc: identificador único do documento.
  - Arquivo: nome do arquivo armazenado que representa o documento.
- Atributos
  - Id_atributo: identificador único do atributo.

[24] http://office.microsoft.com/pt-br/access/default.aspx

- Fk_categoria: chave (estrangeira) da categoria.
- Nome: nome dado ao atributo.
- Desc_tp: indica se o atributo e é numérico ou textual.
- Desc_txt: coluna, na dimensão ou tabela de domínio, que representa o atributo, ou seja, que será lida e acessada pelos mecanismos.

- Categorias
  - Id_categoria: identificador único da categoria.
  - Categoria: nome da categoria.
  - Fk_faceta: chave (estrangeira) da faceta.
- Facetas
  - Id_faceta: identificador único da faceta.
  - Faceta: nome da faceta.
- Relacionamentos
  - Fk_termo: chave (estrangeira) do termo.
  - Fk_dimensao: chave (estrangeira) da dimensão.
  - Fk_atributo: chave (estrangeira) do atributo.
  - Fk_documento: chave (estrangeira) do documento.
  - Contador: número de vezes que o termo, qualificado pelo atributo e pela dimensão, aparece no documento.

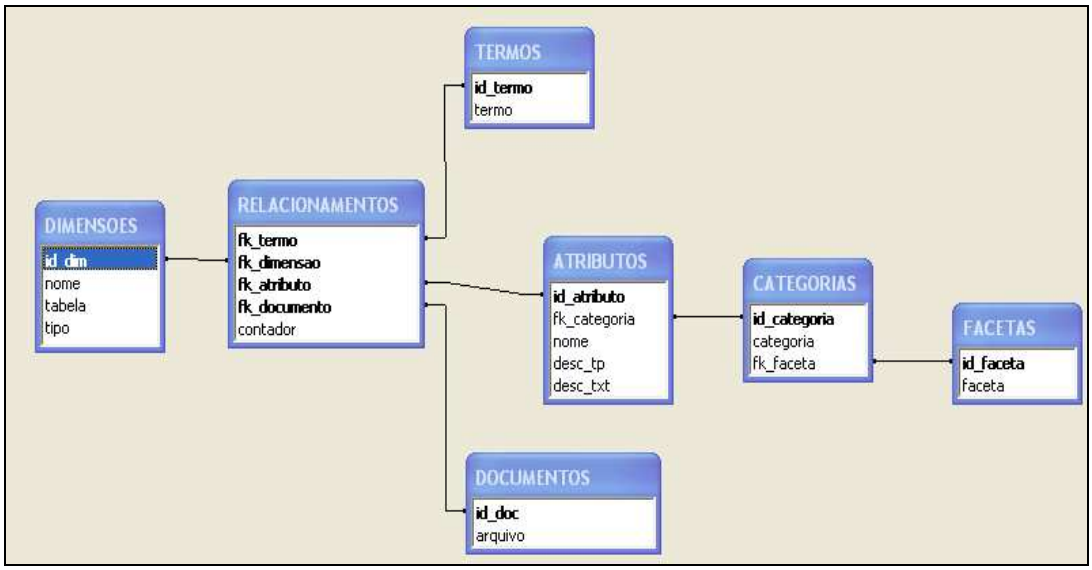

Figura 40. Repositório de metadados criado

5.4.3 Camada de Análise

Para apresentar a taxonomia facetada ao usuário, foi desenvolvida uma página em JSP[25] que realiza uma leitura no repositório de metadados e gera, para o navegador, o código HTML resultante do processamento.

Utilizando esta página HTML, o usuário pode navegar pela taxonomia. Após o usuário navegar na taxonomia facetada criada, escolher os termos e a opção de visualização, o Buscador de Documentos é acionado. Foram desenvolvidas classes em JAVA que recebem como parâmetro os termos escolhidos e, realizando uma leitura no repositório de metadados, retornam os documentos encontrados. São retornados todos os documentos que, em seu conteúdo, apresentam os termos selecionados pelo usuário, de acordo com a seguinte lógica: para termos pertencentes ao mesmo atributo, o documento que apresentar qualquer um destes é retornado, ou seja, é realizado um “or”; para termos pertencentes a diferentes atributos, somente serão retornados documentos que apresentem os dois, ou seja, é realizado um “and”. A interface desenvolvida não informa de maneira explícita esta composição, ou seja, em próximas versões do protótipo, seria muito interessante desenvolver mecanismos capazes de informar ao usuário como esta consulta foi criada, além de permitir que o mesmo faça suas próprias composições.

Utilizando como exemplo a seleção de termos da Figura 41, seriam retornados documentos que contém o termo “asma” e pelo menos uma ocorrência de “Amil” ou “Unimed”. A forma como o usuário irá visualizar as informações recuperadas vai diferir de acordo com a opção realizada no momento da busca. As opções, assim como as respectivas explicações, serão apresentadas na seção 5.5.

[25] http://java.sun.com/products/jsp/

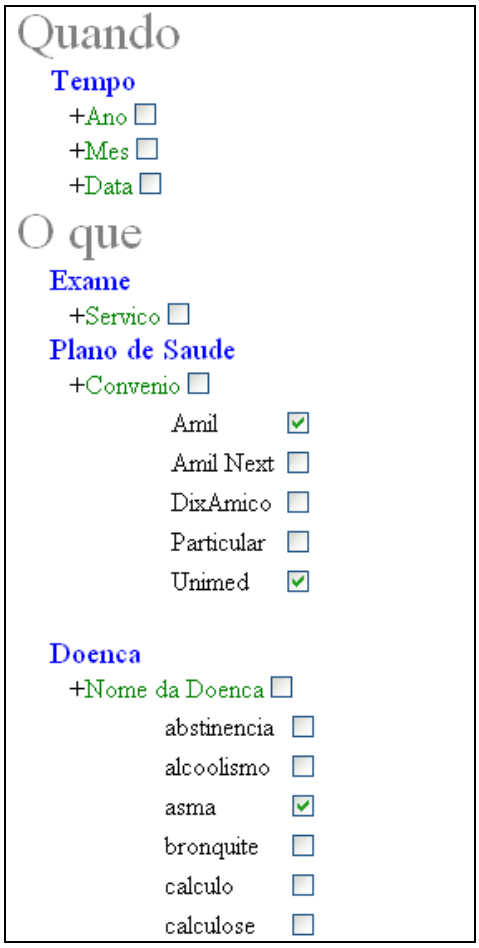

Figura 41. Exemplo de seleção de termos

Conforme mencionado anteriormente, uma hierarquia representa relações “pai e filho” entre os atributos de uma dimensão, tendo o elemento “pai” a função de agregar informações de seus respectivos filhos. Por limitações de tempo, os níveis de hierarquias entre atributos não estão sendo representados explicitamente pelo protótipo na taxonomia facetada. Podemos observar na Figura 41 que para a hierarquia de tempo os atributos “Ano” e “Mês” são apresentados no mesmo nível. Somente a relação entre faceta, categoria, atributo e termos está sendo representada hierarquicamente.

No conceito de navegação facetada apresentada por Hearst (2009), conforme o usuário vai selecionando as categorias desejadas, combinações de categorias não mais existentes na base vão sendo desabilitadas para escolha, impedindo que seja realizada uma busca que não retorne dados e restringindo o universo de exploração durante a navegação. Além disto, o número de documentos que se enquadra em cada categoria é exibido. Como exemplo de funcionamento de uma navegação facetada, podemos imaginar uma coleção onde não existam

documentos que se enquadrem nas categorias “Paciente” e “Fornecedor”. Durante a navegação sobre a taxonomia, ao escolher a categoria “Paciente”, a categoria “Fornecedor” ficaria indisponível para seleção. O sítio de vencedores do prêmio Nobel[26] apresenta a funcionalidade de maneira bem clara. Não temos em nosso protótipo esta funcionalidade, sendo necessárias algumas alterações no Gerador da Taxonomia, mecanismo responsável pela lógica de exibição da taxonomia, para implementá-la.

Caso o ambiente de trabalho possua uma ferramenta OLAP para acesso ao DW, o Construtor de Consulta pode ser desenvolvido através da adaptação desta, adicionando-se funcionalidades na mesma, de maneira a oferecer o caminho de análise exploratória proposto. Na solução criada neste trabalho, nenhuma ferramenta OLAP foi utilizada para acesso ao DW. Todo o Construtor de Consulta foi desenvolvido especificamente para mostrar que a idéia de ligação entre o universo estruturado e o não-estruturado funcionaria se colocada em prática. Assim, a consulta criada para acesso ao ambiente estruturado é submetida ao banco de dados e o resultado exibido em uma página Web, na qual foi adicionada a funcionalidade de seleção dos termos para posterior envio ao ambiente não-estruturado, conforme veremos na seção 5.5.

Em ferramentas OLAP, uma operação de *slice and dice* consiste em extrair subconjuntos do cubo, ou seja, limitar a análise por determinadas dimensões e determinados atributos de uma dimensão. No protótipo desenvolvido, através da seleção dos atributos das categorias na faceta, estamos fazendo cortes no modelo, tanto em uma dimensão como em mais de uma. Quando realizamos a seleção de termos, tanto no momento da navegação sobre a taxonomia facetada, quanto na seleção sobre os dados vindos do DW, estamos realizando cortes no domínio dos atributos.

Os filtros, ou seleções, são critérios adotados pelo usuário, no momento da exploração do universo de informações, para restringir o conteúdo a ser recuperado. Estes são de vital importância devido ao grande volume de dados tratado pelas ferramentas analíticas. No protótipo desenvolvido, ao navegar na taxonomia facetada, o usuário pode expandir os atributos desejados, sendo apresentados os termos pertencentes a este, disponibilizando-se a seleção dos mesmos. Da mesma maneira, na exibição das informações vindas do DW, caixas

[26] http://well-formed-data.net/experiments/elastic_lists/

de seleção são disponibilizadas ao lado de cada termo retornado. Ao selecionar os termos e submetê-los à análise, estamos realizando uma seleção, um filtro sobre o universo de informações disponível. Esta operação está disponível em todas as etapas da exploração conjunta para o protótipo criado.

5.5 Visualização das informações

A construção de mecanismos que permitam uma análise conjunta aos dados dos universos não-estruturados e estruturados vai além da construção de links entre estes dois ambientes. A maneira como os dados são apresentados ao usuário é um fator determinante no sucesso de qualquer abordagem de sistemas com interação entre o Homem e o mesmo. Segundo Judelman (2004), o grande desafio hoje não é necessariamente produzir novo conhecimento, mas desenvolver novas técnicas, ou aprimorar as existentes, melhorando a maneira como trabalhamos com este conhecimento, de modo a aproveitarmos ao máximo o que este pode nos oferecer. Conforme observamos no capítulo 2, alguns trabalhos cujo foco está na exploração conjunta entre os dois universos apresentam soluções próprias para a visualização da informação que são disponibilizadas.

Foram criadas, com a linguagem JSP, páginas para o protótipo desenvolvido. Estas são acessadas através de um servidor Tomcat[27], via HTTP, pela porta 8080. Entretanto, podemos utilizar outro servidor Web, tendo como única restrição que este ofereça suporte à linguagem referida (JSP). As subseções seguintes irão explicar com maiores detalhes como foi desenvolvido o processo de visualização das informações neste trabalho. Na subseção 5.5.1 temos a apresentação de como a interface foi projetada. A visualização da informação pode ser dividida de acordo com o universo que está sendo apresentado.

Desenvolvemos três maneiras de exibição das informações provenientes do universo não-estruturado,, cada uma tendo vantagens e desvantagens de acordo com o propósito da busca. As visualizações podem ter como origem a navegação sobre a taxonomia ou a análise das informações vindas do DW (universo estruturado). As diferentes maneiras de visualização do universo não-estruturado serão explicadas nas subseções 5.5.2 , 5.5.3 e 5.5.4 . A subseção

[27] http://tomcat.apache.org/

5.5.5 irá apresentar o mecanismo utilizado para a exibição das informações do universo estruturado, ou seja, vindas do DW.

5.5.1 Interface de visualização

No protótipo desenvolvido, uma interface (pertencente à camada de análise, descrita na seção 5.3) foi criada para apresentar as informações provenientes dos dois universos (estruturado e não-estruturado), de maneira a possibilitar uma exploração conjunta das mesmas. A interface é composta por três áreas: navegação, ação e resultados, representadas na Figura 42 pelas numerações I, II e III, respectivamente.

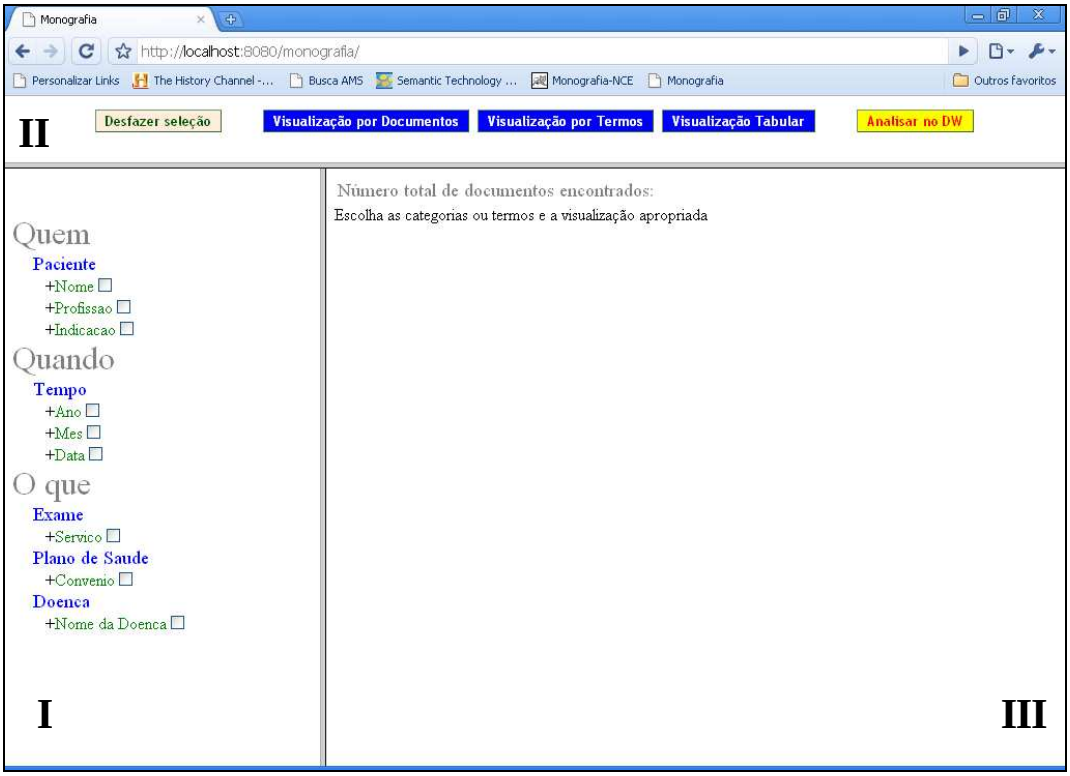

Figura 42. Interface de visualização

Na área de navegação (I) o usuário pode explorar a taxonomia facetada desenvolvida para atender ao universo de informação que está sendo explorado. É através da área de ação (II) que o usuário pode escolher o tipo de visualização dos documentos resultantes, assim como disparar a ação de análise, junto ao DW (universo estruturado), dos termos selecionados após a análise dos documentos. A área de resultados (III) apresenta os mesmos, de acordo

com a opção escolhida e os universos de origem e destino da consulta, propiciando também a escolha de termos para análise entre os dois universos.

5.5.2 Visualização por documentos

O módulo de visualização por documentos irá privilegiar o documento, ou seja, a partir deste são exibidas as informações encontradas. Conforme mencionado anteriormente, os fatos em uma análise no universo não-estruturado são os documentos encontrados para os atributos ou termos em questão. Através de uma exibição que privilegia o documento, estamos endossando esta teoria, apresentando os mesmos como resultados diretos da busca realizada.

Para cada documento, são exibidos os termos encontrados. Cada termo será agrupado e precedido na exibição pela categoria à qual pertence e pelo atributo ao qual foi mapeado. Ao lado de cada termo é apresentada ao usuário a ocorrência deste no documento e uma caixa de seleção. Através desta é possível a seleção do termo de maneira que este possa compor a consulta a ser enviada para o universo estruturado, no caso o DW.

A Figura 43 apresenta a tela de visualização por documentos. Observamos a taxonomia facetada à esquerda da imagem. Foram selecionados pelo usuário os atributos “Profissão”, “Indicação” e “Convênio”, tendo o mesmo clicado no botão “Visualização por Documentos”, na área de ação. Caso o usuário clique no sinal de “mais”, logo à esquerda do atributo, a lista dos termos disponíveis para este é apresentada. A Figura 44 mostra a área de navegação do protótipo tendo o usuário expandido o atributo “Convênio”. A busca no universo não-estruturado pode ser realizada através do atributo ou do termo, sendo que na primeira serão buscados todos os termos disponíveis neste. Na exibição dos resultados, primeiramente é informado ao usuário, na parte superior da tela, quantos documentos foram retornados utilizando-se os critérios selecionados. Na parte localizada à direita da imagem, temos então o resultado da consulta. O nome de cada documento encontrado é exibido. Através de um clique em cima do nome, o documento é aberto, em outra janela, ao usuário. Abaixo do nome de cada documento, temos as combinações das categorias encontradas neste, assim como seus respectivos atributos. Após cada combinação é então apresentada uma lista

dos termos, encontrados no documento, que foram classificados nesta. Ao lado de cada um observamos o número de ocorrências entre parênteses e uma caixa de seleção logo ao lado.

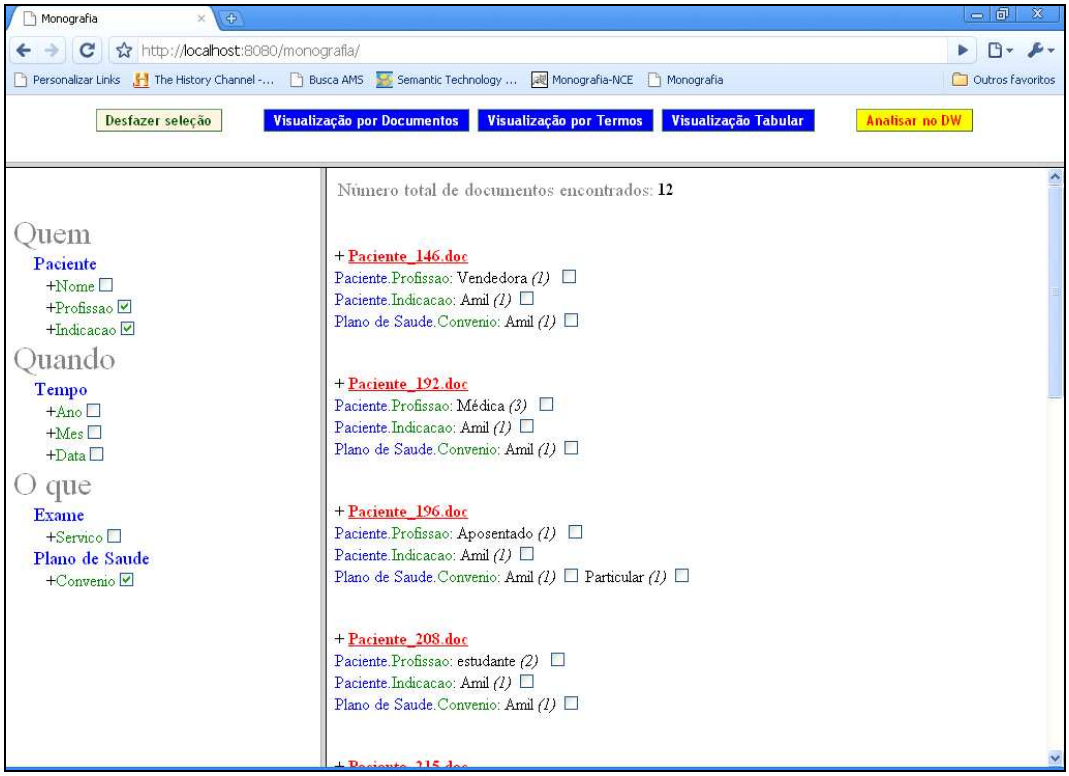

Figura 43. Visualização por documentos

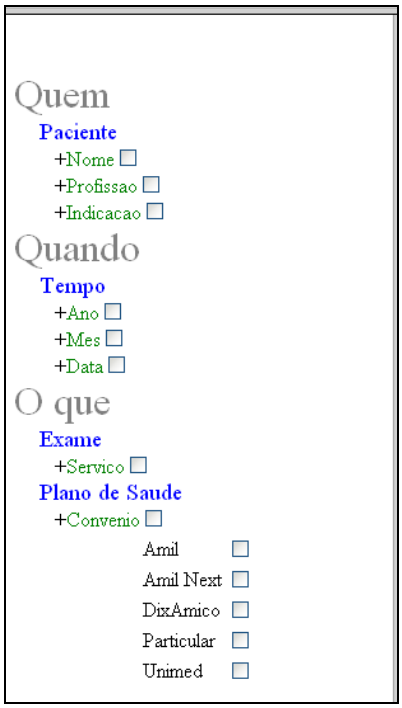

Figura 44. Expansão do atributo "Convênio"

O botão “Analisar no DW”, apresentado na área de ação, irá coletar os termos selecionados pelo usuário e disparar a consulta, que será criada pelo Construtor de Consultas, para o universo estruturado, levando à exibição dos resultados encontrados neste ambiente, conforme veremos na seção 5.5.5 . A consulta é construída seguindo o mesmo padrão da busca na coleção, ou seja, entre termos de um mesmo atributo é realizado um “and”; entre termos de diferentes atributos é realizado um “or”. Somente os termos apresentados na área de resultados são utilizados na criação da consulta, não sendo possível enviar diretamente os termos selecionados na taxonomia. Próximos desenvolvimentos podem explorar esta lacuna, agregando mais esta possibilidade ao processo exploratório.

5.5.3 Visualização por termos

O módulo de visualização por termos irá privilegiar os termos encontrados em cada documento. Uma lista das combinações de categorias e seus atributos será exibida como elemento de agrupamento dos termos. Dentro de cada combinação, os termos encontrados para estas serão exibidos. Ao lado de cada um, uma caixa de seleção é disponibilizada, de maneira que este possa ser incluído na consulta ao DW a ser gerada.

Cada termo encontrado irá então agrupar os documentos nos quais este está presente. Da mesma forma que na visualização por documentos, ao clicar em cima do documento, o arquivo do mesmo é exibido ao usuário. Entretanto, o número de ocorrências agora será apresentado ao lado do documento e não do termo, devido ao modo de agrupamento desta visualização. Logo abaixo do termo é exibida uma mensagem indicando em quantos documentos o termo foi encontrado. A Figura 45 apresenta a tela da mesma consulta realizada na Figura 43, tendo o usuário clicado no botão “Visualização por Termos”. Podemos observar que os mesmos dados foram retornados, só que sob uma ótica distinta.

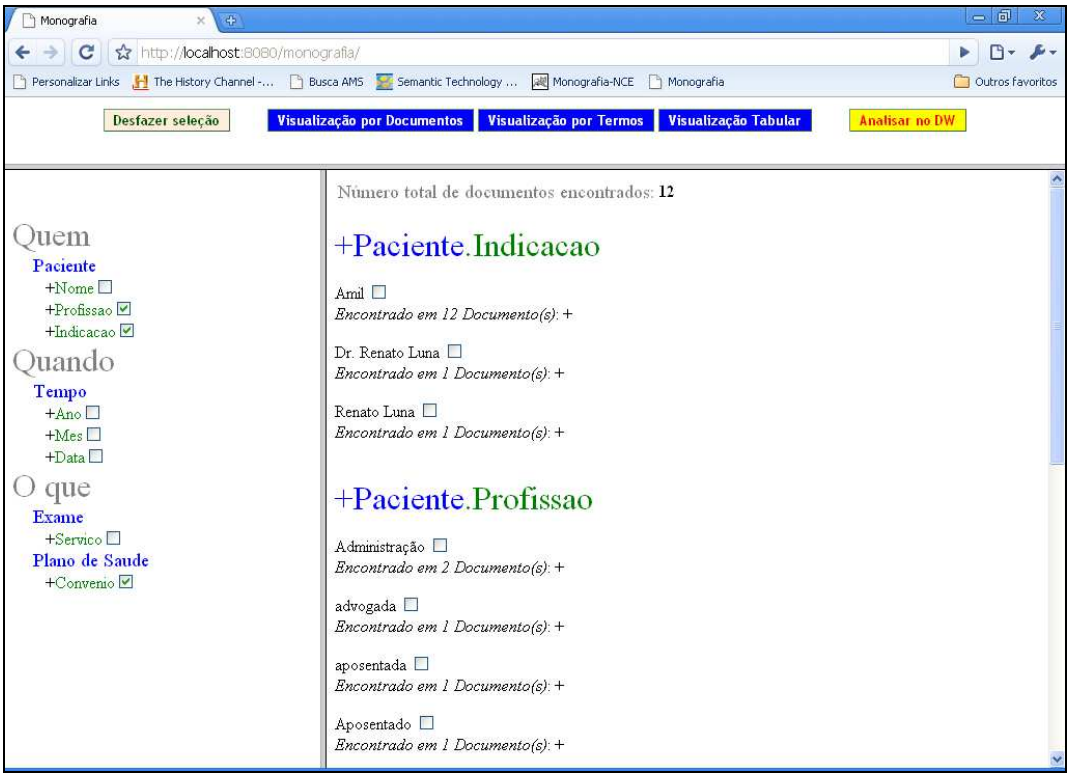

Figura 45. Visualização por termos

5.5.4 Visualização tabular

As duas formas de visualização apresentadas propiciam ao usuário que está realizando a análise uma maior facilidade de acesso à informação desejada, assim como ao próximo passo a ser dado. Ao obter o resultado do universo não-estruturado, o usuário pode ter como objetivo final um ou mais documentos que satisfaçam sua necessidade, ou, no caso que se aplica ao problema foco deste trabalho, pode necessitar o complemento da análise com dados presentes no universo estruturado (DW).

Cada visualização criada apresenta vantagens e desvantagens, conforme o objetivo do usuário. Ao realizar a apresentação do resultado pelos critérios estabelecidos, a visão por termos e a visão por documentos irão propiciar ao usuário perspectivas diferentes dos resultados obtidos: termos e documentos, respectivamente. Enquanto a primeira facilita as buscas cujo objetivo seja o enriquecimento das análises com as informações estruturadas, pois destaca o termo, objeto de criação da consulta ao DW, a segunda apresenta-se como melhor opção no caso da busca ter como objetivo os documentos propriamente ditos, uma vez que estes são o foco desta visualização.

Criada posteriormente, a “Visualização Tabular” tende a unir as vantagens apresentadas pelas visões anteriores. Ao dispor as informações na forma de uma tabela, esta visão permite ao usuário uma listagem direta dos documentos, assim como disponibiliza uma matriz de termos por categorias (e atributos) encontrados. Acreditamos que esta visão pode ser bem utilizada tanto por usuários que desejam somente obter os documentos que satisfaçam aos critérios estabelecidos, quanto aos usuários que necessitam enriquecer suas análises com as informações dos dois universos. A Figura 46 apresenta o resultado da mesma consulta realizada para as visualizações anteriores, tendo o usuário selecionado a “Visualização Tabular”.

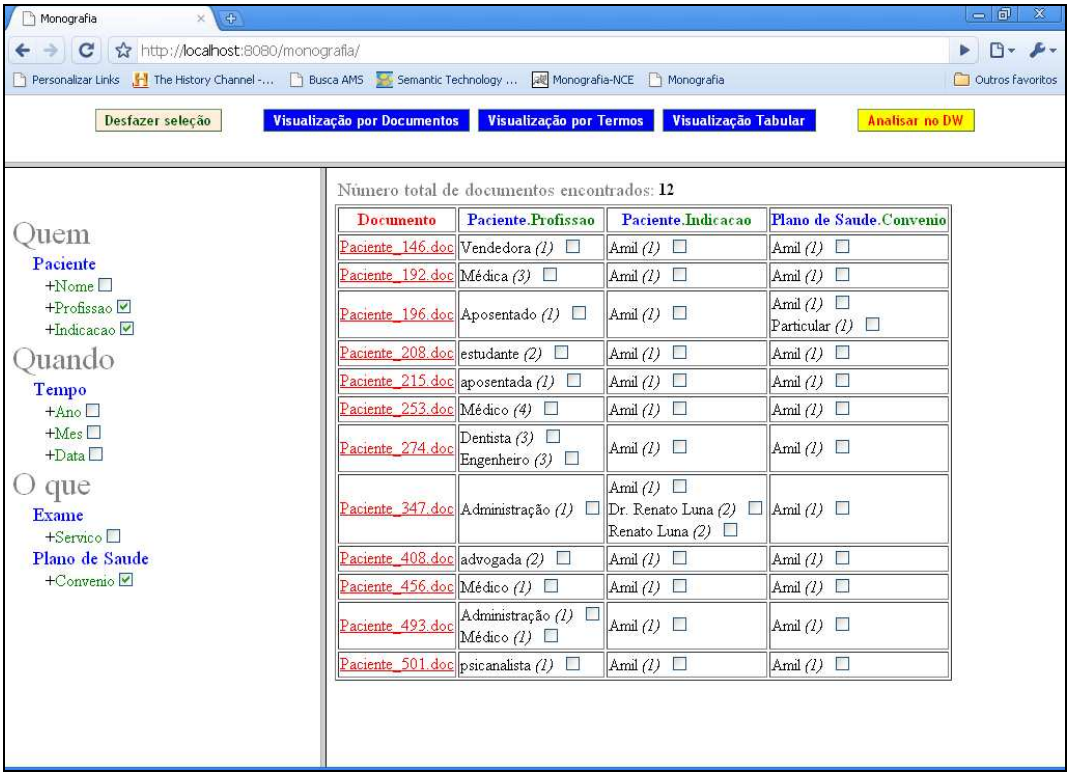

Figura 46. Visualização Tabular

5.5.5 Visualização dos dados do DW

Quando os termos do universo não-estruturado são submetidos ao enriquecimento da análise, através da agregação dos dados do universo estruturado, ou quando as informações a serem exibidas têm origem no DW (de natureza estruturada), as mesmas podem ser exibidas de acordo com a interface gráfica da ferramenta OLAP responsável pelo acesso à informação presente no DW. Atualmente as ferramentas disponíveis no mercado possuem diversos recursos para apresentação das informações, tanto em softwares de desktop, que rodam nas máquinas dos clientes, como em sofisticadas programações para páginas Web. Entretanto, conforme mencionado anteriormente, a apresentação dos dados provenientes do ambiente estruturado neste trabalho foi realizada através de programação própria. Nesta, os dados foram apresentados de forma tabular, de maneira a facilitar o entendimento da solução proposta e o caminho de navegação deste universo para o não-estruturado.

A Figura 47 apresenta o resultado de uma consulta realizada através da seleção de termos, resultantes da exploração do universo não-estruturado, que foram submetidos ao DW. A primeira informação exibida é uma listagem dos termos, com seus respectivos atributos, que deram origem à consulta. Na Figura 47, podemos observar que os termos de origem da consulta foram “Aposentado”, “Amil”, “Particular” e “Médico”. Logo abaixo da listagem, uma planilha é exibida contendo o resultado da consulta ao DW. Cada linha da planilha representa uma ocorrência na tabela de fatos, sendo exibido um somatório na última linha da mesma. Ao lado de cada termo, é disponibilizada uma caixa de seleção. Através destas é possível a realização da análise dos termos, retornados pelo DW, no universo não-estruturado. A Figura 48 apresenta a “Visualização Tabular” (sobre o universo não-estruturado) resultante da escolha dos termos “Médico” e “Amil” nos resultados do DW (universo estruturado).

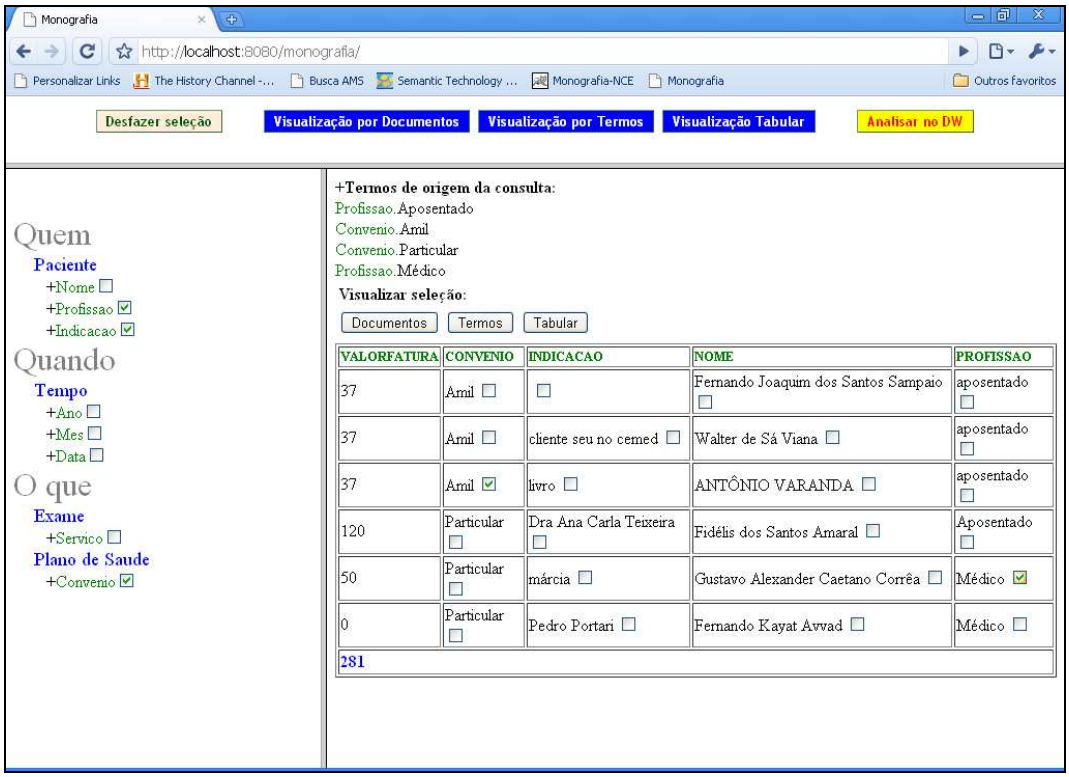

Figura 47. Visualização dos dados do DW

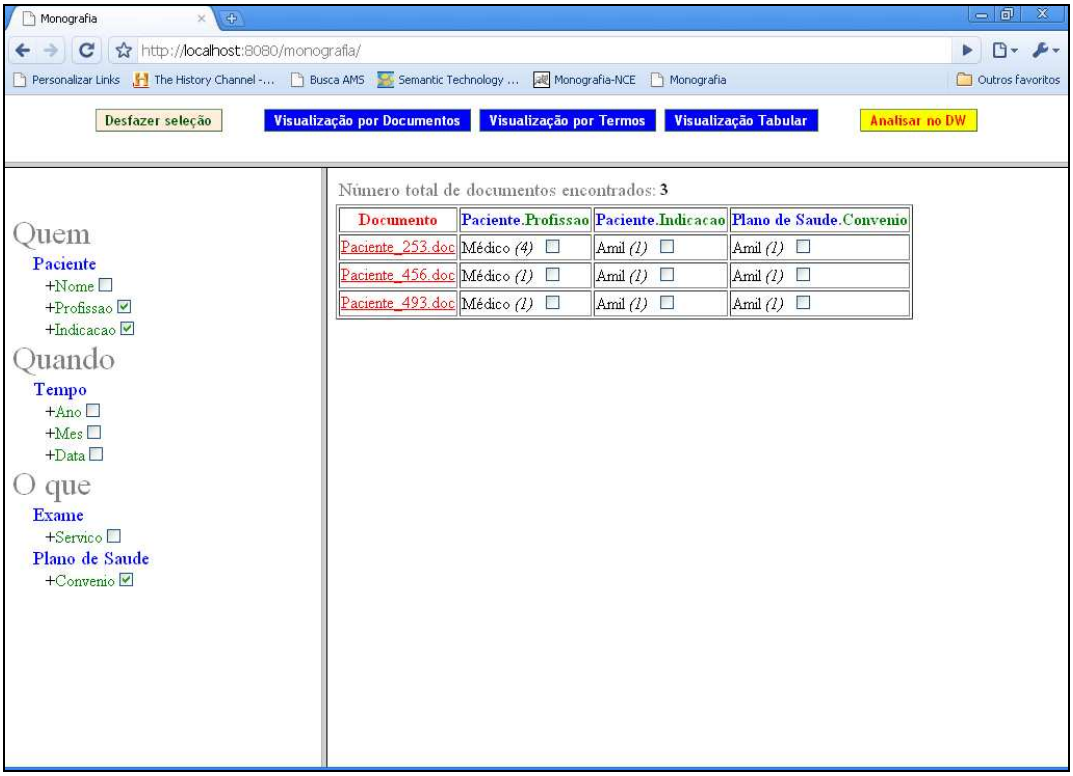

Figura 48. Resultado da seleção dos termos vindos do DW

5.6 Considerações Finais

A arquitetura apresentada ao longo deste capítulo teve como objetivo apoiar nossa idéia de exploração integrada, em ambientes analíticos, entre dados de natureza estruturada e de natureza não-estruturada. Através da construção do protótipo, implementando os mecanismos elaborados pela arquitetura, esta exploração pode ser realizada, conforme veremos em mais detalhes no capítulo 6. Tanto a arquitetura como principalmente o protótipo podem ser expandidos e melhorados em trabalhos futuros, de maneira a enriquecer o poder analítico da exploração oferecida.

A concepção em camadas da arquitetura visou tornar mais fácil este aprimoramento. Novos mecanismos podem ser concebidos e adicionados à camada correspondente sem detrimento dos demais. A visualização dos dados é um assunto que pode ser ainda muito explorado dentro da abordagem concebida. As três formas de visualização das informações não-estruturadas e, principalmente, a visualização dos dados estruturados podem ser aprimoradas. Desenvolvemos no protótipo, conforme mencionado, somente mecanismos e funcionalidades essenciais para a construção da ligação entre os universos. Seu aprimoramento não oferece ao desenvolvedor grandes dificuldades e seus mecanismos podem ser aproveitados em futuros desenvolvimentos.

6. Aplicação do protótipo e Comparativo

Após desenvolver os mecanismos apresentados pela arquitetura, construindo o protótipo, aplicamos o mesmo em um exemplo prático, de maneira a verificar que nosso objetivo (uma exploração conjunta de dados de natureza estruturada e não estruturada em ambientes analíticos) foi alcançado. A seção 6.1 apresenta esta aplicação, contendo dados sobre o processamento e um exemplo ilustrativo de exploração conjunta.

A abordagem proposta neste trabalho possui semelhanças e diferenças em relação aos trabalhos estudados e apresentados no capítulo 2. A principal semelhança, entre todas as propostas, está na tentativa de enriquecimento do ambiente analítico, tradicionalmente alimentado por dados vindos de universos estruturados, com dados provenientes de universos não-estruturados. A incorporação, entretanto, ocorre de maneira diferente entre os trabalhos. Além disto, a maneira como a análise é realmente enriquecida também irá diferir, de acordo com a proposta apresentada. Deste modo, analisamos na seção 6.2 as semelhanças e diferenças existentes entre a abordagem proposta neste trabalho e duas das principais soluções estudadas: o DW 2.0™ (seção 2.1) e o DoctorOlap (seção 2.2).

6.1 Aplicação do protótipo no domínio da medicina

Para a aplicação dos mecanismos propostos pela abordagem e desenvolvidos no protótipo, utilizamos como universo estruturado o Data Mart construído para armazenar os dados estruturados do DoctorOLAP (MOREIRA, CORDEIRO, CAMPOS, 2009). Foram utilizadas somente as tabelas cujo conteúdo é referente aos dados de origem estruturada, não fazendo parte do escopo dos testes, como componentes do universo, as tabelas contendo os dados de origem não-estruturada. Estes permanecem armazenados nos documentos da coleção. O modelo dimensional utilizado foi apresentado pela Figura 38. Como representantes do universo não-estruturado, foram utilizados documentos em formato de texto, criados a partir da extração dos campos de livre digitação do eDoctor (sistema de origem dos dados para o DoctorOLAP). Os documentos foram agrupados em um único diretório, totalizando 506 arquivos, servindo de base para o processamento necessário.

Antes de iniciar a execução dos mecanismos foi necessária a criação de uma taxonomia facetada para ser utilizada no processo de exploração conjunta. O processo teve início então na construção desta taxonomia, tendo como base o modelo dimensional. Para a construção desta, o primeiro passo foi a definição das facetas.

Utilizamos nesta aplicação as facetas “Onde”, “Como”, “O que”, “Quando” e “Quem”. A utilização destas facetas oferece ao usuário um agrupamento lógico das categorias criadas. Poderíamos ter optado por apresentar diretamente as categorias ao usuário. Entretanto, optamos pelo agrupamento por este oferecer uma maior organização do conteúdo. Após este primeiro passo, aplicamos o processo descrito na Seção 4.1 para obter a taxonomia, com suas respectivas relações e ligações.

Por se tratar de um DW Clínico, informações sobre as doenças em questão são muito importantes para as análises. Entretanto, não existe uma dimensão “doença” no modelo dimensional utilizado. Para suprir esta ausência foi utilizada uma tabela interna do sistema, seguindo as diretrizes estabelecidas na seção 4.2, contendo as doenças existentes para o domínio. Conforme mencionado anteriormente, esta adição torna as informações sobre doenças disponíveis para análises sobre o universo não-estruturado. Por não possuir uma correspondência direta no modelo dimensional do DW, o usuário não poderá selecionar uma doença e enviar a consulta para o universo estruturado. Para navegar entre os universos, neste caso, outras informações devem ser obtidas, como o paciente em questão, por exemplo.

O Anexo 4 contém os arquivos de configuração responsáveis pelas definições de facetas e categorias. Podemos visualizar o arquivo de configuração de dimensões no Anexo 5. O processamento destes arquivos de configuração deu origem a uma taxonomia facetada, que serviu de mecanismo exploratório, apresentada na Figura 49.

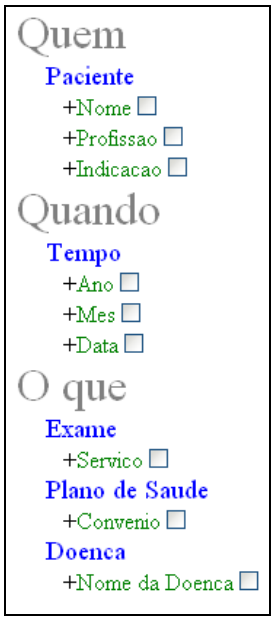

Figura 49. Taxonomia construída para aplicação dos mecanismos do protótipo

Após a construção da taxonomia facetada, a camada de processamento foi acionada. O tempo total de processamento foi de 39 minutos, tendo como resultado um índice contendo 1159 entradas de termos. A Tabela 2 mostra o número de registros gerados, para cada tabela do repositório de metadados, após a execução dos mecanismos.

Tabela 2. Nº de registros no repositório de metadados após execução dos mecanismos

TABELA	Nº DE REGISTROS
FACETAS	5
CATEGORIAS	7
ATRIBUTOS	9
TERMOS	1159
DIMENSOES	5
DOCUMENTOS	506
RELACIONAMENTOS	1957

Demonstraremos a seguir um exemplo de exploração conjunta do universo estruturado e do não-estruturado. O processo tem início através da navegação sobre a taxonomia facetada construída para o domínio. Podemos começar a exploração consultando quais documentos possuem os termos “2008”, “Unimed” e “Dix Amico” em seu corpo. Para tal, o atributo “Ano”, da categoria “Tempo” é expandido, sendo selecionado o termo 2008. Os termos “Unimed” e “Dix Amico” são selecionados a partir da expansão do atributo “Convênio”, da

categoria “Plano de Saúde”. A Figura 50 mostra os resultados obtidos a partir da ação disparada pelo clique no botão “Visualização Tabular”, presente na área de ação. Podemos observar que 12 prontuários foram retornados e que não existe, para o ano de 2008, nenhum documento que possua o termo “Dix Amico”.

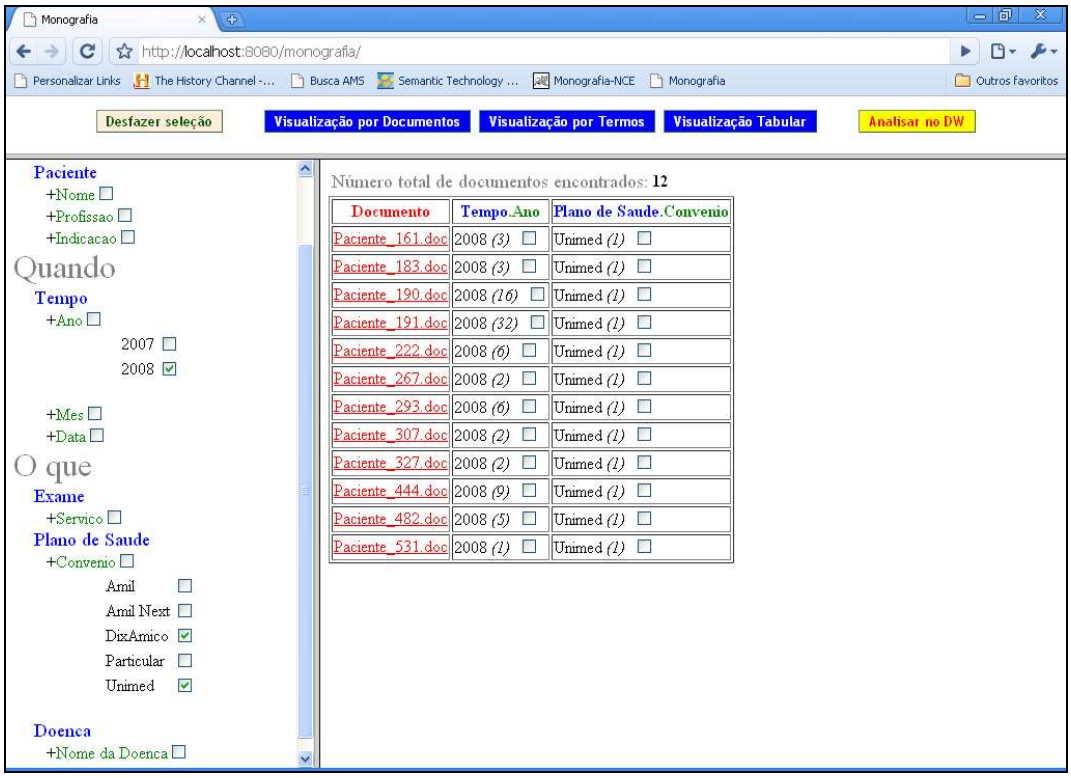

Figura 50. Documentos que possuem os termos "2008" e "Unimed" ou "Dix Amico"

Temos, na Figura 50, os resultados obtidos através da consulta realizada aos documentos. No universo sendo explorado, a chave de ligação entre os dois ambientes (estruturado e não-estruturado) é o paciente. Para saber quais são os pacientes envolvidos para as informações retornadas, selecionamos o atributo “Nome” e submetemos novamente a consulta com os demais parâmetros. O resultado pode ser visualizado na Figura 51.

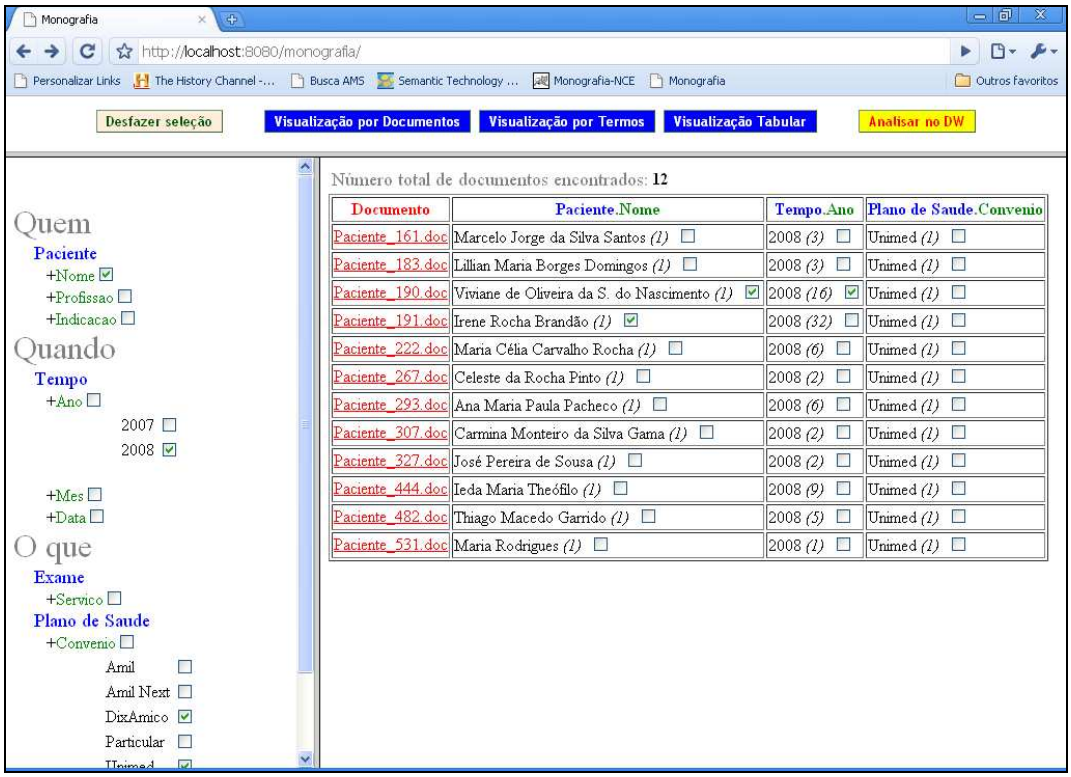

Figura 51. Adição do atributo "Nome" à consulta

Analisando os resultados da Figura 51, podemos observar que existem dois pacientes cujo termo “2008” aparece com maior freqüência em relação aos demais. Isto indica uma tendência que estes foram mais vezes na clinica médica que os demais. Podemos analisar os registros criados por estas idas no universo estruturado. Para tal, selecionamos os termos que representam os nomes dos pacientes e o termo “2008” na área de resultados, conforme a Figura 51. Ao clicar em “Analisar no DW”, na área de ação, devemos escolher as métricas que desejamos analisar. Uma lista das métricas disponíveis será apresentada conforme a Figura 52. Após a seleção da métrica “Valor da Fatura”, clicamos no botão “OK” e a consulta é enviada ao DW, obtendo como resultado o exibido na Figura 53.

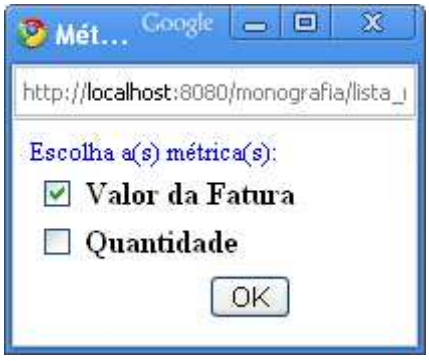

Figura 52. Métricas disponíveis, para análise no DW, de acordo com a seleção realizada

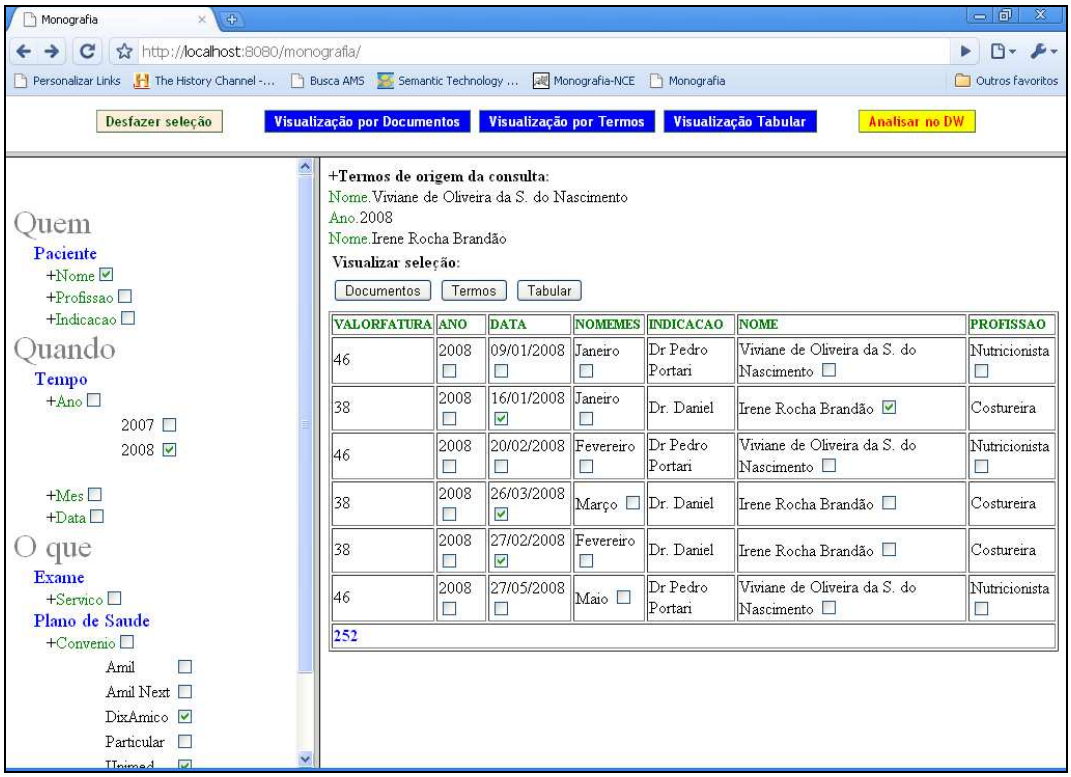

Figura 53. Informações vindas do universo estruturado para a seleção realizada

Neste momento, na área de resultados, observamos os dados vindos do universo estruturado, mais especificamente do modelo dimensional construído para o DoctorOLAP. No topo da área podemos visualizar os termos utilizados para a consulta ao DW. Estes termos funcionam como filtros no modelo dimensional. Os resultados retornados são todas as ocorrências da tabela de fatos que satisfazem os filtros dimensionais utilizados.

Podemos observar que as duas pacientes geraram faturas que totalizaram 252 unidades de moeda. A ausência da caixa de seleção ao lado de alguns termos representa que estes não possuem correspondência no universo não-estruturado, ou seja, apesar de estarem cadastrados no sistema, nenhum prontuário possui estes termos em seu corpo, o que impossibilita uma consulta ao universo não-estruturado através destes. Observamos que a paciente “Irene” aparece em três datas distintas. Podemos então querer visualizar estes prontuários. Para tal, selecionamos os termos que representam as datas e o termo que representa o nome da paciente. Apesar do termo aparecer mais de uma vez nos resultados, somente uma seleção do mesmo é necessária, ou seja, ao selecionar um termo, as caixas de seleção dos termos iguais são desabilitadas. A Figura 54 ilustra o resultado obtido.

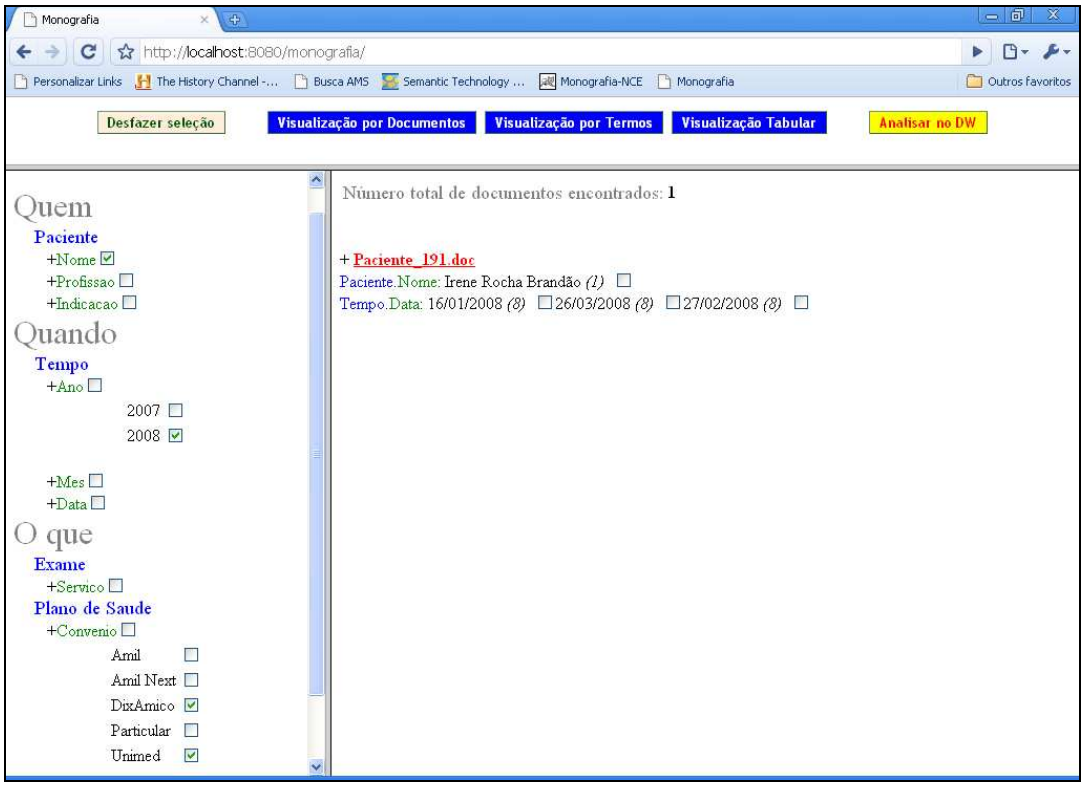

Figura 54. Prontuários obtidos pela seleção da Figura 53

Apesar de existirem três datas para a paciente em questão, somente um prontuário foi retornado. Isto ocorreu devido à lógica de geração dos documentos. Como foi gerado um documento para cada usuário, contendo todas as informações de livre digitação, todas as ocorrências fazem parte então de somente um documento. Caso fosse gerado um documento para cada data, seriam retornados três resultados. Podemos clicar no nome do prontuário (no caso da Figura 54 “Paciente_191.doc”), abrindo o documento em questão.

Através destes passos exploramos conjuntamente informações cuja origem era o universo não-estruturado e outras cuja origem era o universo estruturado. O protótipo desenvolvido, aplicado ao domínio de informação conseguiu alcançar o objetivo da exploração conjunta, embora seja passível de diversas melhorias a serem desenvolvidas em versões futuras.

6.2 Comparativo

Nesta seção, iremos apresentar os principais aspectos envolvidos na incorporação de dados de natureza não-estruturada, em ambientes analíticos, nos mecanismos desenvolvidos em nossa solução e nos mecanismos apresentados por duas das principais soluções estudadas: o DW2.0™ (ver seção 2.1) e o DoctorOLAP (ver seção 2.2). Procuramos apontar as semelhanças e principais diferenças existentes entre as soluções.

O DW 2.0™ permite que qualquer domínio de informação seja explorado. Para tal, o usuário deve realizar uma série de cadastros que descrevam o assunto a ser explorado. Este cadastro é realizado em parceria com um especialista do domínio, pois é neste que reside o conhecimento necessário para a classificação do conteúdo. Através da criação de um glossário, os termos relevantes para o assunto são identificados e submetidos a mecanismos que irão processar os documentos, cujo conteúdo consiste de dados de natureza não-estruturada. Também é permitido o cadastro de padrões, para dados de natureza semi-estruturada. Cada padrão cadastrado fará parte do processamento dos documentos. O DoctorOLAP foi desenvolvido para atender especificamente ao domínio de Data Warehouses Clínicos, sendo seu desenvolvimento direcionado para solucionar problemas neste tipo de universo.

A abordagem proposta por este trabalho não depende de nenhum domínio específico, sendo passível de aplicação em qualquer área que necessite da agregação de valor, propiciada pelas informações do universo não-estruturado, a um DW existente. As propostas e técnicas podem ser aplicadas integralmente a qualquer assunto de interesse.

De acordo com Inmon, Strauss e Neushloss (2008), modelos de dados representam um grande papel no universo estruturado, mas sua relevância em relação ao universo não-estruturado não é tão significativa. No DW 2.0™, os dados não-estruturados são lidos, processados e armazenados em um banco de dados relacional, para que possam ser objetos de análises por parte dos usuários. Foi desenvolvida, no DoctorOLAP, uma modelagem composta de dimensões e tabelas de fatos, criadas para receber o conteúdo oriundo do universo não-estruturado, de maneira a incorporar este ao ambiente analítico. Após realizar

etapas de pré-processamento, como eliminação de *stop words* e radicalização, foi desenvolvido um processo de ETC textual para carregar as informações obtidas no modelo criado.

Em nossa abordagem, desenvolvemos um repositório de metadados, baseado em um metamodelo, que irá conter as informações necessárias sobre o conteúdo não-estruturado e sobre o modelo dimensional que armazena as informações provenientes do universo estruturado. Nenhuma modificação no modelo dimensional existente é necessária. O repositório de metadados existe de maneira independente do modelo dimensional, podendo ser inclusive criado em um banco de dados diferente.

Podemos observar na Figura 55 as diferenças quanto à forma de armazenamento dos dados nas três soluções. No DW 2.0™ os dados vindos do universo estruturado são armazenados em um DW estruturado, enquanto os dados vindos do universo não-estruturado são armazenados em um DW não-estruturado. No DoctorOLAP os dois universos são carregados no mesmo modelo dimensional, estando entretanto em tabelas de fatos distintas. Em nossa abordagem, os dados de natureza não-estruturada não são carregados em um modelo, sendo somente carregadas informações sobre estes no repositório de metadados desenvolvido. Os dados de natureza estruturada permanecem no modelo dimensional existente.

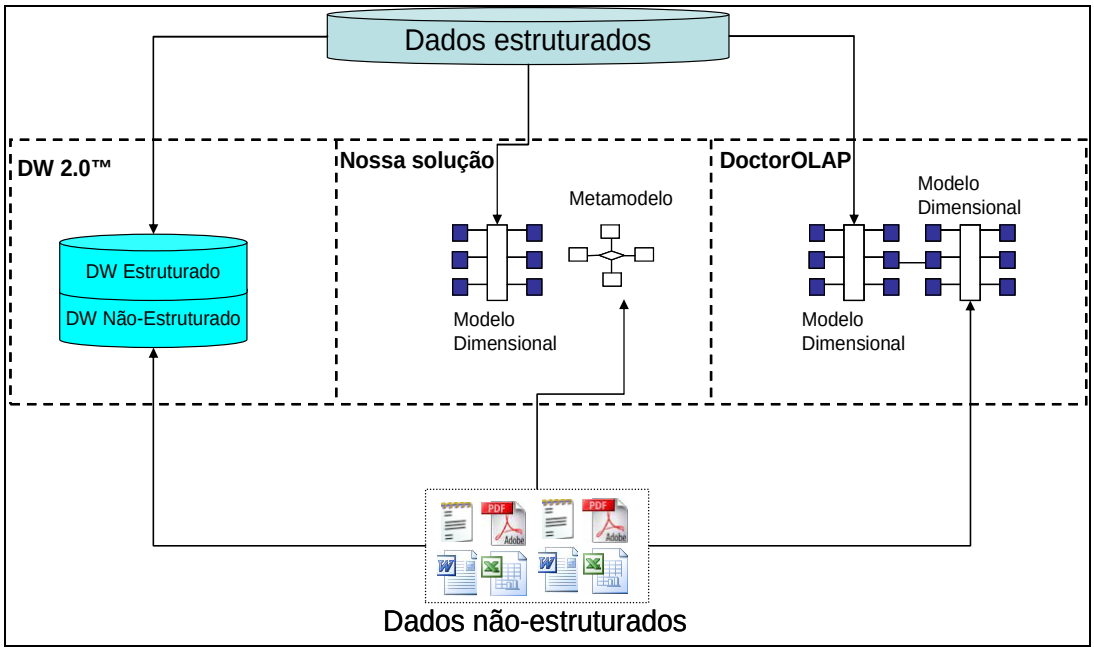

Figura 55. Diferenças entre o local de armazenamento dos dados

Através de algumas fases de pré-processamento de texto (descritas na seção 2.1), o DW 2.0™ elimina termos irrelevantes para o domínio em questão e cria uma lista de termos a serem utilizados para a exploração das informações não-estruturadas. No momento de processamento dos documentos, conforme mencionado anteriormente neste capítulo, uma tabela contendo os termos é criada. Além dos termos, outras informações como o documento no qual este aparece, e o local onde o mesmo se encontra no texto também são armazenados. Podemos considerar que esta tabela relacional de termos funciona como um indexador para os documentos.

No DoctorOLAP, foram construídas tabelas de apoio para a identificação dos termos relevantes e suas associações com categorias e facetas. A dimensão “Termos” contém os termos relevantes para o universo de clínicas médicas, obtidos através da leitura dos prontuários médicos. Foi utilizada a ferramenta *Foundation* (seção 2.2) para a obtenção destes. Cada termo foi então associado à categoria pertinente, e esta à respectiva faceta. Nesta etapa a equipe contou com a ajuda de especialistas no domínio. Podemos então considerar que a dimensão “Termo”, no DoctorOLAP, funciona como um indexador para os dados não-estruturados, sendo as demais informações associadas a esta através da tabela de fatos, tendo sido desenvolvido um processo de ETC textual para o carregamento destas. Assim como no DW 2.0™, esta tabela irá conter outras informações relevantes sobre o termo, como o documento, a localização do termo no mesmo e seu radical.

O conteúdo não-estruturado foi indexado primeiramente, em nossa abordagem, através das informações presentes nas dimensões do modelo dimensional, não sendo necessário associar cada termo em uma categoria da taxonomia facetada. Basta associar as categorias às dimensões existentes que, no processamento das mesmas, os termos presentes nestas serão automaticamente mapeados para a categoria em questão. Entretanto, caso algum termo relevante não esteja presente no modelo dimensional do DW, acabaríamos por perder esta informação caso a indexação ficasse restrita ao conteúdo dimensional. Conforme observamos nos capítulos 4 e 5, a abordagem proposta também prevê a incorporação de recursos específicos de domínio, contendo conhecimento prévio sobre o mesmo, que serão utilizados pelo processo de indexação para oferecer ao usuário um maior poder exploratório.

O DW 2.0™ vem evoluindo ao longo do tempo, enriquecendo sua abordagem e ampliando as funcionalidades disponíveis nas ferramentas que o apóiam. Entretanto, até o momento de seu estudo para a realização deste trabalho, maiores detalhes sobre a ligação entre os dois universos (estruturado e não-estruturado) não estava disponível publicamente, embora acreditemos que muito possa estar incorporado na ferramenta Forest Rim (INMON, 2008a). No DoctorOLAP, a ligação entre as informações oriundas dos dois universos (estruturado e não-estruturado) é realizada através da dimensão ”Paciente”. As informações não-estruturadas são obtidas através do processamento dos campos de livre digitação, sendo agrupadas e adicionadas a um arquivo para cada paciente. Cada documento criado irá conter o nome e uma identificação do mesmo, para que os dados, ao serem carregados no modelo dimensional, possam estar associados ao paciente em questão. As tabelas de fatos presentes no modelo dimensional estarão então interligadas através da dimensão ”Paciente”.

Podemos considerar que o DW 2.0™ apresenta uma abordagem de mais alto nível para exploração conjunta, em ambientes analíticos, de dados estruturados e não-estruturados. Conforme mencionado, sua abordagem vem sendo detalhada e refinada ao longo dos últimos anos, sendo apoiada pela constante evolução das ferramentas associadas. O DoctorOLAP procurou explorar esta interligação apresentada pelo DW 2,0™ em um domínio específico, de clínicas médicas, através de uma modelagem estendida que contempla os dois universos.

Na abordagem proposta por este trabalho, a ligação entre os dois universos é realizada através das informações contidas no repositório de metadados, a partir da execução dos mecanismos pertencentes à camada de processamento. Ao escolher um termo (ou atributo), o sistema criado é capaz de identificar, através da realização de uma leitura no repositório de metadados, de qual dimensão este faz parte. Além do repositório de metadados, é necessária, em um passo seguinte, a leitura do arquivo de mapeamento do DW, para a identificação da tabela de fatos que faz parte do domínio escolhido, assim como das ligações entre estas e as dimensões, estando esta leitura mais ligada a fatores físicos, como identificação da chave primária e chaves estrangeiras do modelo.

A Tabela 3 apresenta um resumo das comparações feitas entre as três soluções. As afirmações realizadas tomaram como base as informações obtidas até a conclusão deste trabalho.

Tabela 3. Resumo do comparativo entre soluções

Solução	Universo	Modelos de Dados	Índice (conteúdo não-estruturado)	Ponte	Visualização
DW 2.0™	Qualquer	Dados estruturados em modelo dimensional; Dados não-estruturados inicialmente carregados em tabelas relacionais	Tabela de termos	Abordagem ainda muito geral, sem entrar no detalhe.	Disponibilização das Informações para acesso via ferramenta OLAP e geração de SOMs.
DoctorOLAP	DW clínico	Dados estruturados e não-estruturados em um modelo dimensional	Dimensão "termos"	Dimensão "paciente"	Informações estruturadas e não-estruturadas através da ferramenta OLAP Dundas.
Nossa Solução	Qualquer	Dados estruturados em modelo dimensional; Metadados com informações sobre dados não-estruturados	Índice invertido	Repositório de metadados	Construção de um protótipo para permitir a exploração conjunta.

7. Conclusão

Dados de natureza não-estruturada representam a grande maioria das informações disponíveis atualmente. Ao não viabilizar que estes sejam analisados em sistemas analíticos, juntamente com os dados de natureza estruturada, estamos deixando de lado parte significativa do conhecimento de uma instituição. Usuários de sistemas gerenciais, que tratam e disponibilizam somente informações de natureza estruturada, podem ser prejudicados por esta ausência, podendo até mesmo tomar decisões incorretas que afetem a instituição.

Ao longo deste trabalho foram apresentadas técnicas, estudos e soluções objetivando o enriquecimento de ambientes analíticos, tradicionalmente caracterizados por possuírem dados de natureza estruturada, com dados de natureza não-estruturada. Estudamos, conforme o capítulo 2, algumas propostas de exploração conjunta entre estes dois universos, além de trabalhos relacionados ao uso de taxonomias, especialmente as facetadas, em sistemas de informação. Devido às semelhanças com as dimensões de um modelo dimensional, taxonomias facetadas foram escolhidas como mecanismo capaz de realizar a exploração, possibilitando ao usuário uma análise de dados das duas naturezas.

7.1 Principais Contribuições

Nosso trabalho teve como principais contribuições a criação de mecanismos, de domínio público, mais genéricos (baseados na leitura de arquivos XML), capazes de serem utilizados em cenários diversos e, principalmente, dar ênfase ao papel de taxonomias facetadas como mecanismo exploratório. Procuramos explorar sua correspondência com as dimensões de um DW, já habitualmente utilizadas em operações analíticas sobre dados de natureza estruturada.

A abordagem proposta, com a idealização da arquitetura para apoiá-la, mostrou ser capaz de oferecer a um usuário de sistemas analíticos uma exploração conjunta de informações vindas do mundo estruturado e do não-estruturado, possibilitando ao mesmo um caminho de mão-dupla em suas análises. Através da utilização de arquivos de configuração em XML e mecanismos de leitura do modelo dimensional, com a possibilidade de utilização

de recursos específicos de domínio, a ponte entre os dois universos foi construída. Nenhuma modificação no modelo dimensional é necessária para que a abordagem possa ser implementada, sendo este um fator positivo da mesma.

As informações geradas pelo processo são armazenadas em um repositório de metadados, desenvolvido para a solução proposta, servindo este de apoio para todos os passos da exploração. Prática comum nos trabalhos relacionados, não realizamos nenhum tipo de processamento dos textos (ETC Textual) para inseri-lo em modelos, ou seja, a solução procura realizar análises sobre o universo não-estruturado mantendo sua natureza original.

Podemos aplicar a solução apresentada a qualquer domínio, através da construção de uma taxonomia facetada específica que o represente. Escolhemos o domínio da medicina, conforme o capítulo 6, por possuirmos acesso os insumos necessários aos mecanismos da abordagem, ou seja, um modelo de dados dimensional e um conjunto de documentos correlatos. A taxonomia construída teve como base as dimensões presentes no modelo, sendo enriquecida posteriormente por uma tabela contendo o nome das doenças de relevância para o assunto. Conseguimos, então, evidenciar a capacidade exploratória de taxonomias facetadas em ambientes analíticos, servindo esta de ligação entre os universos (estruturado e não-estruturado), confirmando a hipótese principal do trabalho.

7.2 Limitações e Trabalhos Futuros

A incorporação de algumas técnicas de processamento de texto ao mecanismo exploratório ficou ausente da solução proposta, sendo esta uma questão importante, pois traria ao usuário uma maior flexibilidade. De maneira a dar continuação ao trabalho realizado, a adição destas técnicas torna-se um campo de muita riqueza para futuros trabalhos. Ao realizar a busca no universo não-estruturado, dado um termo selecionado no universo estruturado, por exemplo, o mecanismo poderia selecionar não somente documentos que contenham este termo, mas também que apresentem ocorrências de seus sinônimos. Estaríamos oferecendo ao usuário um ambiente analítico ainda mais representativo.

Ao retornar os documentos que satisfazem aos critérios estabelecidos pelo usuário da busca, os termos encontrados nestes são destacados, juntamente com a quantidade de vezes

que ocorrem nos mesmos. Desenvolvemos esta medida para oferecer ao usuário uma maior noção da relevância do documento em relação à busca realizada. Entretanto, o conceito de relevância pode ser mais bem explorado em trabalhos futuros, sendo necessário um maior estudo sobre o mesmo em sistemas de recuperação de informação, de maneira a incorporar novos conceitos ao mecanismo proposto por este trabalho.

O protótipo desenvolvido teve como objetivo evidenciar que as técnicas sugeridas, se aplicadas, dariam ao usuário a exploração conjunta objetivada. Vimos, no capítulo 6, que este atingiu seu objetivo, embora algumas funcionalidades, habitualmente encontradas em ambientes analíticos, tenham ficado de fora do desenvolvimento desta primeira versão. Grande parte destas lacunas foi devido ao fato de termos desenvolvido nossa própria interface com o DW. Fizemos esta escolha para facilitar a implementação dos mecanismos necessários à construção da ponte entre os dois universos. Entretanto, apesar de ser mais trabalhosa em um primeiro momento, a adição de uma ferramenta OLAP de mercado, com a realização das adaptações necessárias, supriria grande parte das lacunas mapeadas. Acreditamos que próximos trabalhos que tenham como interesse aprofundar o mecanismo exploratório proposto, poderão tirar proveito das funcionalidades disponíveis em ferramentas OLAP já existentes, com pequenas adaptações.

REFERÊNCIAS

BAEZA-YATES, R. ; RIBEIRO-NETO, B. **Modern information retrieval**. Harlow: Addison Wesley Longman, 1999.

BORDAWEKAR, R. ; LANG, C.A. Research articles and surveys: analytical processing of XML documents: opportunities and challenges. **ACM SIGMOD Record**, New York, v. 34, n. 2, p. 27-32, Jun. 2005.

BRODER, A. et al. Google Patents. **Efficient multifaceted search in Information Retrieval Systems.** US 7,496,568 B2, 26 fev. 2009. Disponível em: <http://www.google.com.br/patents?hl=pt-BR&lr=&vid=USPATAPP11564915&id=PDaqAAAAEBAJ&oi=fnd&dq=information+retrieval+faceted+taxonomies>. Acesso em: jun. 2009.

BROUGHTON, V. The need for a faceted classification as the basis of all methods of information retrieval. **Aslib Proceedings** London, v. 58, n. 1/2, p. 49-72, 2006.

CHAMPMAN, L. et al. **Knowledge discovery and data mining (KDDM) survey report**. Albuquerque: Sandia National Laboratories, 2005.(Sandia Report, SAND2005-1219)

Dijck, P.V. **XFML**. Disponível em: <http://www.xml.com/pub/a/2003/01/22/xfml.html>. Acesso em: mar. 2009.

DAGAN, I. ; CHURCH, K. ; Termight**:** Identifying and translating technical terminology. In: CONFERENCE ON APPLIED NATURAL LANGUAGE PROCESSING, 4. 1994. Stuttgart. **Proceedings …** , Stuttgart: Association for Computational Linguistics,1994.p. 34-40.

FURNAS et al. The vocabulary problem in human-system communication. **Communications of the ACM,** New York**,** 30, n. 11, p. 964-971, Nov. 1987.

GOLFARELLI, M. ; RIZZI, S. ; VRDOLJAK, B. Data warehouse design from XML sources. In: INTERNATIONAL WORKSHOP ON DATA WAREHOUSING AND OLAP, 4., 2002, Atlanta. **Proceedings …** , New York: ACM, 2001, p. 40-47.

HEARST, M. Integrating navigation with search. In: _____**Search user interfaces.** Cambridge: Cambridge University Press, 2009. chap. 8.

INMON, B. **Building DW 2.0**. Disponível em: <http://inmoncif.com/news/pdf/buildingFN.pdf>. 2007a. Acesso em: dez. 2007.

INMON, B. **Unstructured visualization**. Disponível em: <http://www.inmoncif.com/registration/whitepapers/ids%20visualization.pdf>. 2007b. Acesso em: dez.2009.

_____. **Unstructured applications:** unlocking the potential. Disponível em: <http://www.inmoncif.com/registration/whitepapers/unlockingunstructuredapps.pdf>. 2007c. Acesso em: dez. 2009.

_____. **Corporate information factory** Disponível em: <http://inmoncif.com/registration/news/dw2.php>. 2007. Acesso em: out. 2009.

_____. **A Solution for textual analytics**. Disponível em: < http://www.textual-etl.com/whitepaper/Solution_for_Textual_Analytics_Brochure_8-11-2008.pdf>. 2008. Acesso em: fev. 2010.

_____. **Textual ETL engine™ and foundation visualization**. Disponível em: < http://www.textual-etl.com/whitepaper/Overview_Textual_ETL_Engine_and_Foundation_Visualization_8-11-2008.pdf>. 2008b. Acesso em: fev. 2010.

INMON, B. ; NESAVICH, A. **Tapping into unstructured data:** integrating unstructured and textual analytics into business intelligence. Upper Saddle River: Prentice Hall, 2008.

INMON, B. ; STRAUSS, D. ; NEUSHLOSS, G. **DW 2.0:** the architecture for the next generation of data warehousing. Amsterdam: Elsevier, 2008.

INMON, W. H. **Building the data warehouse**. New York: John Wiley & Sons, 2005.

JUDELMAN, G. **Knowledge visualization:** problems and principles for mapping the knowledge space. MSc Thesis Dissertation (Master of Science in Digital Media) – International School o New Media University of Lübecke. Germany. 2004..

JUKIC, N. Modeling strategies and alternatives for data warehousing projects. **Communications of the ACM**, New York, v. 49, n. 4, p. 83-88, Apr. 2006.

KIMBALL, R. ; ROSS, M. **The data warehouse toolkit:** the complete guide to dimensional modeling. New York: John Wiley& Sons, 2002.

KIMBALL, R. et al. **The data warehouse lifecycle toolkit**. 2. ed. New York:John Wiley& Sons, 2008.

LEE, J. et al. Integrating structured data and text: a multi-dimensional approach. In: INTERNATIONAL CONFERENCE ON INFORMATION TECHNOLOGY:

CODING AND COMPUTING, 1., 2000, Las Vegas, **Proceedings …** Los Alamitos: IEEE, 2000. p. 264 – 269.

LI, Y. ; BELKIN, N.J. A faceted approach to conceptualizing tasks in information seeking. **Information Processing & Management**, Elmsford, v. 44, n. 6, p. 1822-1837, Nov. 2008.

MANNING, C. D. ; RAGHAVAN, P. ; SCHÜTZE, H. **Introduction to information retrieval**. Cambridge: Cambridge University Press, 2008.

MARCHIONINI G. Exploratory search: from finding to understanding. **Communications of the ACM**, New York, v. 49, n. 4, p. 41 - 46, Apr. 2006.

MOREIRA, J. ; CORDEIRO, K. ; CAMPOS, M.L. DoctorOLAP: ambiente para análise multifacetada de prontuários médicos. In: SIMPÓSIO BRASILEIRODE BANCO DE DADOS, 24., 2009, Fortaleza. **Anais ...**, Fortaleza: SBC, 2009.

MCCABE, M.C. et al. On the design and evaluation of a multi-dimensional approach to information retrieval. In: INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 23., 2000, Atenas. **Proceedings …** New York: ACM, 2000. p. 363-365.

McCALLUM, A. Information extraction: distilling structured data from unstructured text. **ACM Queue**, New York, v. 3, n.9, p. 48-57, Nov. 2005.

NASSIS, V. et al. A requirement engineering approach for designing XML-view driven, XML document warehouses. In: Annual international Computer Software and Applications Conference, 29., 2005, Eidinburg. **Proceedings …** Los Alamitos: IEEE, 2005. v. 1, p. 388-395.

PARK, B-K. ; HAN, H. ; SONG, I-Y. XML-OLAP: A Multidimensional Analysis Framework for XML Warehouses. In: TJOA, A. M. ; TRUJILLO, J. (Ed). **Data warehousing and knowledge discovery**. Berlin: Springer, 2005. p. 32-42, (Lecture Notes in Computer Science, v. 3589).

PÉREZ-MARTÍNEZ, J. M. **Contextualizing a data warehouse with documents**. 2007. Thesis (Doutorado) - Dpto. de Lenguajes y Sistemas Informáticos, Universitat Jaume I, Castellon, 2007

PÉREZ-MARTÍNEZ, J. M. et al. A relevance-extended multi-dimensional model for a data warehouse contextualized with documents. In: ACM INTERNATIONAL

WORKSHOP ON DATA WAREHOUSING AND OLAP , 8., 2005, Bremen, **Proceedings …** New York: ACM, 2005. p. 19-28.

PÉREZ-MARTÍNEZ, J. M. et al. IR and OLAP in XML document warehouses. In: LOZADA, D. E. ; FERNÁNDEZ-LUNA, J. M. (Ed.). **Advances in Information retrieval**, Berlin: Springer, 2005. p. 536-539. (Lecture in Notes in Computer Science, v. 3408).

RUSU, L. I. ; RAHAYU, W. ; TANIAR, D. Warehousing dynamic XML documents. In: TJOA, A. M. ; TRUJILLO, J. (Ed). **Data warehousing and knowledge discovery**. Berlin: Springer, 2006. p. 175-184, (Lecture Notes in Computer Science, v. 4081).

SACCO, G. Analysis and validation of information access through mono, multidimensional and dynamic taxonomies. In: LANSEN, H. L. et al (Ed). **Flexible query answering systems**. Berlin, Springer, 2006. p. 659-670. (Lecture Notes in Computer Science, v. 4027).

SANKAR K.; TALWAR V.; MITRA P. Web mining in soft computing framework : Relevance, state of the art and future directions**. IEEE Transactions on Neural Networks**, New York, v. 13, nº. 5, p. 1163-1177, Sept. 2002.

TUCKER, M. Dark matter of decision making. **Intelligent Enterprise Magazine**, Manhasset, v. 2, nº 13, p. 20-26, 1999.

TUNKELANG, D. Dynamic category sets: an approach for faceted search. In: SIGIR FACETED SEARCH WORKSHOP, Seattle, 2006. **Proceedings ...** New York: ACM 2006. Disponível em: < http://www.cs.cmu.edu/~quixote/DynamicCategorySets.pdf>. Acesso em: fevereiro de 2010.

TZITZIKAS, Y. ; ANALYTI, A. Faceted taxonomy-based information managemen**t**. In: INTERNATIONAL CONFERENCE ON DATABASE AND EXPERT SYSTEMS APPLICATIONS, 18., 2007, Regensburg. **Proceedings …** Linz: DEXA Society, 2007. p. 207-211.

TZITZIKAS, Y et al. An algebraic approach for specifying compound terms in faceted taxonomies. INFORMATION MODELLING AND KNOWLEDGE BASES, 15., EUROPEAN-JAPANESE CONFERENCE ON INFORMATION MODELLING AND KNOWLEDGE BASES, 13., Kytakyushu. 2003. **Proceedings …** . Amsterdam: IOS Press, 2004.

UDDIN, M. N. ; JANECEK, P. Faceted classification in web information architecture. **The Electronic Library**, v. 25, n.2, p. 219 - 233, 2007.

VIERA, A.F.G.; VIRGIL, J. Uma revisão dos algoritmos de radicalização em língua portuguesa. **Information Research**, Lund, v. 12, n. 3, Apr. 2007. paper 315, 7. Disponível em: <http://informationr.net/ir/12-3/paper315.html>. Acesso em: maio, 2009.

WATSON, H. J. ; WIXOM, B.H. The current state of business intelligence. **IEEE Computer** , Los Alamitos, v. 40, n. 9, p. 96-99, Sept. 2007.

WIWATWATTANA, N. et al. X^ 3: a cube operator for XML OLAP. In: INTERNATIONAL CONFERENCE ON DATA ENGINEERING, 23., 2007, Istanbul. **Proceedings …** Los Alamitos: IEEE, 2007. p. 916-925.

ZHOU, X. et al. Approaches to text mining for clinical medical records. In: ACM SYMPOSIUM ON APPLIED COMPUTING, 21., 2006, Dijon. **Proceedings …** New York, 2006. p. 235-239.

ZHOU, X. et al**.** Building clinical data warehouse for traditional chinese medicine knowledge discovery. In: 2008 INTERNATIONAL CONFERENCE ON BIOMEDICAL ENGINEERING AND INFORMATICS, 2008, Sanya, **Proceedings …** Washington: IEEE,2008. v.1, p. 615-620.

Anexo 1. Arquivo de Mapeamento para a dimensão da Figura 16

```
<dimensoes>
   <dimensao>
      <nomedim>Localidade</nomedim>
      <origem>I</origem>
      <tabela>Localidade</tabela>
      <categoria>
         <nomecat>Pais</nomecat>
         <atributo id='pais'>
            <nomeatrib>Pais</nomeatrib>
            <coluna>nome_pais</coluna>
            <tipo>1</tipo>
         </atributo>
      </categoria>
      <categoria>
         <nomecat>Estado</nomecat>
         <atributo id='uf' pai='pais'>
            <nomeatrib>Nome</nomeatrib>
            <coluna>nome_estado</coluna>
            <tipo>1</tipo>
         </atributo>
         <atributo id='sigla' pai='pais'>
            <nomeatrib>Sigla</nomeatrib>
            <coluna>sigla_estado</coluna>
            <tipo>1</tipo>
         </atributo>
      </categoria>
      <categoria>
         <nomecat>Cidade</nomecat>
         <atributo id='cidade' pai='uf'>
            <nomeatrib>Nome</nomeatrib>
            <coluna>nome_cidade</coluna>
            <tipo>1</tipo>
         </atributo>
         <atributo id='num'>
            <nomeatrib>Nº Habitantes</nomeatrib>
            <coluna>num_habitantes</coluna>
            <tipo>1</tipo>
         </atributo>
      </categoria>
   </dimensao>
</dimensões>
```

Anexo 2. Mapeamento completo para a Figura 22

```
<dimensoes>
   <dimensao>
      <nomedim>Localidade-Pais</nomedim>
      <origem>I</origem>
      <tabela>Pais</tabela>
      <categoria>
         <nomecat>Pais</nomecat>
         <atributo id='pais'>
            <nomeatrib>Pais</nomeatrib>
            <coluna>nome_pais</coluna>
            <tipo>1</tipo>
         </atributo>
      </categoria>
   </dimensao>
   <dimensao>
      <nomedim>Localidade-Estado</nomedim>
      <origem>I</origem>
      <tabela>Estado</tabela>
      <categoria>
         <nomecat>Estado</nomecat>
         <atributo id='uf'>
            <nomeatrib>Nome</nomeatrib>
            <coluna>nome_estado</coluna>
            <tipo>1</tipo>
         </atributo>
         <atributo id='sigla'>
            <nomeatrib>Sigla</nomeatrib>
            <coluna>sigla_estado</coluna>
            <tipo>1</tipo>
         </atributo>
      </categoria>
   </dimensao>
   <dimensao>
      <nomedim>Localidade-Cidade</nomedim>
      <origem>I</origem>
      <tabela>Cidade</tabela>
      <categoria>
         <nomecat>Cidade</nomecat>
         <atributo id='cidade'>
            <nomeatrib>Nome da Cidade</nomeatrib>
            <coluna>nome_cidade</coluna>
            <tipo>1</tipo>
         </atributo>
         <atributo id='num'>
            <nomeatrib>Nº Habitantes</nomeatrib>
            <coluna>num_habitantes</coluna>
            <tipo>1</tipo>
         </atributo>
      </categoria>
   </dimensao>
</dimensoes>
```

Anexo 3. Mapeamento para dimensão sucata da Figura 24

```
<dimensoes>
   <dimensao>
      <nomedim>Cartao de Credito</nomedim>
      <origem>I</origem>
      <tabela>Informacoes_Compra</tabela>
      <categoria>
         <nomecat>Cartão de Crédito</nomecat>
         <atributo id='tp'>
            <nomeatrib>
               Débito ou Crédito
            </nomeatrib>
            <coluna>tipo_cartao</coluna>
            <tipo>1</tipo>
         </atributo>
         <atributo id='bandeira'>
            <nomeatrib>
               Bandeira do Cartão
            </nomeatrib>
            <coluna>bandeira_cartao</coluna>
            <tipo>1</tipo>
         </atributo>
      </categoria>
   </dimensao>
   <dimensao>
      <nomedim>Endereco de Entrega</nomedim>
      <origem>I</origem>
      <tabela>Informacoes_Compra</tabela>
      <categoria>
         <nomecat>Endereço de Entrega</nomecat>
         <atributo id='end'>
            <nomeatrib>
               Endereço de Entrega
            </nomeatrib>
            <coluna>end_entrega</coluna>
            <tipo>1</tipo>
         </atributo>
      </categoria>
   </dimensao>
</dimensoes>
```

Anexo 4. Arquivo de configuração de facetas (A) e categorias (B)

```
<facetas>
    <faceta>
        <nome>Onde</nome>
    </faceta>
    <faceta>
        <nome>Quando</nome>
    </faceta>
    <faceta>
        <nome>Como</nome>
    </faceta>
    <faceta>
        <nome>O que</nome>
    </faceta>
    <faceta>
        <nome>Quem</nome>
    </faceta>
</facetas>
```

(A)

```
<categorias>
    <categoria>
        <nome>Paciente</nome>
        <faceta>Quem</faceta>
    </categoria>
    <categoria>
        <nome>Medico</nome>
        <faceta>Quem</faceta>
    </categoria>
    <categoria>
        <nome>Mes</nome>
        <faceta>Quando</faceta>
    </categoria>
    <categoria>
        <nome>Tempo</nome>
        <faceta>Quando</faceta>
    </categoria>
    <categoria>
        <nome>Exame</nome>
        <faceta>O que</faceta>
    </categoria>
    <categoria>
        <nome>Plano de Saude</nome>
        <faceta>O que</faceta>
    </categoria>
    <categoria>
        <nome>Doenca</nome>
        <faceta>O que</faceta>
    </categoria>
</categorias>
```

(B)

Anexo 5. Arquivo de configuração de dimensões

```
<!--
0 inteiro
1 string
-->
<dimensoes>
    <dimensao name='paciente' tipo='I'>
        <tabela>dimPaciente</tabela>
        <categoria>
            <nomecat>Paciente</nomecat>
                <atributo id='nome'>
                    <nomeatrib>Nome</nomeatrib>
                    <coluna>nome</coluna>
                    <tipo>1</tipo>
                </atributo>
                <atributo id='profissao'>
                    <nomeatrib>Profissao</nomeatrib>
                    <coluna>profissao</coluna>
                    <tipo>1</tipo>
                </atributo>
                <atributo id='indicacao'>
                    <nomeatrib>Indicacao</nomeatrib>
                    <coluna>indicacao</coluna>
                    <tipo>1</tipo>
                </atributo>
        </categoria>
    </dimensao>
    <dimensao name='ano' tipo='I'>
        <tabela>dimTempo</tabela>
        <categoria>
            <nomecat>Tempo</nomecat>
            <atributo id='ano'>
                <nomeatrib>Ano</nomeatrib>
                <coluna>ano</coluna>
                <tipo>0</tipo>
            </atributo>
            <atributo id='mes'>
                <nomeatrib>Mes</nomeatrib>
                <coluna>NomeMes</coluna>
                <tipo>1</tipo>
            </atributo>
            <atributo id='data'>
                <nomeatrib>Data</nomeatrib>
                <coluna>Data</coluna>
                <tipo>1</tipo>
            </atributo>
        </categoria>
    </dimensao>
    <dimensao name='exame' tipo='I'>
        <tabela>dimServico</tabela>
        <categoria>
            <nomecat>Exame</nomecat>
            <atributo id='servico'>
                <nomeatrib>Servico</nomeatrib>
                <coluna>servico</coluna>
                <tipo>1</tipo>
            </atributo>
        </categoria>
    </dimensao>
    <dimensao name='plano' tipo='I'>
        <tabela>dimConvenio</tabela>
        <categoria>
            <nomecat>Plano de Saude</nomecat>
            <atributo id='convenio'>
                <nomeatrib>Convenio</nomeatrib>
                <coluna>convenio</coluna>
                <tipo>1</tipo>
```

```
            </atributo>
        </categoria>
    </dimensao>
    <dimensao name='doenca' tipo='E'>
        <tabela>Doenca</tabela>
        <categoria>
            <nomecat>Doenca</nomecat>
            <atributo id='doenca'>
                <nomeatrib>Nome da Doenca</nomeatrib>
                <coluna>Termo</coluna>
                <tipo>1</tipo>
            </atributo>
        </categoria>
    </dimensao>
</dimensoes>
```