JointOLAP – Sistema de Informação para Exploração Conjunta de Dados Estruturados e Textuais: Um estudo de caso no setor elétrico

João Luiz Moreira[1], Kelli de Faria Cordeiro[1], Maria Luiza M. Campos[1]

[1]Programa de Pós-graduação em Informática
Departamento de Ciência da Computação/iNCE – Universidade Federal do Rio de Janeiro (UFRJ) Rio de Janeiro – RJ – Brasil

joao.moreira@ppgi.ufrj.br, kelli@ppgi.ufrj.br, mluiza@ppgi.ufrj.br

Abstract. *With the increasing generation of structured and unstructured data, internally and externally to the enterprises, approaches have been proposed to process and present this information, enriching their analytical capabilities. But few have addressed it in an integrated way, joint exploration and taking advantage from the analytical processing facilities offered by OLAP tools. Thus, this paper proposes JointOLAP, an integrated architecture for joint exploration of structure and unstructured data applied in a real scenario with the development of an analytical information system, called PowerOLAP. It allows the enrichment of the analyses of disturbances on the Brazilian integrated electrical system with their impact reflected in the news. Examples of analyses illustrate its analytical potential.*

Resumo. *Com a geração cada vez maior de dados, internos e externos às empresas, de natureza estruturada e não estruturada, abordagens para o tratamento e apresentação dessas informações, de modo que possam enriquecer seu potencial analítico, têm sido propostas. Porém poucas abordam, de forma integrada, a exploração conjunta dessas informações aproveitando as facilidades oferecidas pelas ferramentas de apoio ao processamento analítico (OLAP). Desta forma, este trabalho propõe o JointOLAP, uma arquitetura integrada para exploração conjunta dos dados de natureza estrutura e não estruturada aplicada em um cenário real com o desenvolvimento de um sistema de informação analítico, chamado PowerOLAP, o qual permite o enriquecimento das análises sobre as perturbações do sistema elétrico interligado brasileiro com a sua repercussão veiculada nas mídias. Exemplos de análises ilustram seu potencial analítico, com conclusões sobre o relato desta experiência.*

1. Introdução

A informação é reconhecida, cada vez mais, como um ativo de alto valor dentro das empresas e como tal tem sido tratada por meio do desenvolvimento de sistemas de informação que buscam a exploração do seu potencial para apoiar a tomada de decisão nessas empresas. A evolução das interfaces e as funcionalidades analíticas desses sistemas têm atraído mais investimentos nessa área. Aplicações analíticas do tipo *dashboard* com acesso via dispositivos móveis estão entre as mais requisitadas para o

nível estratégico. Essas aplicações utilizam diversas fontes de dados, internas ou externas à empresa, podendo ser estruturadas ou não estruturadas.

Os dados de natureza estruturada são normalmente encontrados em bancos de dados que armazenam informações sobre as transações e operações do negócio da empresa, e os de natureza não estruturada são encontrados em documentos com conteúdo escrito em linguagem natural, como relatórios, normas, padrões, chats e e-mails. Os dados de natureza não estruturada, de interesse das empresas, podem ter sido gerados fora dela, como nas mídias digitais e redes sociais. Como o volume de dados desta natureza tem crescido muito nos últimos anos, os mesmos estão sendo alvo de várias pesquisas sobre formas de explorar o seu potencial analítico [Russom 2007] e em conjunto com os dados de natureza estruturada [Nelson 2010]. Estes visam enriquecer os dados estruturados com as informações não estruturadas. Alguns desses trabalhos [Nesavich & Inmon 2007] [Park & Song 2011] [Costa, Souza, Times, & Benevenuto 2012] destacam a necessidade de uso de técnicas de Processamento de Linguagem Natural (PLN) na extração, limpeza e transformação da informação textual, para posteriormente serem integradas ao ambiente estruturado, mas a vinculação com os documentos de origem acaba não sendo explorada.

Assim, alguns desafios ainda são encontrados nessas abordagens, como uma solução para a exploração conjunta que permita a navegação entre os dados estruturados e não estruturados. Além disso, essa abordagem de integração pode tirar vantagem de mecanismos de tratamento semântico do conteúdo não estruturado, assim como da utilização de instrumentos terminologicos no momento da exploração analítica. A partir da necessidade real de uma instituição e combinando estratégias de propostas anteriores de integração de dados, este trabalho propõe uma abordagem integrada para exploração conjunta de dados estruturados e textuais, chamada JointOLAP, que foi aplicada em um cenário real do Operador Nacional do Sistema Elétrico (ONS).

O ONS é uma organização sem fins lucrativos com a responsabilidade de coordenar e controlar a operação dos sistemas de geração e transmissão do Sistema Interligado Nacional (SIN), buscando assegurar a segurança do suprimento de energia e a otimização econômica da sua comercialização no Brasil [ONS 2009]. Dentre as suas principais preocupações, estão as falhas causadas por perturbações no sistema elétrico e a sua repercussão nas mídias. Para isso, o ONS conta com dois sistemas de informação, um para apoiar as análises das perturbações, com informações de natureza estruturada, e outro para analisar sua imagem institucional, com informações de natureza não estruturada. Esses dois sistemas são isolados e não permitem análise conjunta dos seus dados. Diante desse cenário, esse trabalho aplicou a abordagem do JointOLAP e desenvolveu o PowerOLAP, um sistema analítico que permite a exploração conjunta das informações sobre as perturbações do SIN e sua repercussão nas mídias. Seu objetivo é possibilitar o enriquecimento semântico das análises através da ligação do atual sistema analítico – o cubo de perturbações – com as notícias do setor elétrico.

Algumas abordagens anteriores encontradas na literatura estão descritas na seção 2, seguidas do detalhamento da abordagem integrada proposta por este trabalho na seção 3, e sua aplicação com exemplos de análises conjuntas na seção 4. A seção 5 finaliza o trabalho com as conclusões.

2. Análise Conjunta de Dados de Natureza Estruturada e Não Estruturada

Atualmente, nas organizações, os dados estruturados são geralmente gerenciados de forma independente e distinta dos dados não estruturados, de natureza textual. É estimado que de todas as informações úteis ao negócio em uma organização, 31% delas são não estruturadas [Russom 2007]. Mesmo assim, ainda hoje, a quase totalidade dos ambientes de inteligência de negócio, apoiados por data warehouses (DW) corporativos ou por conjuntos de *Data Marts* (DM) interligados, se baseia apenas em dados estruturados, oriundos dos bancos de dados de nível operacional. Analisar e explorar esse dados de diferentes natureza, de forma conjunta, pode aumentar consideravelmente o potencial das aplicações analíticas oferecidas aos tomadores de decisão dessas organizações [Inmon, Strauss & Neushloss 2008].

Em geral, os dados estruturados passam por um processo de preparação dos dados, nomeado de *Extraction, Transformation and Loading* (ETL), onde os dados são recuperados das fontes, tratados com relação à limpeza e transformações necessárias (agregações, derivações, etc.) e usualmente organizados multidimensionalmente em DMs, com medidas mantidas em uma estrutura central, descritas segundo diferentes perspectivas (chamadas de dimensões). Esses DMs são a base para as análises e explorações feitas através de aplicações construídas em ferramentas OLAP (ferramentas de suporte ao processamento analítico). Por sua vez, dados não estruturados podem ser tratados com técnicas de Processamento de Linguagem Natural (PLN) para extrair, limpar e transformar esses dados, e onde o texto de origem passa por um processo de indexação de seus termos, sofrendo uma série de tratamentos específicos [Almeida & Silva 2009], dentre eles a marcação do texto, eliminação de pontuação e stop words (termos não úteis), resolução de sinônimos e homógrafos e radicalização. Uma vez tratadas, as informações extraídas podem ser disponibilizadas, vindo tanto a complementar DMs existentes ou mesmo alimentando DMs sobre dados não estruturados [Nesavich & Inmon 2007].

A partir dos *DMs* sobre os dados estruturados e não estruturados, é possível efetuar análises em seus domínios, respectivamente, através de uma ferramenta OLAP. Porém, a análise isolada em cada um deles pode exigir árduo trabalho manual para o cruzamento das informações. Para minorar tal problema é preciso estabelecer um mecanismo de ligação (*Linkage*) [Inmon, Strauss, & Neushloss 2008] entre os dois tipos de DMs. Essas ligações podem ser dinâmicas ou estáticas. No primeiro caso, a ligação é criada em "tempo real" por uma transação ou uma consulta executada nos universos separadamente e depois unindo os resultados. No estático, já existem os relacionamentos entre os universos, os quais são acessados diretamente pelas consultas. O caso mais comum é o uso de dimensões compartilhadas entre os DMs, provendo uma ponte de navegação [Moreira, Cordeiro, & Campos 2009]. Heuseler (2010) utiliza o conceito de facetas associadas às dimensões para a geração de ligações dinâmicas entre cubos com dados estruturados e cubos com referências ao conteúdo de documentos, permitindo a navegação bidirecional entre os dois tipos de universos. Uma faceta corresponde a uma perspectiva sobre determinado tema, agrupando termos que apresentam igual relacionamento com o assunto mais geral, através da aplicação de um princípio básico de divisão.

O uso conjunto dos dois tipos de dados também foi viabilizado no sistema SCORE [Roy et al., 2005], onde dados estruturados são recuperados via consultas SQL e palavras chave extraídas dessas consultas são utilizadas para acesso às informações associadas em documentos. Outras propostas interligaram cubos gerados a partir de conteúdos de documentos a cubos de dados estruturados. Os chamados R-Cubes [Perez et al. 2007] usam conteúdo textual extraído de documentos para contextualizar resultados de análises obtidos via consultas OLAP. Na direção inversa, o sistema EROCS [Chakaravarthy et al. 2006] encontra em um banco de dados elementos que podem estar associados a um documento de entrada, permitindo explorá-los para enriquecer as informações contindas no documento. No caso da *Total Business Intelligence Platform* [Park & Song 2011], encontram-se funcionalidades para tanto recuperar informações contextuais associadas a dados estruturados alvo de uma análise, quanto para complementar dados não estruturados extraídos de um documento com resultados de uma análise sobre os dados estruturados de um cubo. Mesmo neste caso, maiores facilidades de análíse dos dados não foram exploradas, especialmente no que concerne ao *front-end* analítico. A proposta neste artigo pretende contemplar algumas dessas facilidades, que serão descritas nas próximas seções.

3. JointOLAP – Sistema de Informação para Exploração Conjunta de Dados Estruturados e Textuais

O JointOLAP é um sistema de informação que permite a realização de análises conjuntas entre os dados de natureza estruturada e não estruturada por meio de interfaces amigáveis e gráficas com os usuários finais. Para isso, mecanismos de tratamento de dados são implementados de uma forma que as informações textuais possam agregar valor aos dados estruturados, enriquecendo as possibilidades de análises, mas que as análises possam partir de qualquer dos universos. Esses mecanismos são independentes de tecnologias e ferramentas, e contemplam, de forma integrada, o tratamento dos dados e, principalmente, a ligação entre eles.

3.1. Arquitetura Integrada Proposta

A arquitetura proposta é composta, basicamente, por três camadas, ilustrada na Figura 1. As duas camadas iniciais possuem elementos fundamentais necessários para a construção da terceira camada, na qual os dados são apresentados e explorados. Esta é composta por uma base de informações analíticas e por interfaces de acesso, formando o sistema de informações analíticas JointOLAP. As camadas iniciais são divididas por elementos do ambiente de dados de natureza estruturada e por elementos do ambiente de dados de natureza não estruturada, que se integram para formar o ambiente para análise conjunta desses dados. Um componente fundamental para este propósito, chama-se *Linkage*, que faz parte do fluxo de tratamento das informações do JointOLAP, responsável por integrar os dados de diferentes naturezas.

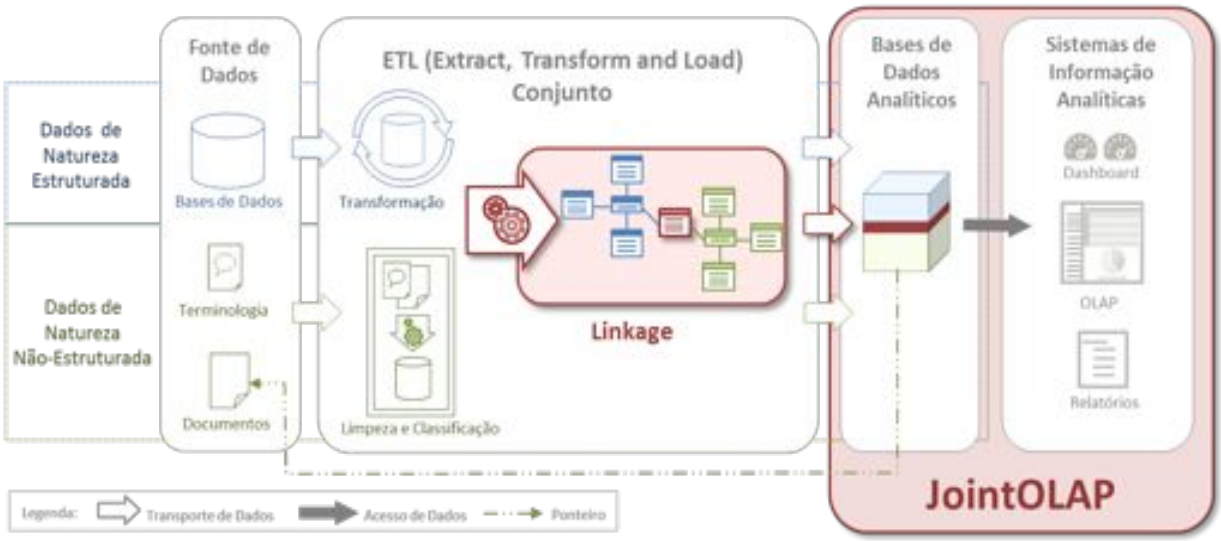

Figura 1 – Arquitetura Integrada do JointOLAP

Os dados de natureza estruturada são encontrados em bases de dados que armazenam, em sua maioria, alvos das transações ou operações de uma instituição. Essas bases, normalmente relacionais, são alimentadas por sistemas que automatizam seus processos de negócio. Já os dados de natureza não estruturada estão localizados em documentos em linguagem natural, normalmente em arquivos com formato doc, txt, pdf, dentre outros. Esses documentos armazenam informações sobre o negócio da instituição, como normas, padrões e notícias veiculadas nas mídias. Outro tipo de fonte de dados do JointOLAP é a terminologia empregada na geração dos documentos escritos em linguagem natural. Essa terminologia pode ser encontrada, por exemplo, em glossários de termos, taxonomias e ontologias do domínio, e exercem um papel fundamental no processamento do texto dos documentos.

O ETL conjunto é dividido, inicialmente, em dois processos distintos, o convencional e o textual. Os dados gerados por esses dois processos são ligados pelo processo *Linkage*. No processo convencional, sobre os dados estruturados, são executadas transformações comuns, como limpeza, transformação e integração dos dados, assim como a carga das dimensões, suas hierarquias e os fatos. O resultado desse processo é o banco de dados analítico sobre os dados estruturados. No processo textual, chamado ETL textual, outros tipos de transformações são executados, como as descritas na seção 2. Depois disso, os termos são categorizados em facetas, que podem ser enriquecidas com conceitos encontrados em instrumentos terminológicos, como o glossário de termos do domínio. Os termos, categorias e facetas podem, também, passar por um processo de aprimoramento da sua expressividade semântica, fazendo uso de ontologias, as quais buscam assegurar a aplicação dos termos de forma comprometida com a sua semântica. A preocupação com o correto uso e intepretação das informações deve ser uma constante no JointOLAP, representando o principal risco do projeto. Esse tratamento irá facilitar a navegação sobre as informações textuais na camada de apresentação. O resultado do ETL Textual é o banco de dados analítico sobre os dados não estruturados, agora estruturado em um banco relacional. No processo *Linkage*, para a construção da ligação entre as bases analíticas geradas, é necessária uma análise prévia sobre as interseções entre os dados. Primeiro são levantadas as entidades que

compõem o modelo multidimensional convencional e depois são listadas as entidades contidas nos documentos. Em seguida, é feito um mapeamento entre as entidades que são comuns em ambos ambientes. Essas entidades irão compor dimensões que serão compartilhadas entre as duas bases. Na maioria das vezes a dimensão Tempo é uma dessas dimensões. Neste caso, serão aplicadas operações para o relacionamento de intervalos de tempo entre os fatos, já aprofundada na literatura [Malinowski & Zimányi 2011].

O resultado do ETL conjunto é uma base com dados organizados em um modelo multidimensional, conforme o metamodelo ilustrado na Figura 2. Os fatos são ligados entre si por dimensões em comum, o que chamamos de Dimensões *Linkage*. Essas dimensões permitem a associação e navegação entre os fatos de ambos universos de dados. A ocorrência dos termos nos documentos são fatos analisados pelas perspectivas dos documentos e das suas categorias e facetas. Estas são modeladas com a aplicação da técnica de modelagem multidimensional, chamada *snowflake* [Kimball & Ross 2002], na qual hierarquias são criadas a partir das dimensões, eliminando-se redundância de dados e criando-se perspectivas de análises agrupadas em diferentes níveis. Em consequência, a imagem do modelo se parece com um floco de neve.

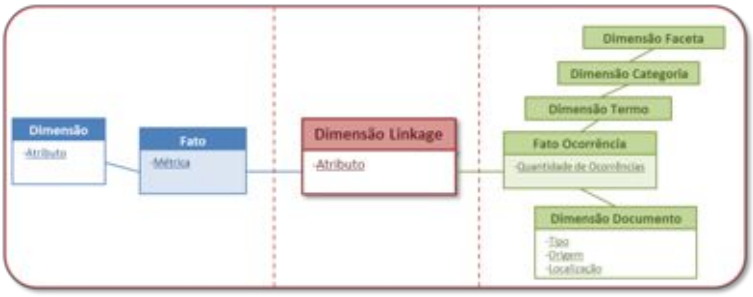

Figura 2 – Metamodelo Multidimensional do JointOLAP

A abordagem proposta neste trabalho possui um sistema de informação analítica, chamado JointOLAP, dotado de funcionalidades que permitem a combinação de dados de diferentes naturezas e fontes, sua agregação em diferentes níveis, e sua análise por diferentes perspectivas, sendo uma delas o Tempo. Pela dimensão Tempo é possível analisar o comportamento de um determinado fato ao longo dos anos e, a partir disso, tomar decisões para potencializar ou inibir aquele fato. Além disso, com as informações textuais disponíveis, essas análises são enriquecidas permitindo a sua correlação com fatos que não foram originalmente descritos e representados de forma estruturada. Detalhes dessas funcionalidades e do funcionamento da exploração conjunta são apresentados no exemplo de aplicação descrito na seção 4. As informações textuais podem ser, também, o ponto de partida de análises. As funcionalidades analíticas do JointOLAP podem ser aplicadas, inicialmente, nas informações oriundas de textos, para depois serem investigadas, com maior detalhe, nas informações oriundas de fontes de dados estruturados. Com isso, o JointOLAP permite não apenas a correlação dos fatos registrados de formas distintas mas também a navegação entre esses dois ambientes.

Além das funções analíticas apresentadas, o sistema de informação do JointOLAP é dotado de funcionalidades mais simples de acesso aos dados, como as encontradas em ferramentas de relatórios, que permitem listagens ordenadas e filtradas

de acordo com a necessidade do usuário final. Outro tipo de ferramenta analítica do JointOLAP, também de simples operação, porém com alto potencial analítico, são os *dashboards* que permitem a visualização dos indicadores calculados com dados da base de informação analítica.

4. Exemplo de Aplicação

O exemplo de aplicação foi realizado no âmbito do Operador Nacional do Sistema Elétrico (ONS), responsável por assegurar a segurança do suprimento de energia e a otimização econômica da sua comercialização [ONS 2009]. Dentre as suas principais preocupações, estão as falhas causadas por perturbações no sistema elétrico e a sua repercussão na vida dos usuários, refletida nas mídias.

4.1. Perturbações no Sistema Elétrico e as Notícias do Setor – Contexto do ONS

O SIN é o conjunto de instalações (unidades geradoras, linhas de transmissão, subestações, etc.) responsáveis pelo suprimento de energia elétrica de 97% do território brasileiro. As atividades para cumprir a coordenação e o controle da operação do SIN são baseadas em regras, critérios e procedimentos técnicos definidos nos Procedimentos de Rede. As perturbações elétricas podem ser causadas por descargas elétricas atmosféricas, inundações, queimadas e falhas humanas. No Glossário de termos técnicos, a perturbação elétrica é definida como: "Ocorrência no SIN caracterizada pelo desligamento forçado de um ou mais de seus componentes, que acarretam quaisquer das seguintes consequências: corte de carga, desligamento de outros componentes do sistema, danos em equipamentos ou violação de limites operativos." [ONS 2009].

A atividade de registro de perturbações é suportada pela utilização de equipamentos nas subestações dos agentes de transmissão. Tais registros são usados na triagem das ocorrências e perturbações, que os categoriza em anormalidades, eventos indesejáveis e desempenhos insatisfatórios. O Sistema Integrado de Perturbações (SIPER) e o Sistema de Apuração da Transmissão (SATRA) são os Sistemas de Informação que apoiam esse processo. Posteriormente essas informações são analisadas com o objetivo de avaliar o comportamento da rede durante as perturbações e apontar soluções para os problemas encontrados.

Para tratar os aspectos analíticos sobre o tema, o ONS possui um Sistema de Informação Analítico baseado em *Data Warehousing*, chamado DM de Perturbações, que apresenta as informações consolidadas dos sistemas que apoiam o processo de registro e classificação e as informações provenientes da Base de Dados Técnicas do setor elétrico. A integração é feita através de um ETL convencional sobre os dados estruturados, sendo disponibilizados no DM em questão.

O resultado de uma perturbação para o sistema pode levar ao corte de suprimento de energia de uma área geográfica, sendo conhecido popularmente como "apagão" (blecaute). As consequências negativas de um blecaute para a população são inúmeras, gerando grande prejuízo financeiro em todos os setores da economia. Por essa razão, a imprensa brasileira dá grande foco ao tema, muitas vezes citando o ONS quando tal situação ocorre, podendo impactar em sua imagem institucional. Por isso, a organização precisa de instrumentos para tratar o tema.

A Imagem Institucional é a forma como a organização é percebida pela sociedade, tendendo a ser classificada como positiva ou negativa, variando de intensidade e dependendo de variáveis como as oportunidades, as ameaças e a sua competência [Cardoso & Polidoro 2011]. A imagem de uma empresa tem origem externa, a mente do público, enquanto a sua identidade é construída a partir de suas políticas e processos.

Diariamente o ONS disponibiliza um resumo de todas as notícias referentes ao setor elétrico na *home page* da sua Intranet, nomeadas de *Clippings*. São três matérias consideradas mais importantes na semana com o link para o acesso à sinopse delas, mostrando quando o ONS é citado. O principal objetivo de ter tal Sistema de Informação é apresentar para os colaboradores da empresa o que está sendo publicado sobre o setor elétrico, evidenciando todas as vezes que a organização é mencionada na imprensa brasileira.

Por causa da natureza do trabalho do ONS a organização é notada pela população, na maioria das vezes, quando algum problema ocorre, como os "apagões" gerados por perturbações elétricas no SIN. Atualmente os Sistemas de Informação apresentam as informações de forma isolada e para uma análise conjunta das informações é necessário grande trabalho manual, muitas vezes impossibilitando o alcance dos resultados desejados.

4.2. Sistema para Análise Conjunta das Perturbações e Notícias com o JointOLAP

Foi idealizado um DM sobre os *clippings*, no qual é possível analisar a terminologia utilizada nas notícias de acordo com o momento de sua publicação. Para extrair, transformar e carregar as informações contidas nos *Clippings* foi utilizado, primeiramente, um ETL Textual e posteriormente um ETL convencional. Na primeira parte desse processo foram executadas as atividades de tratamento textual, gerando uma base de dados transacional dos *Clippings*, contendo registros dos documentos processados, termos, sinônimos, *stop words*, entre outros.

A segunda parte do processo é um ETL convencional, que utiliza a base transacional gerada pela primeira etapa como fonte de dados. Algumas transformações são executadas para o agrupamento dos dados e a carga deles na estrutura modelada com esquema estrela no DM de *clippings*, contendo as dimensões de arquivo (a notícia), a data de publicação, os termos existentes e o veículo de imprensa responsável. Para executar a experimentação da arquitetura integrada proposta nesse trabalho, foi utilizado como universo estruturado as ocorrências de perturbações elétricas no SIN, disponibilizado através do DM de Perturbações. Como universo não estruturado foram utilizadas as notícias do setor elétrico, disponibilizado através do DM de *Clippings*.

4.2.1 *Linkage*

A ligação entre os universos foi concebida através da observação do histórico das notícias, onde foi percebido grande volume de publicações citando a organização quando perturbações ocorrem no setor elétrico. Uma notícia relacionada à ocorrência de uma perturbação geralmente é publicada no mesmo dia ou alguns dias depois que ela aconteceu. Embora não exista a garantia de um relacionamento direto entre a data de publicação de uma notícia e a data de ocorrência de uma perturbação, o uso das

dimensões temporais como *Linkage* traz resultados consistentes [Costa, Souza, Times, & Benevenuto 2012], portanto esse tipo de *Linkage* pode ser caracterizado como forte.

No DM de Perturbações existe a dimensão de fim da perturbação, a data de fim da ocorrência. O DM de *Clippings* apresenta a dimensão temporal de data da publicação da notícia. Baseado nessas duas dimensões, uma tabela-ponte para tratar relacionamentos m para n foi construída (Figura 3) com o objetivo de ligá-las, contendo o relacionamento entre as chaves primárias das dimensões, isto é, um *Linkage* estático. Para apoiar tal implementação foi utilizado o tipo de sincronização de relacionamento temporal "meet" [Malinowski & Zimányi 2011]. Nesse tipo de relacionamento dois valores temporais se encontram em um determinado instante, isto é, quando um termina o outro começa.

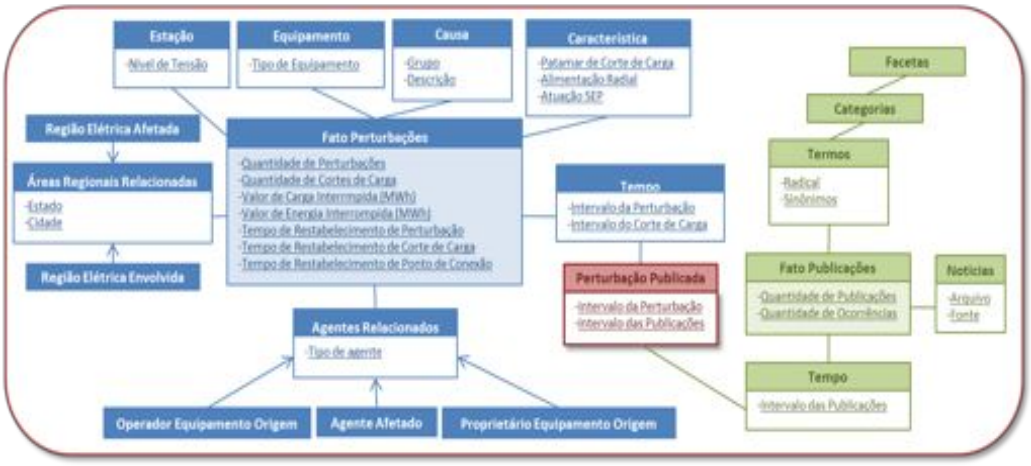

Figura 3 – Modelo Muldimensional do PowerOLAP

A partir da integração física entre os DMs foi possível o cruzamento das informações das dimensões e das medidas dos fatos e através da perspectiva temporal, as análises sobre as perturbações do setor elétrico e sua influência na mídia brasileira puderam ser feitas, possibilitando que o usuário navegue até os documentos (*clippings*) quando necessário.

4.2.2 PowerOLAP – Sistema de Exploração Conjunta das Perturbações e *Clippings*

Algumas análises representativas no domínio da segurança do suprimento de energia e seu impacto à imagem institucional do ONS puderam ser executadas com base nesse trabalho. A seguir, serão apresentados alguns exemplos dessas análises.

Utilizando as informações provenientes do universo estruturado, o usuário, ao analisar o tempo de reestabelecimento médio das perturbações com patamar de carga acima de 99MW nos meses de outubro e novembro de 2009, pôde perceber um aumento considerável de cerca de 1000% do primeiro para o segundo (Figura 4 – Navegação 1). A repercussão de tal fato nas notícias do setor elétrico pôde ser analisada ao cruzar o resultado com os termos e suas ocorrências nos *clippings* onde grande parte das publicações citam os termos "apagão" e “erro de operação” no mês de novembro, citando a organização por diversos momentos. Como o tratamento semântico foi realizado, o resultado da consulta contempla o termo apagão e seus sinônimos. É possível observar a quantidade de vezes que o conceito “apagão” foi citado nas notícias

no dia seguinte à perturbação e nos 15 dias seguintes, e que no quarto dia, a notícia não tem mais tanto destaque (Figura 4 – Navegação 2). A partir disso, os gestores podem se organizar para atuação intensa junto à impressa, com conhecimento sobre as falhas, nos 4 dias subsequentes à mesma. E com a navegação até pela fonte das notícias, é possível saber quais jornais mais falam sobre o assunto, incluindo o acesso até o documento da reportagem (Figura 4 – Navegação 3).

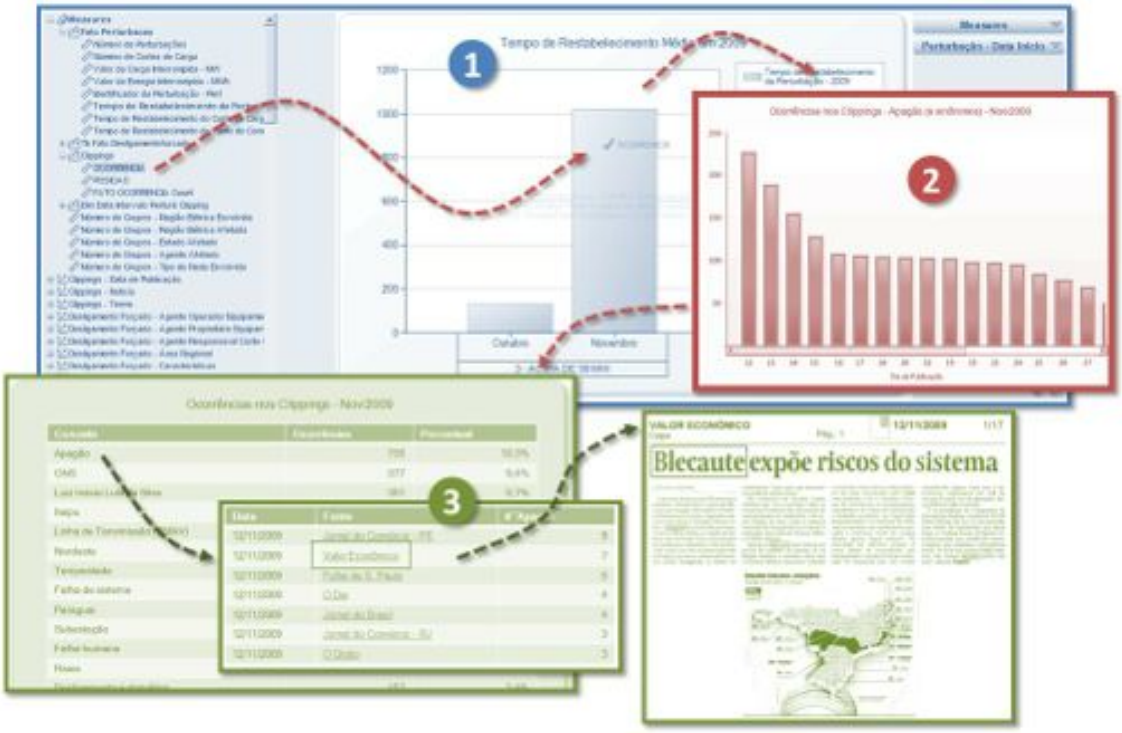

Figura 4 – Exploração conjunta por meio de navegação no PowerOLAP

Foi descoberto também que em junho de 2009 todas as vezes que o conceito "apagão" foi citado, o ONS também foi. Tal resultado foi obtido através da comparação entre a quantidade de publicações com o conceito "apagão" e a quantidade de publicações com conceitos "apagão" e "ONS", ambas representadas como medidas do fato de ocorrências de termos nas publicações. Para os tomadores de decisões da organização esse tipo de análise pode direcionar a necessidade de trabalho no marketing externo com a sociedade, para desassociar a marca ONS do termo "apagão" (e seus sinônimos).

Em outro exemplo de análise conjunta, ilustrada na Figura 5(a), foi verificada a relação das causas das perturbações com o conceito "erro de operação" através do cruzamento da dimensão de causa da perturbação e a medida de ocorrências do conceito "erro de operação" nas notícias publicadas. Foi descoberto que quando ocorrem citações sobre esse conceito nas notícias cerca de 40% das causas das perturbações são as "Falhas Humanas", seguidas das "Operações de Proteção, Medida e Controle" (32%), "Corpos Estranhos e Objetos" (18%), "Fiação AC-DC" (7%) e "Outros" (3%). A partir desse resultado podemos concluir que somente em 72% desses casos a imprensa empregou o termo "erro de operação" de forma concisa às reais causas das perturbações.

Outras análises interessantes puderam ser respondidas, como quantas vezes o ONS é citado quando ocorrem perturbações graves por falhas humanas; o quão normal a imprensa costuma distinguir erro causado pelo agente e erro causado pela operação do setor; quais são os termos mais comuns nas notícias do setor quando ocorre um desligamento forçado causando danos em equipamentos; e em quais regiões do país a imprensa publica mais notícias quando uma perturbação grave ocorre. Além de indicadores para comparar mensalmente as perturbações e as publicações – Figura 5(b).

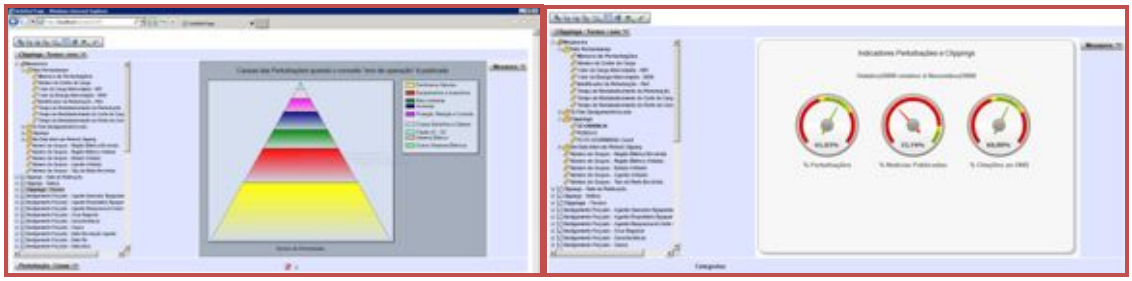

Figura 5 – Dashboards: (a) Causas das perturbações com o conceito "erro de operação". (b) Análise conjunta das perturbações e notícias relativas a um mês

5. Conclusões

A grande disponibilidade de informações de diferentes fontes e natureza, juntamente com a necessidade de explorar essas informações de forma integrada, permitiu a implementação e validação da abordagem proposta em um cenário real, que apoiada por um sistema de informação analítico, apresentou resultados de grande valia para o processo de tomada de decisão da instituição. A exploração conjunta dos dados estruturados e textuais se mostrou viável por meio das funcionalidades analíticas disponibilizadas no sistema, para uso direto pelos usuários finais. Estes se mostram cada vez mais interessados em conhecer e se beneficiar do potencial das análises de instrumentos como o JointOLAP a cada nova funcionalidade apresentada. Tal interesse vem sendo materializado com investimentos no projeto PowerOLAP, tanto na disponibilização de infraestrutura para o desenvolvimento quando no tempo para capacitação e trabalho do seu pessoal técnico e de negócio.

O tratamento conjunto dos dados de diferentes naturezas, associado ao tratamento semântico, reafirmou ser um fator fundamental para a qualidade da informação apresentada. Nesse sentido, está evidente a necessidade de explorar instrumentos terminológicos de alta expressividade semântica como as ontologias. Este é o próximo passo da evolução do JointOLAP, no qual busca-se o aproveitamento dos benefícios de fontes de informação, cada vez mais comuns, que se encontram ou podem ser mapeadas com ontologias do domínio. Fontes como mídias e redes sociais são o principal alvo na próxima versão do JointOLAP.

Embora o experimento tenha sido executado em um ambiente computacional na nuvem – através do serviço da *Amazon Web Services (AWS)* – as potencialidades que as características do *Big Data* podem oferecer, como a eficiência e a escalabilidade para armazenamento, análise e integração de dados, não foram tratadas nesse trabalho.

Referências

Almeida, D. L., & Silva, T. L. (2009). Estratégias e Mecanismos para ETL Textual. *Monografia de Graduação*.

Cardoso, C., & Polidoro, M. (2011). Gestão do Risco da Imagem Institucional. *Congresso de Comunicação Empresarial ABERJE* .

Chakaravarthy, V. T., Gupta, H., Roy, P., & Mohania, M. (2006). Efficiently linking text documents with relevant structured information.

Costa, P. R., Souza, F. F., Times, V. C., & Benevenuto, F. (2012, Julho). Towards integrating Online Social Networks and Business Intelligence.

Heuseler, F. M. (2010). Uma abordagem multifacetada para exploração integrada de dados estruturados e não-estruturados em ambientes OLAP. *Dissertação de Mestrado* .

Inmon, W. (2005). *Building the Data Warehouse.* Wiley.

Inmon, W. H., Strauss, D., & Neushloss, G. (2008). *DW 2.0: The Architecture for the Next Generation of Data Warehousing.*

Kimball, R. (2004). *The Data Warehouse ETL Toolkit.* Wiley.

Kimball, R., & Ross, M. (2002). *The Data Warehouse Toolkit.* Wiley.

Malinowski, E., & Zimányi, E. (2011). *Advanced Data Warehouse Design: From Conventional to Spatial and Temporal Applications.*

Moreira, J. L., Cordeiro, K. F., & Campos, M. L. (2009). DoctorOLAP: Ambiente para análise multifacetada de prontuários médicos. *Artigo apresentado no SBBD 2009* .

Nelson, G. S. (2010). Business Intelligence 2.0: Are we there yet? *SAS Global Forum* .

Nesavich, A., & Inmon, W. H. (2007). *Tapping into Unstructured Data: Integrating Unstructured Data and Textual Analytics into Business Intelligence.*

ONS. (2009). *Procedimentos de Rede – http://www.ons.org.br/procedimentos*

Park, B.-K., & Song, I.-Y. (2011). Toward Total Business Intelligence Incorporating Structured and Unstructured Data.

Perez, J. M., Pedersen, T. B., Berlanga, R., & Aramburu, M. J. (2005). IR and OLAP in XML document warehouses.

Roy, P., Mohania, M., Bamba, B., & Raman, S. (2005). Towards automatic association of relevant unstructured content with structured query results.

Russom, P. (2007). *BI Search and Text Analysis Tics: New Additions to the BI Technology Stack.*