

Mineração de dados e textos e suas possibilidades aplicadas ao processo de produção da notícia

Pablo Barbosa¹
Walter Teixeira Lima Junior²

Resumo: A proposta deste ensaio é discutir os desafios de representar em um formalismo processável computacionalmente o conhecimento que o jornalista usa para articular os critérios de noticiabilidade com o propósito de selecionar e hierarquizar as notícias. Discute como realizar pontes para emular esse conhecimento obtido de forma empírica com as bases da ciência da computação, na área de armazenamento, recuperação e relacionamento de dados em banco de dados, que tentam espelhar o modo que a mente humana trata informações obtidas através do seu sistema sensorial. Sistematizar e automatizar parte do processo jornalístico em uma Base de Conhecimento contribui para eliminar distorções, vícios e aplicar de modo eficiente técnicas de Mineração de Dados e/ou Textos que, por definição, permitem descobrir relações não-triviais.

Palavras-chaves: Jornalismo; notícia; banco de dados; data mining; conhecimento.

Introdução

Por que o ser humano se interessa por notícias ou informações jornalísticas? Essa é uma das questões importantes no campo do jornalismo a ser revelada e pode estar longe de ter uma resposta científica. Nota-se que em todos os países, apesar das diferenças culturais, de sistemas políticos democráticos ou autoritários e de níveis de escolaridade díspares, entre outros pontos comparativos, há no mínimo um sistema midiático baseado na rádio difusão ou no meio impresso. Essa necessidade do homem em conhecer o que está acontecendo ao seu redor e, atualmente, com o advento das redes telemáticas, em qualquer lugar do planeta, é a força que legitima a profissão denominada jornalista, fazendo-a necessária para sociedade.

“Jornalismo é a informação de fatos correntes, devidamente interpretados e transmitidos periodicamente à sociedade, com o objeto de difundir conhecimentos e orientar a opinião pública no sentido da promoção do bem comum” (BELTRÃO, 1992)

¹ Graduando em Jornalismo pela Faculdade de Comunicação da Universidade Federal da Bahia

² Pós-doutor em Comunicação e Tecnologia e professor do Programa de Mestrado da Cásper Líbero

Entretanto, é sabido que o ser humano não se interessa somente por notícias. A absorção de informação é uma necessidade básica e de sobrevivência e a notícia, na sua formatação industrial para consumo em massa, é uma categoria da classe informação. As informações jornalísticas, formatadas para cada meio (texto, imagens e áudio), são impregnadas com os conceitos de importância, utilidade e veracidade e são absorvidas pelo ser humano através dos seus receptores sensoriais (visão, tato, olfato e audição), apenas o paladar, ainda, não é acionado nos produtos editoriais jornalísticos.

Essas informações são importantes para o ser humano em função de serem utilizadas como um redutor de incertezas pela mente. Portanto, elas compõem, com outros tipos, regras para serem utilizadas em um momento de decisão. O processo de cognição dessas informações não tem outro caminho a não ser via absorção delas pelas memórias.

Para receber a mais variada carga e tipo de informação, o ser humano é dotado de um sistema composto de aparelhos sensoriais. Segundo a ciência moderna, não há nada na nossa mente, do mundo que nos circunda, que não tenha passado pelo nosso aparelho sensorial.

O sistema sensorial é parte do sistema nervoso que é responsável pelo processamento da informação sensorial. O sistema sensorial consiste em receptores sensoriais, ligações neurais e partes do cérebro envolvidas na percepção sensorial. Comumente reconhecidos como sistemas sensoriais são a visão, audição, sensação somática (tato), paladar e olfato, mas existem outros, como percepção de equilíbrio.

É esse sistema que permite que tenhamos uma representação particular da realidade e realiza a tradução das informações coletadas no ambiente. Contudo, na atualidade, o jornalista além de recolher informações do ambiente do fato, o profissional pode coletar dados por intermédio dos meios de comunicação, além de utilizar de tecnologias de captação em rede (internet) e tradicionais, como o telefone.

O jornalismo, então, é a prática em que um ser humano (jornalista) recolhe as informações do ambiente e/ou através de tecnologias de captação, seguindo critérios técnicos e mercadológicos tenta traduzir essa representação do real através de plataformas comunicacionais, nas suas respectivas linguagens (impresso, eletrônico e digital), para outros seres humanos (leitores, radiouvintes, telespectadores e internautas), que utilizam sistema sensorial deles para absorver tais informações, que serão memorizadas ou descartadas de acordo com interesse, perfil cultural e

historicidade de cada um. Nesses dois estágios de absorção, fica evidente a redução informativa imposta pela que a cadeia perceptiva entre o jornalista e o usuário final da informação.

Na psicologia e nas Ciências Cognitivas, percepção é o processo de aquisição, interpretação, seleção e organização da informação sensorial. A palavra percepção vem do Latim, *capere*, que significa “tomar”, o prefixo *per* significa “completamente”.

Os métodos de estudo da percepção vão desde as aproximações essenciais com a biologia ou psicologia, através de abordagens psicológicas da filosofia da mente e a epistemologia empírica, tais como de David Hume, George Berkeley ou como a afirmação de Merleau Ponty que tem a percepção como a base de todas as ciências e conhecimento.

Será possível emular?

No campo das Ciências Sociais aplicadas há muitos questionamentos, por parte dos pesquisadores da área, sobre a possibilidade de máquinas computacionais receberem dados e rodarem programas que selecionem, encontrem ou hierarquizem, mesmo que de forma reduzida (modelagem) e por aproximação do processo mental humano, notícias jornalísticas dentro dos valores-notícia apontados por pesquisadores da área, como Gislene Silva e Érica Frazon.

Ora, ao tratar jornalisticamente os fatos na produção material da notícia, a seleção e hierarquização recorrem sim aos valores- notícia. Mas estes agem aqui apenas como uma parte do processo, pois nessas escolhas seqüenciadas entrarão outros critérios de noticiabilidade, como formato do produto, qualidade da imagem, linha editorial, custo, público alvo etc. Valores-notícia, as características do fato em si, em sua origem, são somente um subgrupo de fatores agindo juntamente com esse segundo conjunto de critérios de noticiabilidade, relacionados agora ao tratamento do fato. Estudar a seleção implica, inclusive, rastrear os julgamentos próprios de cada seletor, as influências organizacionais, sociais e culturais que este sofre ao fazer suas escolhas, os diversos agentes dessas escolhas postados em diferentes cargos na redação, e até mesmo a participação das fontes e do público nessas decisões – aqui vale lembrar os estudos de agendamento (agenda-setting), que complexificam as investigações sobre o processo de seleção das notícias. (SILVA, 2005, p.5)

Apesar de toda a complexidade do tema apontada pela pesquisadora da Universidade Federal de Santa Catarina, a sua sistematização dos critérios de noticiabilidade revela partes de como se constitui os valores-notícias. É um trabalho de

engenharia reversa do pensamento humano, descortinando um pouco quais são as causas que fazem um ser humano prestar mais ou menos atenção em uma notícia (anexo). Ou seja, o trabalho é uma tentativa inicial de estruturar e de classificar atributos e suas respectivas escalas de valores da notícia. Esse tipo de organização pode servir de base para iniciar simulações de modelos, utilizando sistemas computacionais com o auxílio de banco de dados. Essa afirmativa parte do pressuposto que há uma lógica humana na busca pela notícia. A lógica não é privilégio de alguma área do conhecimento, ela permeia toda a atividade humana. A lógica, por exemplo, se transformou na linguagem básica das ciências formais. Os sistemas computacionais e os softwares são produtos das ciências formais.

Uma ciência é qualquer corpo organizado de conhecimento que possui princípios. Os primeiros princípios de qualquer ciência são aquelas verdades fundamentais em que se apóia e em que todas as suas atividades se baseiam. A lógica, como ciência, tem seus princípios fundamentais, mas a lógica guarda uma relação única com todas as outras ciências, porque os primeiros princípios da lógica aplicam-se não apenas à lógica, mas a todas as outras ciências. Na verdade, suas bases são mais abrangentes, porque se aplicam à razão humana como tal, embora isso deva ser exercitado (MCLNERNY, 2006, p. 46)

Identificando parâmetros

O processo empírico e anos de refinamento estabeleceram uma ligação entre o que o usuário entende como notícia e o que é transmitido pelos jornalistas através da mídia. Essa conexão é estabelecida pelos valores-notícias mencionados acima. Portanto, construir uma Base de Conhecimento (BC) juntando áreas, então, díspares, não é uma tarefa fácil.

É preciso, em primeiro lugar, articular pessoas (GARCIA, FLÁVIO, FERRAZ, 2005). Encontrar a sinergia entre engenheiros de conhecimento e especialistas em um domínio exige dedicação, planejamento e uma dose de boa vontade.

Enquanto os engenheiros de conhecimento são responsáveis por transformar as idéias, os conceitos e os modos de racionalizar o mundo em um modelo processável computacionalmente, o especialista em um domínio precisa se esforçar para traduzir seu conhecimento em uma linguagem clara e objetiva, além de avaliar e apontar os deslizos cometidos pelo sistema com base nas respostas obtidas. Isso significa dizer que é um ciclo contínuo de aperfeiçoamento.

Entretanto, além do problema da sinergia, outro grande gargalo para o sucesso do desenvolvimento de uma BC é a captura de conhecimento. Diversas técnicas são empregadas para solucionar a questão, mas nenhuma delas ainda é perfeita e por isso, exige criatividade do engenheiro de conhecimento e paciência do especialista para corrigir as falhas. As mais comuns empregadas são:

Técnicas de aquisição de conhecimento manuais baseadas em entrevistas, em acompanhamentos ou em modelos; técnicas de aquisição semi-automáticas baseadas em teorias cognitivas ou em modelos existentes; tecnologia de aprendizado de máquina tentando induzir regras a partir de exemplos catalogados; tecnologia de mineração de dados que tenta extrair regras e comportamento a partir da análise de grandes massas de dados; tecnologia de mineração de textos que tenta extrair conhecimento de grandes massas de dados não-estruturados. (GARCIA, VAREJÃO, FERRAZ, 2005, p. 68)

Contudo, após extrair e representar computacionalmente o conhecimento de um domínio, o trabalho não se torna mais fácil. Ainda é preciso atender alguns requisitos para que a Base de Conhecimento seja eficiente:

[Ela deve] ser compreensível ao ser humano, pois caso seja necessário avaliar o estado de conhecimento do sistema, a Representação do Conhecimento deve permitir sua interpretação; Abstrair-se dos detalhes de como funciona internamente o processador de conhecimento que a interpretará; Ser robusta, isto é, permitir sua utilização mesmo que não aborde todas as situações possíveis; Ser generalizável, ao contrário do conhecimento em si que é individual. Uma representação necessita de vários pontos de vista do mesmo conhecimento, de modo que possa ser atribuída a diversas situações e interpretações. (REZENDE, PUGLIESI, VAREJÃO, 2005, p. 29)

Mais do que sistematizar e automatizar parte do processo jornalístico, a construção de uma Base de Conhecimento (BC) com as melhores práticas permite comparar os registros do banco de dados com as regras estabelecidas e prover um armazenamento posterior dos padrões encontrados, beneficiando assim outros processos. Dentre eles, destacamos os benefícios de aplicar a Mineração de Dados e/ou Textos para auxiliar a apuração, complementação e até o furo jornalístico.

É importante ressaltar que apesar dos temas Mineração de Dados ou *Data Mining* (DM) e Mineração de Textos ou *Text Mining* (TM) serem amplamente discutidos no campo da ciência da computação, o esforço de relacioná-los em aplicações no jornalismo é recente. Portanto, naturalmente, enfrenta dificuldades de compatibilização.

Basicamente, falar de *Data Mining* é buscar padrões ocultos em massas de dados que encontramos em *data warehouses*³ corporativos ou Bases de Conhecimento de Sistemas Inteligentes. Como conceito que envolve Estatística, Inteligência Artificial e *Machine Learning* (Aprendizado de Máquina), o DM garimpa informações de valor estratégico que estão “invisíveis” nos registros, permitindo a identificação de tendências para uma visão antecipada de cenários futuros e a descoberta de novos padrões entre dados, nem sempre perceptíveis ao analista humano.

Há várias definições de Mineração de Dados, mas a que vem sendo mais aceita é a de Usama Fayyad (1996).

Extração de conhecimento de Base de Dados é o processo não-trivial de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis embutidos nos dados (FAYYAD, PIATESKYSHAPIRO, SMYTH, 1996).

Para efeito comparativo, Rowe & Cilione (2000) entendem que a “mineração de dados é um processo analítico desenhado para explorar grandes massas de dados à procura de padrões consistentes e/ou relacionamentos sistemáticos entre variáveis, e então validar os achados aplicando os padrões detectados para novos conjuntos de dados. O processo consiste de quatro estágios básicos: (1) preparação dos dados, (2) exploração, (3) construção do modelo (ou definição de padrão), e (4) validação/verificação⁴”.

³ O *data warehouse* (ou armazém de dados) é um sistema de computação utilizado para armazenar informação relativa às atividades de uma organização em bancos de dados, de forma consolidada. O desenho da base de dados favorece os relatórios e análise de grandes volumes de dados e obtenção de informações estratégicas que podem facilitar a tomada de decisão. O processamento de dados em um *data warehouse* é sempre referenciado como *Online Analytical Processing* (OLAP) ou Processo Analítico em Tempo Real, em contraste com o *Online Transaction Processing* (OLTP) - usado para armazenar as operações de negócios. Outra diferença, é que os dados em um *data warehouse* não são voláteis, ou seja, eles não mudam, salvo quando é necessário fazer correções de dados previamente carregados. Os dados então são somente para leitura e não podem ser alterados.

O *data warehouse* possibilita a análise de grandes volumes de dados, armazenados pelos sistemas transacionais (OLTP). São as chamadas séries históricas que possibilitam uma melhor análise de eventos passados para a tomada de decisões presentes e a previsão de eventos futuros.

Pela sua capacidade de sumarizar grandes volumes de dados e de possibilitar análises os *data warehouses* são atualmente o núcleo dos sistemas de informações gerenciais e apoio à decisão das principais soluções de business intelligence do mercado. Disponível em: http://pt.wikipedia.org/wiki/Data_warehouse Acesso em: 27/06/2007

⁴ Tradução dos autores: “Data mining is an analytic process designed to explore large amounts of data in search for consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data. The process consists of four basic stages: (1) data preparation, (2) exploration, (3) model building (or pattern definition), and (4) validation/verification.” (Rowe, Cilione, 2000). Para mais informações, ver: ROWE, Ken; CILIONE, Patrick. *Data Mining and Neural Networks Analysis*. Disponível em:

Como essência das duas definições, percebemos que sua utilidade está relacionada ao desenvolvimento de sistemas automatizados para a produção jornalística, mas como explica Walter Lima Jr (2006, p. 125) “é preciso que os bancos de dados sejam precisos e não históricos, e que tenham certa inteligência artificial para lidar com as modificações semânticas das palavras, por exemplo. Com o *data mining* é possível extrair padrões⁵ válidos, por exemplo, para investigar se o índice de desemprego diminui quando se aproxima uma eleição e por que isso acontece”.

A possibilidade descrita acima é pertinente por um simples motivo: mesmo em bancos relacionais, quando bem projetados, há uma extração de diversas informações utilizando *Structured Query Language* (SQL)⁶, porém, o processo exige, necessariamente, a elaboração de questões para serem resolvidas. Por mais criativo que o analista seja, ele conseguirá elaborar apenas algumas perguntas para que o sistema mapeie o banco e traga resultados práticos no final, ou seja, devido o volume de dados envolvidos, padrões e comportamentos relevantes são ignorados e neste ponto, o *Data Mining* demonstra suas vantagens.

O processo de mineração identifica por meio de *tarefas* (que são classes de problemas) e *técnicas* (que são grupos de soluções que utilizam algoritmos para os problemas propostos nas tarefas) as perguntas e as respostas na base de dados. Em síntese, é possível não só relacionar eventos com base no histórico, mas a partir daí, atuar de modo preditivo. É válido ressaltar que não há intenção de dizer que o jornalista passará a dar credibilidade a especulações ou pior, a fatos inventados com esse tipo de aplicação, mas é notável o ganho de precisão para acompanhar às minúcias de um desdobramento.

Outra aplicação também relevante para o jornalismo é a Mineração de Textos (MT). Segundo Tan (1999), esta nova área é definida como o processo de extrair padrões ou conhecimentos, interessantes e não-triviais, a partir de um conjunto de documentos textuais. Há uma semelhança com a definição de Usama Fayyad (1996) sobre DM, mas é fato que a inspiração da MT adveio, efetivamente, do processo de *Data Mining*. Entretanto, diferente do DM que consiste em extrair informação de

<http://web.archive.org/web/20050616182712/http://acspri.anu.edu.au/newsletter/news42/DataMining.htm>

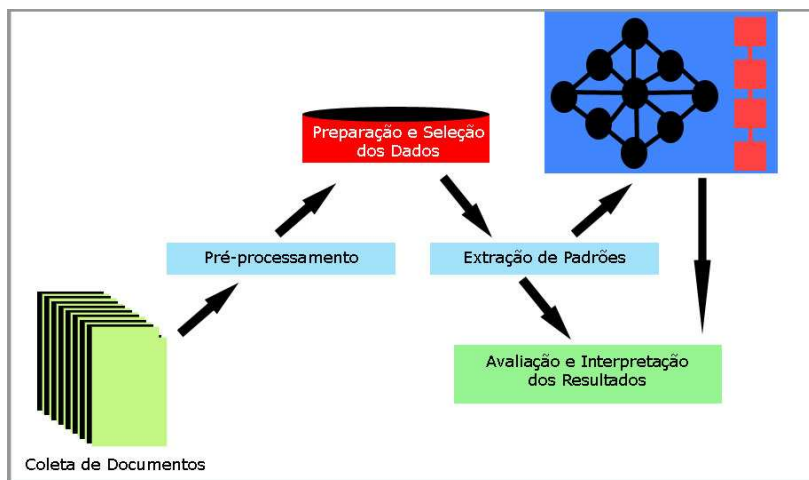
Acesso em: 27/06/2007

⁵ Padrões são unidades de informação que se repetem ou, então, são seqüências de informações que dispõem de uma estrutura que se repete.

⁶ Structured Query Language, ou Linguagem de Consulta Estruturada ou SQL, é uma linguagem de pesquisa declarativa para banco de dados relacional (base de dados relacional). Muitas das características originais do SQL foram inspiradas na álgebra relacional. Disponível em

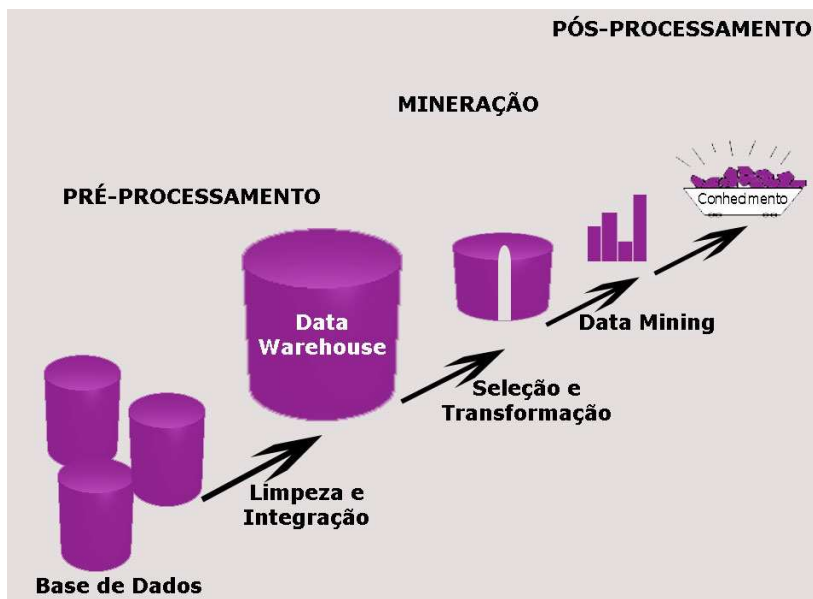
bancos de dados estruturados, a MT extrai informação de dados não-estruturados ou semi-estruturados. Essa diferença permite trabalhar com vários fatores que acarretam uma complexidade de tarefas como, por exemplo, lidar com os diferentes tipos de linguagem, estilo ou conteúdo do documento escrito.

De um modo geral, as duas aplicações são discutidas em diversos livros e artigos como parte de um processo maior chamado *Knowledge Discovery in Databases* (KDD), que significa Descoberta de Conhecimento em Base de Dados e *Knowledge Discovery from Texts* (KDT), que seria Descoberta de Conhecimento em Textos. Ambos possuem semelhanças no tratamento dos dados e prevêem várias etapas como seleção, pré-processamento, transformação, mineração dos dados, gerador de relatórios, bem como a interpretação dos resultados obtidos. Alguns autores⁷ apresentam variações destas etapas e agrupam em pré-processamento, mineração e pós-processamento. Abaixo seguem dois gráficos que ilustram o processo de KDD e KDT:



PROCESSO DE KDT

⁷ Tais como: Rezende, Pugliesi, Melanda & de Paula (2005) e Fayyad (1996)



PROCESSO DE KDD – Inspirado em uma imagem do livro de Morgan Kaufmann. Data Mining: Concepts and Techniques, 2000, p. 22

Em síntese, o pré-processamento é a etapa que visa gerar uma representação conveniente para os algoritmos de mineração, a partir da base de dados. Inclui a seleção (automática e/ou manual de atributos relevantes), amostragem, transformações de representação, etc. A mineração de dados, por sua vez, é a aplicação de algoritmos aos dados pré-processados e o pós-processamento resume-se a seleção e ordenação das descobertas interessantes, mapeamentos de representação de conhecimento, geração de relatórios e interpretação por parte dos usuários e/ou especialistas.

Para compreender a real vantagem de utilizar mecanismos que auxiliem o processo de apuração e complementação de material, pensemos em dois grandes grupos midiáticos brasileiros que apostaram na construção de bancos de dados: o Grupo Abril e o grupo que inclui a Folha de S.Paulo. Eis o que informa o site de um desses grupos:

...o banco de dados Folha é um acervo jornalístico que contém mais de oito décadas da história recente do Brasil. Seu objetivo é dar suporte aos jornalistas do Grupo Folha da Manhã e propiciar o atendimento a pesquisadores, estudantes e empresas na realização de pesquisas. O acervo inclui a coleção de jornais editados pelo grupo, arquivo de recortes com cerca de 100 mil pastas temáticas e 20 milhões de imagens em arquivos físico e digital⁸.

Já o Grupo Abril, lembra Walter Lima (2006, p. 121), tem o seu Dedoc, inaugurado em 1968. “Antes, tudo era manual. Em 1984, iniciou-se o processo de informatização. Primeiro foi a vez da revista Veja, com acesso ao resumo de todas as

⁸ Disponível em <http://bd.folha.uol.com.br/> Acesso em: 10 de julho de 2007

matérias e pesquisa de palavras-referência. Todas as revistas do grupo estão num banco de dados chamado Fólio News. Na época do levantamento, a revista Veja, carro-chefe da editora, por exemplo, tinha 43.687 matérias; Anamaria, 19.587; Exame, 12.958; e Cláudia, 11.262”.

Apesar deste volume de informação nos veículos, não existe um sistema eficiente e eficaz de relacionamento de informação (baseado em palavras-chaves) que permita aos jornalistas utilizar sua capacidade de conectar informações para executar tarefas relativamente simples, tais como, traçar panoramas. Supondo que o *deadline* para elaborar um caderno especial sobre a trajetória de um determinado movimento social, desde a década de 80, seja de um dia e o editor escolha apenas um jornalista para esta tarefa, o tempo será escasso para tal tarefa.

A dificuldade neste caso não é falta de informação, mas como fazer a triagem da melhor forma e em menor tempo. Qual o melhor modo de argüir o sistema para trazer o que se deseja e sem perder tempo vasculhando dados que não são importantes para o contexto da matéria? Melhor, como fazer encontrar o que deseja com apenas uma sentença? Ou ainda, por quê o sistema não trouxe isso há tempos de modo automático sem a necessidade de uma pauta com o tempo de apuração curto?

O que esse ensaio sugere, mas que não foge do universo das redações, é que o tempo é curto e os jornalistas precisam de mecanismos que auxiliem a encontrar as informações relevantes, precisas e balisadoras para a sua matéria. Recorrer ao departamento de Arquivo não é mais possível. Devido a constante política de redução de custos, esse tipo de seção foi extinta em muitos veículos. Anteriormente, esse era o melhor modo de obter a memória dos principais acontecimentos. Na história do jornalismo é fácil encontrar matérias importantes que surgiram com a ajuda desse tipo de setor.

Jornalistas e conexões informativas

Jornalistas, que se denominam de investigativos, como Cláudio Júlio Tognolli, utilizou muitos dos serviços oferecidos pelo banco de dados da Editora Abril, mas também, segundo ele, desenvolveu habilidades para extrair as informações de sistemas de busca na web (LIMA JR, 2006).

Walter Lima Jr (2006) afirma que “A técnica de Tognolli baseia-se em sempre começar procurando pelo Google Imagens, e nunca pelo Google Texto, pois, segundo o

jornalista, o mecanismo fornece um “substrato caótico” de imagens mais interessante do que o outro sistema:”

Portanto, se eu tenho um determinado repertório baseado em livres associações, sobre uma pessoa, e eu quero pesquisar essa pessoa na internet, eu penso por alguns minutos nela e a associo a uns vinte ou trinta vocábulos. Bem simples. E coloco “o nome dela e and crime”, “and carro”, “and guitarra”, mas baseado na minha visão daquela pessoa. Então, eu faço um esquema booleando, usando and, com livres-associações (Tognolli, 2004)

O problema no exemplo citado é que cada profissional possui uma técnica refinada para alcançar o seu objetivo. Com a utilização de aplicações, tais como, *Data Mining* ou *Text Mining* isso seria reforçado e viabilizaria uma atualização de alguns potenciais que definem o jornalismo digital⁹, a exemplo da memória e personalização ou customização de conteúdos.

O *Text Mining* (TM), por exemplo, possibilita exibir uma representação estruturada dos documentos, frequentemente no formato de uma tabela atributo-valor. Essa tabela atributo-valor caracteriza-se pela alta dimensionalidade, uma vez que cada termo do documento pode representar um possível elemento do conjunto de atributos da tabela, logo, é imprescindível a seleção dos dados, a fim de reduzir a extensão da tabela atributo-valor, de modo que se alcance dados efetivamente expressivos.

Para alcançar este objetivo, durante a etapa de pré-processamento é realizado, dentre outras coisas, uma análise léxica, uma eliminação de termos considerados irrelevantes, ou *stopwords*¹⁰ bem como a normalização morfológica dos termos (remoção de prefixos e sufixos).

Mas, afinal, como o *Text Mining* pode auxiliar o jornalismo a partir de uma representação estruturada dos documentos? Como encontrar termos essenciais? Adiantamos que a resposta não é simples. Em linhas gerais há duas formas de abordagem para se trabalhar com os dados textuais: uma é através da análise semântica e a outra através da estatística. A primeira diz respeito a uma avaliação da seqüência dos

⁹ Jornalismo digital é todo produto discursivo que constrói a realidade por meio da singularidade dos eventos, que tem como suporte as redes telemáticas ou qualquer outro tipo de tecnologia por onde se transmitam sinais numéricos e que incorpore a interação com os usuários ao longo do processo produtivo. É uma das atividades que se desenvolve no ciberespaço (MACHADO, 2000, p.19). Quanto as características que definem o jornalismo digital Schwingel (2005, p. 01) diz o seguinte: é composto pela hipertextualidade, multimidialidade, atualização contínua, memória, personalização ou customização de conteúdos, interatividade e a supressão dos limites de tempo e espaço para a postagem de informações em sua natureza primeira.

¹⁰ Ebecken, Lopes & Costa (2005, p. 347) dizem que *stopwords* ou *stoplist* é o nome dado a retirada de palavras ou termos que não contém conhecimento no texto, tais como, palavras auxiliares ou conectivas (e, para, a, eles) que não traduzem a essência dos textos.

termos no contexto da frase, enquanto a segunda dedica-se a contabilizar o número de vezes que um termo aparece no texto.

Análise Semântica

Utilizando fundamentos e técnicas baseadas no processamento de linguagem natural, a análise semântica propicia, quando incrementada por um processamento lingüístico mais complexo, identificar corretamente a função de cada termo. Para isso são necessários alguns tipos de conhecimento. São eles:

Morfológico: conhecimento da estrutura, da forma e das inflexões das palavras. **Sintático:** conhecimento estrutural das listas de palavras e como as palavras podem ser combinadas para produzir sentenças. **Semântico:** o que as palavras significam independentes do contexto, e como significados mais complexos são formados pela combinação de palavras. **Pragmático:** o conhecimento do uso da língua em diferentes contextos, e como o significado e a interpretação é afetada pelo contexto. **Discursivo:** como as sentenças imediatamente precedentes afetam a interpretação da próxima sentença. **Mundo:** conhecimento geral do domínio ou o mundo que a comunicação da linguagem natural se relaciona. (EBECKEN, LOPES, COSTA, 2005, p. 339-340)

Os problemas dessa abordagem são basicamente dois: customizar a aplicação de acordo com o idioma e apresentar um índice elevado erros ao tentar chegar à essência de uma figura de linguagem. Entretanto, o segundo problema é amenizado sob a justificativa que a boa prática jornalística evita ironias, metonímias e todo tipo de estrutura que possa comprometer o bom entendimento.

Análise Estatística

Já a análise estatística prevê o conhecimento a partir de um estudo quantitativo dos termos. A principal vantagem dessa aplicação é permitir que esta estratégia seja utilizada em qualquer idioma, contudo, como a repetição de termos não é aconselhável em um texto jornalístico deve-se recorrer a métodos que ultrapassem esse obstáculo. A solução é definir um dicionário ou *thesaurus* como um vocabulário controlado que representa termos variantes, tais como, sinônimos, abreviações, acrônimos e ortografias alternativas (EBECKEN, LOPES, COSTA, 2005, p. 349). É importante ressaltar que esta alternativa também pode ser aplicada à abordagem semântica, mas o tempo de

processamento da análise estatística é inferior para relacionar grandes quantidades de textos.

Tanto o *Text Mining* quando o *Data Mining* recorrem a diversas classes de tarefas para descobrir relações não triviais e assim atuar de modo eficiente no objetivo proposto. No caso específico, pretendemos que as aplicações atuem na apuração, complementação e até no furo jornalístico.

Ebecken, Lopes & Costa (2005) destacam as seguintes tarefas do *Text Mining*:

O processo de **agrupamento ou clustering** [grifo meu] torna explícito o relacionamento entre documentos, enquanto a **categorização** [grifo meu] identifica os tópicos-chaves de um documento. A **extração de características** [grifo meu] é usada quando é preciso conhecer pessoas, lugares, organizações e objetos mencionados no texto. A **sumarização** [grifo meu] estende o princípio de extração de características concentrando-se mais em sentenças inteiras que em nomes ou frases. A **indexação temática** [grifo meu] é útil quando se quer ser capaz de trabalhar preferencialmente com tópicos do que com palavras-chaves. (EBECKEN, LOPES, COSTA, 2005, p. 351)

Já o *Data Mining* divide suas classes tarefas em dois grupos: preditivas e descritivas. São elas:

Preditivas - Classificação: a tarefa objetiva o agrupamento de dados em classes pré-definidas. Estimativa (regressão): objetiva definir um valor (numérico) de alguma variável desconhecida a partir dos valores de variáveis conhecidas. **Descritivas** - Associação: estuda um padrão de relacionamento entre itens de dados. [Ela] é usada para identificar padrões em dados históricos. Clusterização (segmentação): as informações podem ser particionadas em classes de segmentos similares. Neste caso, nada é informado ao sistema a respeito das classes existentes. O próprio algoritmo descobre as classes a partir das alternativas encontradas na base de dados, agrupando assim um conjunto de objetos em classes de objetos semelhantes. (VIANA, 2004, p.18)

Esse tipo de tratamento de dados e os demais demonstrados ao longo do ensaio, mostra que há uma efetiva possibilidade de transmissão de procedimentos e padrões, já utilizados em outras áreas, para um conjunto de sistemas digitais de armazenamento e relacionamento de dados e informações visando à construção de conhecimento no campo do jornalismo.

Considerações finais

O uso de tecnologias no campo do jornalismo não é algo novo e surpreendente. Entretanto, com o advento de suportes digitais através de banco de dados e softwares que tratam as informações de modo ágil, relacional, customizado e hierarquizado, a prática jornalística ganha novas possibilidades e amplia sua capacidade de trato informativo qualitativo. Os processos de apuração, complementação e busca do fato jornalístico relevante e inédito, por exemplo, são oxigenados devido ao esforço de construção dessa Base de Conhecimento baseada nas melhores práticas do campo e na aplicação de um ferramental tecnológico de ponta, tais como, o *Data Mining* e o *Text Mining*.

O que propomos é um trabalho de engenharia reversa do pensamento humano, que descortina um pouco quais são as causas que fazem um ser humano prestar mais ou menos atenção em uma notícia.

Ou seja, o trabalho é uma tentativa inicial de estruturar e de classificar atributos e suas respectivas escalas de valores da notícia. Esse tipo de organização pode servir de base para iniciar simulações de modelos, utilizando sistemas computacionais com o auxílio de banco de dados. Essa afirmativa parte do pressuposto que há uma lógica humana na busca pela notícia.

Nesse sentido, acreditamos que utilizar os conhecimentos já consolidados nos campos das Ciências da Computação e Ciências Cognitivas para solucionar os problemas sistemáticos dos processos rudimentares de produção do jornalismo em base de dados, por exemplo, proporcionará ao jornalismo a equalização e sintonia das suas práticas e objetivos com uma sociedade globalizada, que parte já possui dispositivos multimidiáticos de obtenção de informação com alto poder computacional, em real time, em alta definição, com conexão sem fio, a caminho da sociedade do conhecimento.

Referências:

- ALVARES, Reinaldo V. **Mineração de Dados: Introdução e Aplicações**. Artigo publicado na revista SQL Magazine, edição 10, ano 1, 2004
- BELTRÃO, Luiz. **Iniciação à filosofia do jornalismo**. São Paulo: EDUSP, 1992
- DA SILVA, Cassiana F. **Uso de Informações Linguísticas na etapa de pré-processamento em Mineração de Texto**. Dissertação de mestrado defendida no Programa de Pós-Graduação em Computação Aplicada da Universidade do Vale do Rio do Sinos, São Leopoldo (RS), 2004.
- DAVIS, R., SHROBE, H. and SZOLOVITS, P. **What is a Knowledge Representation?** AI Magazine, v.14, no.1, p. 17-33, Menlo Park, USA. 1993

- Disponível em: <http://groups.csail.mit.edu/medg/ftp/psz/k-rep.html> Acesso em 23 de junho de 2007
- HAN, J., KAMBER, M. **Data mining: concepts and techniques**. USA: Morgan Kaufmann, 2001.
- KORFHAGE, Robert R. **Information Retrieval and Storage**. New York: John Wiley & Sons, p. 349, 1997.
- KOWALSKI, Gerald. **Information Retrieval Systems: Theory and Implementation**. Boston: Kluwer Academic Publishers, p. 282, 1997.
- LAGE, Nilson. **Ideologia e técnica da notícia**. Florianópolis: Ufsc-Insular, 2001.
- LEME, Maria Isabel da Silva. **Aquisição de conhecimento**. Bol. psicol. [online]. dic. 2005, vol.55, no.123, p.233-239. Disponível em: http://pepsic.bvs-psi.org.br/scielo.php?script=sci_arttext&pid=S0006-59432005000200008&lng=es&nrm=iso Acesso em 24 de junho de 2007
- LIMA JR. Walter Teixeira. **Jornalismo Inteligente na era do data mining**. Publicado na Revista do Programa de Pós-graduação da Faculdade Cásper Líbero, ano IX – no.18, p. 121-126, 2006.
- MACHADO, E. **La estructura de la noticia en las redes digitales: un estudio de las consecuencias de las metamorfosis tecnológicas en el periodismo**. Tese de doutorado defendida no Programa de Doutorado em Jornalismo e Ciências de Comunicação da Universidade Autônoma de Barcelona. Barcelona (Espanha), Junho de 2000.
- MCLNERNY, D. Q. **Use a Lógica**. Rio de Janeiro: Best Seller, 2006.
- REZENDE, Solange (org) **Sistemas Inteligentes – Fundamentos e Aplicações**. Barueri, São Paulo: Manole, 2005
- SALTON, G.;MACGILL, M. **Introduction to Modern Information Retrieval**. New York: McGRAW-Hill, p.448, 1983.
- SCHWINGEL, Carla. **Jornalismo digital de quarta geração a emergência de sistemas automatizados para o processo de produção industrial no Jornalismo Digital**. Apresentado no GT de Estudos de Jornalismo da Compós em Porto Alegre (RS), 2005.
- SILVA, Gislene. **Valores-notícia: atributos do acontecimento (Para pensar critérios de noticiabilidade I)**. Trabalho apresentado ao NP 02 - Jornalismo, do IV Encontro dos Núcleos de Pesquisa da Intercom, Porto Alegre, 2004.
- TAN, A. **Text mining: the state of the art and the challenges**. In: Pacific-Asia Workshop on Knowledge Discovery from Advanced Databases - PAKDD'99, p. 65-77, Beijing, abril 1999.
- TOGNOLLI, Cláudio Júlio. Entrevista concedida a Walter Teixeira Lima Júnior em 10 de setembro de 2004.

Anexo

Elencos de valores-notícia

<i>Stieler</i> : novidade, proximidade geográfica, proeminência e negativismo.
<i>Lippman</i> : clareza, surpresa, proximidade geográfica, impacto e conflito pessoal.
<i>Bond</i> : referente à pessoa de destaque ou personagem público (proeminência); incomum (raridade); referente ao governo (interesse nacional); que afeta o bolso (interesse pessoal/econômico); injustiça que provoca indignação (injustiça); grandes perdas de vida ou bens (catástrofe); conseqüências universais (interesse universal); que provoca emoção (drama); de interesse de grande número de pessoas (número de pessoas afetadas); grandes somas (grande quantia de dinheiro); descoberta de qualquer setor (descobertas/invenções) e assassinato (crime/violência).
<i>Galtung e Ruge</i> : freqüência, amplitude, clareza ou falta de ambigüidade, relevância, conformidade, imprevisão, continuidade, referência a pessoas e nações de elite, composição, personificação e negativismo.
<i>Golding-Elliot</i> : drama, visual atrativo, entretenimento, importância, proximidade, brevidade, negativismo, atualidade, elites, famosos.
<i>Gans</i> : importância, interesse, novidade, qualidade, equilíbrio.
<i>Warren</i> : atualidade, proximidade, proeminência, curiosidade, conflito, suspense, emoção e conseqüências.
<i>Hetherington</i> : importância, drama, surpresa, famosos, escândalo sexual / crime, número de pessoas envolvidas, proximidade, visual bonito / atrativo.
<i>Shoemaker et all</i> : oportunidade, proximidade, importância / impacto, conseqüência, interesse, conflito /polêmica, controvérsia, sensacionalismo, proeminência, novidade / curiosidade / raro.
<i>Wolf</i> : importância do indivíduo (nível hierárquico), influência sobre o interesse nacional, número de pessoas envolvidas, relevância quanto à evolução futura.
<i>Erbolato</i> : proximidade, marco geográfico, impacto, proeminência, aventura / conflito, conseqüências, humor, raridade, progresso, sexo e idade, interesse pessoal, interesse humano, importância, rivalidade, utilidade, política editorial, oportunidade, dinheiro, expectativa / suspense, originalidade, culto de heróis, descobertas / invenções, repercussão, confiança s.
<i>Chaparro</i> : atualidade, proximidade, notoriedade, conflito, conhecimento, conseqüências, curiosidade, dramaticidade, surpresa.
<i>Lage</i> : proximidade, atualidade, identificação social, intensidade, medatismo, identificação humana.

Fonte: SILVA, Gislene. Valores-notícia: atributos do acontecimento. Trabalho apresentado ao NP 02 – Jornalismo, do V Encontro dos Núcleos de Pesquisa da Intercom, 2005