Mineração de Opiniões aplicada à Análise de Investimentos

Thomas Jefferson P. Lopes
Centro Universitário Senac
Av. Eng. Eusébio Stevaux, 823
Santo Amaro, São Paulo
+55 11 5682-7300
thomas@thlopes.com

Gabriel Koji Lemos Hiratani
Centro Universitário Senac
Av. Eng. Eusébio Stevaux, 823
Santo Amaro, São Paulo
+55 11 5682-7300
gabrielkoji@gmail.com

Fabrício J. Barth
Centro Universitário Senac
Av. Eng. Eusébio Stevaux, 823
Santo Amaro, São Paulo
+55 11 5682-7300
fabricio.jbarth@sp.senac.br

Orlando Rodrigues Jr.
Centro Universitário Senac
Av. Eng. Eusébio Stevaux, 823
Santo Amaro, São Paulo
+55 11 5682-7300
orlando.rodrijr@sp.senac.br

Juliana Maraccini Pinto
Centro Universitário Senac
Av. Eng. Eusébio Stevaux, 823
Santo Amaro, São Paulo
+55 11 5682-7300
juliana.mpinto@sp.senac.br

ABSTRACT
Investors, before making an investment decision, rely on many sources of information, including the Web, which in current times has become an important mean of mass production and dissemination of information for the financial/stock market.

In this work, we study techniques of Opinion Mining, applied to investment analysis, based on news sources disseminated on the Web, in order to achieve a better consumption and processing of this information, extracting the relevant part of the huge amount of unstructured data (text) generated on the Web.

Categories and Subject Descriptors
H.1.1 [**Systems and Information Theory**]: Value of Information – *mining opinion from news text.*

General Terms
Algorithms, Measurement, Economics, Experimentation, Human Factors.

Keywords
Opinion Mining, Information Extraction, Investment Analysis

1.INTRODUÇÃO
No mercado financeiro, um investidor, tentando minimizar perdas e maximizar ganhos, procura informações referentes às ações em diversas fontes, e de diversos tipos, como: notícias, balanços, gráficos, relatórios e também sugestões de compra e venda de ativos feitos pelas corretoras e fóruns com outros investidores e analistas [1]. Porém, essa tarefa pode tomar muito tempo ao ser executada por uma pessoa, visto a quantidade de informação que deve ser analisada. Torna-se mais crítica quando se leva em conta a informação que pode ser gerada e transmitida por meio da Web, em blogs, mídias das empresas, e principalmente, sites de notícias sobre o mercado financeiro.

Pesquisas na área de investimentos já constataram que, no mercado moderno de ações, as notícias divulgadas podem influenciar o preço das ações na bolsa [2], mesmo que de forma rápida e meramente especulativa. Quem define esse valor é quem negocia (o Mercado em si) no momento da compra e venda [3], logo essa opinião global certamente está relacionada ao valor de um ativo no Mercado. Então, teoricamente, seria possível utilizando métodos computacionais, extrair as informações relevantes dessas fontes de notícias, a fim de ajudar os investidores em sua tarefa de coletar informação que auxilie num investimento.

O objetivo principal deste trabalho é apresentar um processo que permita sumarizar opiniões sobre um ativo encontradas em fontes de informação disponíveis na Web - criar uma visualização dentro de um período de tempo. Para esse processo, é essencial capturar informações relevantes sobre ativos financeiros na Web, e identificar dentro dessas informações opiniões relacionadas com os ativos analisados.

Este trabalho está estruturado da seguinte maneira: na seção 2 são descritos os conceitos sobre mineração de opiniões; na seção 3 é

apresentada a proposta deste trabalho; na seção 4 os resultados parciais, e; na seção 5 são descritas as considerações finais.

2.FUNDAMENTAÇÃO TEÓRICA

A informação na Web pode ser categorizada em dois tipos principais: fatos e opiniões [4]. No entanto, neste trabalho, nossa fundamentação se baseará em minerar (extrair informação relevante de um montante de dados) esses dois aspectos, mas tratando-os como se fossem um. Mesmo que fatos sejam geralmente caracterizados por palavras específicas, essas também identificam uma orientação positiva ou negativa, mesmo que não sejam necessariamente opiniões. Partiremos do princípio que ambos possam ser processados igualmente.

2.1Opinião

Segundo Das e Chen [5], dentro de um fórum preparado para coletar informações dos pequenos investidores é possível extrair uma opinião global, sumarizada, das opiniões expressadas num conjunto de textos relacionados a um ativo. Então essa técnica pode ser aplicada em outras fontes textuais como notícias, por exemplo [3]. Mineração de opinião é uma disciplina da área da lingüística computacional, onde a preocupação é definir a opinião que um documento expressa sobre tal tópico ou produto ao qual se refere [6,7]. Para tratar textos de notícias sobre investimentos, não basta apenas saber de que empresa se trata, mas também extrair uma informação que represente se esse texto tem um impacto positivo ou negativo para a mesma [8]. Técnicas comumente utilizadas para extrair uma opinião de textos são: *Pointwise Mutual Information* - PMI, Naive Bayes, Maximum Entropy e Support Vector Machines [5,8,9]. Todas são baseadas em métodos estatísticos, trabalham com um *corpora* previamente anotado, com o mesmo intuito: extrair a orientação semântica do texto: positivo, negativo ou neutro. No escopo desse trabalho, a primeira técnica utilizada para compor o cálculo dessa orientação será o PMI., utilizado para medir a relação entre palavras.

2.2Pointwise Mutual Information

PMI é uma medida da área de Teoria da Informação, utilizada para medir a relação entre uma ou mais palavras, dentro de um conjunto de texto: ele compara a probabilidade de encontrar dois itens juntos com as probabilidades de estarem separados, ajudando a identificar uma associação verdadeira [10]. Se duas palavras *ent* e *pal* tem probabilidades P(*ent*) e P(*pal*), essa medida se define pela seguinte equação [11]:

$$PMI(ent, pal) = \log\left(\frac{(P(ent \wedge pal))}{(P(ent)P(pal))}\right) \quad (1)$$

Onde a probabilidade *P* é fornecida pela contagem da ocorrência das palavras no corpus (conjunto de dados) analisado. A probabilidade conjunta entre entidade *ent* e uma dada palavra *pal* é analisada contando-se as ocorrências de ambas as palavras numa janela (intervalo) definida de texto, que pode ser uma distância *d* entre palavras, uma frase, um parágrafo ou até mesmo um documento [9].

Assim, através desse método pode-se estimar quanto uma palavra *i* está relacionada com uma determinada palavra *j*. Pode-se expandir para o cálculo a relação não somente a uma palavra específica, mas a uma classe de palavras, nesse caso, classificadas como positivas e negativas.

Tendo o valor de PMI entre palavras determinadas, pode-se calcular a orientação semântica de sentimento. Sabendo quanto uma entidade *ent* se relaciona com essas duas classes de palavras, efetua-se o cálculo da orientação de sentimento segundo [9]:

$$O(ent) = PMI(ent, positivas) - PMI(ent, negativas) \quad (2)$$

Onde *positivas* e *negativas* são conjuntos de palavras com peso positivo e negativo, respectivamente. A equação 2 nos dá um panorama de quanto uma entidade, dentro de um conjunto de documentos, está relacionada às palavras selecionadas.

3.PROPOSTA

Este artigo propõe um processo para sumarizar opiniões sobre um ativo encontradas em fontes de informação disponíveis na Web. Este processo é composto pelas etapas descritas a seguir:

1. Selecionar, extrair e armazenar conteúdo relevante sobre o mercado nacional a partir de *feeds* (fontes de notícias em padrão RSS[1], formato para distribuição e aquisição de informações entre diversas fontes, como jornais, sites, etc.) pré-cadastrados. A coleção é baseada em portais mais conhecidos sobre finanças no mercado nacional. Cada item do feed tem a descrição, a data de publicação e o link original da notícia. A extração do conteúdo completo da notícia será feita através do próprio *feed*, pelo link da notícia original. Esse conteúdo será processado para que o texto contenha o mínimo de ruído possível: publicidade, por exemplo.
2. Identificar as entidades relevantes nos textos: palavras para empresas, e; adjetivos, verbos e palavras-chave. Os termos representantes das empresas são destacados no texto, através da linguagem de marcação XML, para efeitos de interface. Adjetivos e outras palavras selecionadas são manualmente classificados como positivas (+1), negativas (-1) e neutras (0), e são catalogadas no banco de dados. Esse processo também vai identificar e destacar essas palavras, de forma análoga ao destaque feito para empresas.
3. Após as palavras serem detectadas, filtram-se as notícias relacionadas com as empresas do mercado que é de interesse do usuário. Através da lista de empresas pré-definidas, filtram-se apenas as notícias que citam tais empresas. Essa lista envolve a razão social das empresas, códigos de ativos e possíveis outros nomes pelas quais a empresa pode ser referida. Todas essas informações são usadas como sinônimos. A lista de interesse utilizada neste trabalho é composta por empresas listadas no grupo N1 de Governança Corporativa Bovespa [12]. Figurar nesse grupo significa divulgar um certo nível de dados sobre a saúde financeira da empresa, o que nivela as empresas no quesito informação disponível.
4. Calcula-se o impacto das notícias, se positivo ou negativo, em relação à(s) empresa(s) de interesse, utilizando a equação 2.
5. Ao término de todos os cálculos, é apresentada uma avaliação quantitativa sobre o impacto da notícia para a

[1] Really Simple Syndication, http://news.yahoo.com/rss

entidade e uma sumarização do impacto de todas notícias dentro de um bloco definido (período) para cada uma das entidades reconhecidas nesse bloco.

4.IMPLEMENTAÇÃO

Para validar a proposta descrita neste trabalho foram realizados experimentos implementados utilizando a linguagem de programação **Python**[2]. Para a extração dos dados partindo dos feeds RSS, foi utilizado o parser HTML/XML **Beautiful Soup**[3]. O armazenamento dos dados é feito numa base de dados **MySQL**[4], para catalogar os feeds que serão processados, é utilizado o padrão de marcação **OPML**[5]. A coleção de feeds é composta por 10 portais de investimentos, que são os mais conhecidos do mercado nacional e com maior popularidade. Essa coleção gera em média 350 notícias diárias.

Cada item (notícia) do *feed* é armazenado no banco, seu link original visitado, utilizando a **UrlLib2**[6]**,** e seu conteúdo HTML é manipulado para extrair somente o texto (imagens, scripts, estilos e links são ignorados). Armazena-se junto o conteúdo extraído da página destino do link presente no *feed*, em formato texto puro. É feita uma comparação linha-a-linha entre as notícias do mesmo portal a fim de limpar o conteúdo não referente a notícia em si (texto de publicidade, menus, etc.).

Foi criado um dicionário, com palavras extraídas da leitura de uma parte do mesmo conjunto de notícias utilizado ao longo do trabalho. Essas palavras são armazenadas no banco, junto com seu valor (1 para positivas, -1 para negativas e 0 para neutras).

A seguir, o texto armazenado é processado, identificando e marcando as palavras reconhecidas. Essa identificação é feita no texto através de marcação (separando palavras identificadoras de empresas e de palavras com pesos). Além disso, são mantidas estruturas de dados para palavra, sentença, bloco (conjunto de sentenças) e Notícia, onde são armazenados também os valores de peso para cada estrutura. Todas são implementadas de modo que seja possível acessar facilmente tanto o valor individual quanto um valor agregado num conjunto.

Após esse processo, monta-se um dicionário contendo as palavras do texto e a contagem de ocorrência de cada uma dentro do conjunto de notícias, para efetuar os processos relacionados a estatística dentro dos textos. Cada classe de objetos (sentenças e blocos) também permite acesso a um dicionário de suas palavras.

Seguindo as idéias expostas por Bing em [4], a partir desse ponto é calculada a orientação das notícias utilizando três intervalos distintos de texto: sentença, bloco e notícia. Identificadas as entidades no conjunto de textos, aplica-se a cada uma delas o cálculo descrito pela equação 2. Assim, tem-se dentro de um conjunto de notícias, a orientação semântica de sentimento diária para cada empresa, que será comparada diretamente com o desempenho diário (variação do valor) das ações na bolsa.

5.RESULTADOS PARCIAIS

Aplicou-se essa implementação num montante de notícias coletadas diariamente, no intervalo de 1º de Julho à 22 de Agosto de 2008, sua maioria durante o horário do pregão (das 10h as 17h). Para cada dia nesse intervalo, foram processadas as notícias como um conjunto único. Na tabela 1 é possível visualizar alguns exemplos da quantidade de informação diária tratada.

Tabela 1. Alguns sumários do Processamento Diário

Data	Notícias	Entidades	Sentenças
22/06/2008	321	37	16709
08/07/2008	343	31	14995
06/08/2008	374	44	18903
20/08/2008	404	36	21891

Na Tabela 2 vemos um demonstrativo do cálculo de orientação (descrita na equação 2) para cada entidade detectada no conjunto de notícias diário, comparando o valor obtido variando a janela de co-ocorrência das entidades/palavras. Valores positivos indicam maior relação com palavras positivas, enquanto valores negativos indicam maior relação com palavras negativas.

Tabela 2. O(*ent*) para algumas entidades no dia 06/08

Entidade	Sentença	Bloco	Notícia
RAPT4	5,22	5,82	-0,99
Bradesco	0,73	0,76	-0,58
VALE3	0,00	5,41	-0,78
Eletrobrás	-6,69	-7,29	-0,78
GGBR4	4,12	4,72	-0,60

Analisando os resultados e comparando o conteúdo das notícias, é possível verificar que o cálculo de orientação proposto por [9] pode realmente estimar bem o relacionamento entre palavras, mas a escolha das palavras e seus pesos (-1, 0, 1) podem impactar no resultado. A quantidade de palavras catalogadas também pode influenciar o resultado dos cálculos. Na figura 1 é possível visualizar um gráfico com as orientações em conjunto com a variação do ativo no mesmo período.

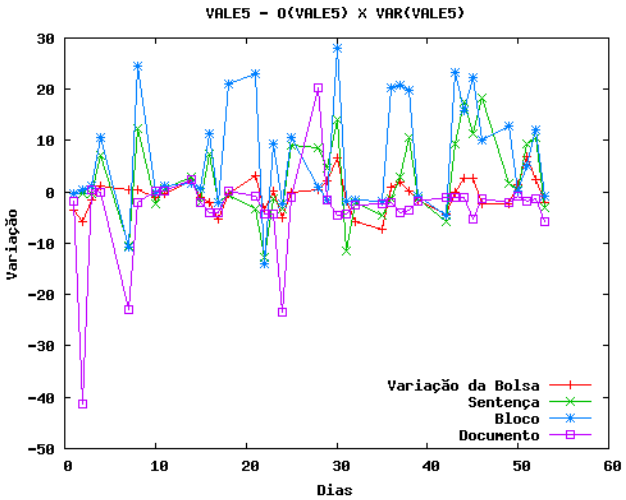

Figura 1: Gráfico com valores de O(VALE) e Variação do Ativo VALE3

[2] http://www.python.org/

[3] http://www.crummy.com/software/BeautifulSoup

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

[5] http://www.opml.org

[6] http://docs.python.org/lib/module-urllib2.html

Além da comparação visual feita com os gráficos, foi aplicada uma análise utilizando a correlação de *Pearson* entre os valores de orientação calculados e a variação dos ativos. Os resultados desta análise de correlação podem ser vistos na tabela 3.

Tabela 3. Correlação entre os valores de orientação e a variação dos ativos

Entidade	Sentença	Bloco	Notícia
GGBR4	0,56	0,66	0,24
USIM5	0,56	0,43	0,10
VALE5	0,52	0,54	0,31
BBDC4	0,32	0,23	0,09
BRKM6	0,31	0,29	0,05
ELET6	0,18	0,19	-0,09
DURA3	0,13	0,15	0,04
BRTP4	0,12	0,00	0,21
SDIA4	0,02	0,32	0,32
ITAU4	0,02	0,21	0,09
CESP6	-0,02	0,19	-0,19
UBBR4	-0,11	0,06	0,02
VGOR4	-0,11	-0,22	0,15

6.CONSIDERAÇÕES FINAIS

A padronização dos formatos de divulgação de notícias, como o RSS, torna possível a aplicação de nossa proposta. Possivelmente, pode torná-la aplicável em outras áreas que façam uso da divulgação de notícias ou conteúdo por esses meios.

O método PMI, utilizado para calcular a orientação semântica, fornece uma estimativa válida, mas desde que o conjunto de textos a ser analisado e a lista de palavras selecionadas tenha um tamanho razoável. Ainda é necessário realizar mais experimentos para definir um tamanho mínimo do *corpora* e do conjunto de palavras selecionadas.

Calcular a orientação apenas sobre um dos sinônimos de uma empresa não é tão eficaz. Por exemplo, a empresa *Vale* é muitas vezes referenciada também apenas como um de seus códigos de ativo (i.e., VALE3 e VALE5), então ocorre que tais palavras, mesmo se referindo a mesma empresa, tenham um valor de orientação diferente. A melhor forma de consolidar isso é somar os valores dos identificadores que representam a mesma empresa para então usar o valor agregado na comparação com a variação do valor do ativo.

O cálculo de correlação não retornou valores conclusivos, mas consideráveis para algumas entidades, demonstrando que realmente há um nível relevante de correlação entre a variação do ativo e os valores obtidos no cálculo de orientação. Talvez um ajuste no intervalo de notícias analisadas (sub-períodos, i.e., semanas) estabeleça correlações mais fortes.

Entidades mais visadas na bolsa possuem maior valor de correlação. Isso pode indicar também uma relação da quantidade de notícias com o volume negociado dessas entidades (e conseqüentemente, um retorno mais preciso no cálculo de orientação), que deverá ser melhor pesquisado.

A janela escolhida para as co-ocorrências também pode variar e influenciar o resultado. Parece ainda não ser interessante usar apenas uma (i.e., sentença), mas sim um comparativo entre todos, até que seja encontrada uma janela ideal para o escopo do trabalho e domínio da aplicação. Também será importante pesquisar intervalos maiores e outros aspectos dos ativos, como por exemplo: uma semana ao invés de apenas um dia, e valores de fechamento, mínimo e máximo além da variação.

7.REFERÊNCIAS

[1] Markowitz, H. (1997). Portfolio Selection. 6a reimpressão da 2a edição. Massachusetts: Blackwell.

[2] Kothari, S. P., Jerold B. Warner. (2006). The econometrics of event studies, Handbook of Corporate Finance: Empirical Corporate Finance (Elsevier/North-Holland).

[3] Oliveira, Miguel Delmar de. (1986). Introdução ao Mercado de Ações. Rio de Janeiro: Comissão Nacional de Bolsas de Valores - CNBV, 1986.

[4] Liu, Bing (2006). Web Data Mining, Exploring Hyperlinks, Contents and Usage Data, Springer.

[5] Sanjiv R. Das, Mike Y. Chen (2007). Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web, MANAGEMENT SCIENCE, Vol. 53, No. 9, September 2007, pp. 1375–1388.

[6] Esuli, Andrea. (2006). Opinion Mining. Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche, Pisa, Italy.

[7] Tatemura, Junichi. (2000). Virtual reviewers for collaborative exploration of movie reviews. In Proc. of the 5th International Conference on Intelligent User Interfaces, pp. 272-275.

[8] Pang, Bo; Lee, Lillian and Vaithyanathan, Shivakumar. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques, in Proceedings of EMNLP, 2002 pp. 79-86.

[9] Turney, Peter D. (2002). Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. Institute for Information Technology National Research Council of Canada

[10] Thomas, M. Cover. and Joy A. Thomas (1991). Elements of Information Theory. (John Wiley).

[11] Church, K. W. and Hanks, P (1989). Word Association Norms, Mutual Information and Lexicography. Proceedings of the 26th Annual Conference of the Association for Computational Linguistics.

[12] Bovespa. (2006). Nível 1 – Bovespa Brasil, Bolsa de Valores do Estado de São Paulo, (Folder explicativo).