

Revisando o Método de Análise da Semântica Latente para Propósitos de Mineração de Opiniões sobre Produtos

Wilson Pires Gavião Neto¹, Sidnei Renato Silveira¹

¹Sistemas de Informação: Ciência e Tecnologia Aplicadas
Centro Universitário Ritter dos Reis

wgaviao@gmail.com, sidnei@uniritter.edu.br

Resumo. *Este trabalho insere-se em um projeto cujo objetivo é tirar proveito de avaliações numéricas de itens de produtos para melhorar a detecção de padrões em opiniões expressas textualmente. Especificamente apresenta-se uma variação de Análise da Semântica Latente que busca contemplar objetivos de detecção de padrões em um conjunto de opiniões de consumidores. Resultados preliminares revelam um grande potencial neste sentido.*

1. Introdução

Facilidades e segurança são alguns dos motivos que levam ao grande crescimento do comércio eletrônico nos últimos anos. Como forma de fidelizar clientes, as empresas buscam constantemente proporcionar melhores experiências de compra aos consumidores bem como aumentar sua satisfação com os produtos adquiridos. Neste contexto, já é uma prática comum na internet a oferta de serviços que possibilitam aos consumidores expressarem suas opiniões sobre produtos comprados. Com mais e mais usuários tornando-se confortáveis com a Web, uma quantidade crescente de pessoas está postando seus comentários online. Como resultado, o número de comentários/críticas que um produto recebe cresce rapidamente. Produtos mais populares podem receber centenas de comentários, como pode-se verificar em sites como www.amazon.com. Além disso, muitos dos comentários postados constituem-se em textos longos com poucas frases contendo opiniões sobre o produto. Neste contexto, torna-se difícil para um potencial consumidor ler e formar uma decisão sobre qual produto comprar. Pelo lado da empresa, também torna-se difícil o monitoramento de opiniões postadas pelos consumidores de seus produtos.

O cenário descrito acima constitui-se em uma das motivações para o surgimento de uma linha de estudos conhecida como *mineração de opiniões* (HU; LIU, 2004). Três campos de pesquisa predominam na área de mineração de opiniões (BODENDORF; KAISER, 2010):

Orientação da opinião ou análise do sentimento: A partir de um conjunto de documentos, a análise de sentimento procura classificá-los de acordo com a orientação das opiniões, como por exemplo, opiniões negativas e positivas (TURNER, 2002; PANG; LEE, 2008).

Mineração baseada em atributos de produtos: Esta linha de estudo considera opiniões sobre característica/itens de produtos/objetos em sentenças (POPESCU; ETZIONI, 2005).

Mineração de opiniões comparativas: Neste contexto, busca-se desenvolver métodos para identificar sentenças comparativas em opiniões sobre produtos/objetos bem como a preferência expressa no texto (GANAPATHIBHOTLA; LIU, 2008).

Este trabalho insere-se em um projeto cujo objetivo é tirar proveito de avaliações numéricas de itens de produtos para melhorar a detecção de padrões em opiniões expressas textualmente. A Figura 1 mostra um exemplo de um site em que proprietários de automóveis expressam opiniões textuais sobre seus carros, associando, ainda, notas para itens predeterminados dos veículos. Neste contexto, este artigo discute potencialidades do método de *Análise de Semântica Latente* (ASL) (DEERWESTER et al., 1990) para propósitos de identificar padrões em um conjunto de depoimentos sobre um produto. Busca-se com isso minimizar o esforço na atividade de analisar todos os depoimentos para se ter uma idéia sobre prós e contras de um produto. Deve-se ainda observar a (razoável) suposição feita pela proposta de sumarização de opiniões deste trabalho: "Características freqüentemente comentadas por consumidores refletem algo relevante sobre o produto".

"Bom desempenho mas tem seu preço"	
Volkswagen Gol G5 1.0 2010	
24/8/2010 18:50:00	
Luiz	
Caçapava SP	
Dono há menos de 1 ano	
Prós:	
melhor desempenho da categoria, design, estabilidade, posição de dirigir confortável, peças de reposição mais baratas que similares, melhor preço e rapidez na revenda.	
Contras:	
bebe muito para 1.0, contudo tem as peças de reposição mais baratas. Ruidos no painel.com ar ele sofre um pouco	
Defeitos apresentados:	
nenhum	
Opinião Geral:	
bebe bastante para 1.0 , é o preço de ter 76cv , ganha em todos comparativos da concorrência até do uno 1.4 , contudo tem as peças de reposição e seguro mais baratos que o gol 1.6 o que acaba equilibrando.excelente carro!!! o melhor Gol de todos os tempos!!recomendo	
Estilo	★★★★★ 10
Acabamento	★★★★★ 9
Posição de dirigir	★★★★★ 9
Instrumentos	★★★★★ 9
Interior	★★★★★ 8
Porta malas	★★★★★ 7
Desempenho	★★★★★ 9
Motor	★★★★★ 10
Câmbio	★★★★★ 10
Freios	★★★★★ 10
Suspensão	★★★★★ 10
Consumo	★★★★★ 6
Estabilidade	★★★★★ 9
Custo/Benefício	★★★★★ 8
Recomendação	★★★★★ 9
Avaliação Geral	★★★★★ 8,87

Figura 1. Exemplo de comentários/depoimentos onde consumidores expressam sua opinião não só em termos textuais mas também atribuindo notas a determinados itens do produto. Especificamente, trata-se do site Carros na Web (<http://www.carrosnaweb.com.br/opinioao.asp>), onde proprietários de veículos relatam suas experiências.

O restante deste artigo está organizado como segue. A seção 2 apresenta conceitos que fundamentam a ASL. Na seção 3 discute-se a ASL para propósitos de mineração de opiniões. A seção 4 apresenta o *setup* dos experimentos realizados além de resultados preliminares. Por fim, a seção 5 apresenta conclusões parciais e trabalhos futuros

2. Análise de Semântica Latente - ASL

Considere o problema onde um usuário deseja recuperar documentos/textos em bases conceituais, sendo que palavras individuais não provêm evidências confiáveis sobre os tópicos discutidos nos documentos. Usualmente há várias maneiras de expressar em palavras um dado conceito. Deste modo, termos literais presentes em uma consulta de usuário pode não retornar documentos relevantes. Adicionalmente, a maioria das palavras possuem múltiplos significados, fazendo com que termos em uma consulta permitam, de fato, que documentos não relevantes sejam recuperados. O técnica de Análise de Semântica Latente explora a relação existente entre termos e os textos nos quais eles aparecem para construir um espaço vetorial onde similaridades de significado podem ser estabelecidas. Este novo espaço é conhecido como *Espaço Conceito* ou *Espaço Semântico* (DEERWESTER et al., 1990), sendo que a proximidade entre significados é proporcional ao ângulo entre vetores neste espaço (BERRY; DRMAC; JESSUP, 1999).

Basicamente, a ASL consiste na construção de uma matriz **A** que informa a co-ocorrência de termos e documentos (*termos* \times *documentos*). Após, a matriz **A** é decomposta algebricamente segundo a decomposição em valores singulares (SVD) como forma de aproximar a matriz **A** por combinações lineares (BERRY; DRMAC; JESSUP, 1999).

Formalmente, seja **A** uma matriz onde o elemento (i, j) descreve a ocorrência do termo i no documento j (por exemplo, a frequência de i em j). Se **A** tem dimensões $m \times n$, onde m é o número de termos e n é a quantidade de documentos, a SVD de **A** é definida como:

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^T \quad (1)$$

Onde **U** = $[u_{ij}]$ é uma matriz ortonormal $m \times m$ cujas colunas são chamadas de vetores singulares a esquerda; **D** = $diag(\sigma_1, \sigma_2, \dots, \sigma_n)$ é uma matriz diagonal $m \times n$ cujos elementos são chamados de valores singulares não negativos, os quais aparecem ordenados de forma decrescente; e **V** = $[v_{ij}]$ é uma matriz ortonormal $n \times n$ cujas colunas são chamadas de vetores singulares a direita.

Se $posto(\mathbf{A}) = k$ a SVD pode ser interpretada como o mapeamento do espaço de **A** em um espaço conceito (reduzido) de k dimensões, as quais são linearmente independentes.

Ainda que $posto(\mathbf{A})$ seja maior que k , quando seleciona-se apenas os k maiores valores singulares $(\sigma_1, \sigma_2, \dots, \sigma_k)$ e seus correspondentes vetores singulares em **U** e **V**, pode-se obter uma aproximação de posto k para a matriz **A** (GONG; LIU, 2001). Deste modo, pode-se tratar vetores de termos e documentos como um *espaço conceito*. Neste novo espaço, os vetores de termos em **U** tem k entradas, cada um dando a ocorrência do termo i em um dos k conceitos. Da mesma forma, os vetores de documentos em **V** revelam a relação entre o documento j com cada conceito k . Usualmente formaliza-se o espaço conceito como

$$\mathbf{A}_k = \mathbf{U}_k \mathbf{D}_k \mathbf{V}_k^T \quad (2)$$

Deste modo, pode-se então verificar em \mathbf{V}_k o quão relacionados estão dois dados documentos j e l (usualmente utilizando-se a distância de cossenos (BERRY; DRMAC; JESSUP, 1999)), possibilitando a aplicação de técnicas de *clustering* de documentos no espaço conceito.

3. Análise e Proposições

Como apresentado na seção 2, a matriz de termos-documentos \mathbf{A} carrega a informação sobre a ocorrência de termos em documentos. Geralmente emprega-se uma matriz de pesos sobre \mathbf{A} de maneira a caracterizar a presença de um termo i em um documento j . Usualmente emprega-se uma matriz de pesos conhecida como *TF-IDF* (*term frequency - inverse document frequency*).

Como parte da *TF-IDF* e com objetivo de medir a importância de um termo i em um documento j , emprega-se a frequência normalizada de um termo i em j , como segue:

$$TF_{i,j} = \frac{\#_{i,j}}{\sum_k \#_{k,j}}, \quad (3)$$

onde, $\#_{i,j}$ é a quantidade de ocorrências do termo i no documento j , sendo o denominador a soma das ocorrências de todos os termos no documento j .

Neste sentido propõe-se não empregar *TF-IDF* como matriz de pesos no contexto deste trabalho. Uma vez que trata-se de opiniões expressas por consumidores de forma direta e orientada por itens predefinidos, é razoável assumir que, uma vez citado, o item já torna-se relevante dentro da opinião, não necessitando aparecer inúmeras vezes dentro do mesmo depoimento para adquirir maior importância. Além disso, busca-se por padrões em opiniões, sendo assim, se dois consumidores citam os mesmos itens (dentro de, por exemplo, uma seção de "defeitos apresentados"), é desejável que o sistema retorne seus depoimentos como sendo altamente similares.

Conforme discussão acima, uma vez que o termo i aparece no documento j , propõe-se que o elemento correspondente de \mathbf{A} , isto é a_{ij} , assuma o valor da importância do termo i no contexto de todos os depoimentos. Assim, emprega-se nos experimentos apenas a frequência do termo i considerando todos os n depoimentos:

$$TO_{i,j} = \frac{\sum_{j=0}^n \#_{i,j}}{n}. \quad (4)$$

4. Experimentos e resultados preliminares

Os dados utilizados nestes experimentos são oriundos de 603 opiniões de proprietários de automóveis extraídas do site Carros na Web (www.carrosnaweb.com.br/opiniao.asp). As opiniões estão dispostas na web como mostrado na Figura 1, sendo que utilizou-se apenas o conteúdo das tags "Contras" e "Defeitos apresentados".



Com o objetivo de preparação dos dados (por exemplo, remover sufixos de termos) empregou-se o algoritmo proposto em (ORENGO; HUYCK, 2001). A Figura 2 ilustra graficamente o conjunto de termos extraído das 603 opiniões, onde o tamanho dos termos é proporcional a quantidade de vezes que estes são mencionados.

Considerando uma quantidade de dimensões $k = 18$ conceitos para a aproximação da matriz **A** via SVD, ilustra-se na Figura 3 o espaço conceito gerado de documentos **V**, considerando as dimensões conceito 2,3 e 4. Pode-se verificar a presença de *clusters* de documentos, evidenciando a existência de padrões nas opiniões.

Dentro de um mesmo *cluster* verificou-se a presença de opiniões como as que seguem:

Opinião 1: *"apesar de previsível, o consumo no início assusta um pouco. por ser automático leva um certo tempo para encontrar a pressão ideal no pedal. nos primeiros meses, cheguei a fazer 5,5 km/l, agora, já acostumado e conhecendo o carro, na cidade faço 7,8 a 8 km/loutro ponto fraco é o preço das peças."*

Opinião 2: *"bebe que só um alcoólatra. podem ver. o megane 2.0 16v é mais econômico q ele. vc economiza na compra, mas gasta o mês inteiro com ele bebendo."*

Verifica-se acima um grande potencial da metodologia ASL, tendo vista que as opiniões acima possuem poucos termos em comum mas tratam sobre consumo de combustível, e foram agrupadas no mesmo *cluster*.

5. Conclusões e trabalhos futuros

Apresentou-se neste trabalho uma variação da abordagem de Análise da Semântica Latente com vistas a detectar padrões em opiniões de consumidores sobre itens de produtos. De acordo com os, ainda preliminares, experimentos realizados, verificou-se um grande potencial da abordagem discutida, uma vez que constatou-se claras estruturas de *cluster* (potenciais padrões) bem como uma surpreendente habilidade de agrupar opiniões com raros termos em comum, mas que discutem o mesmo tópico. Como trabalhos futuros, experimentos mais detalhados serão realizados no sentido de validar as limitações da abordagem proposta.

Referências

BERRY, M. W.; DRMAC, Z.; JESSUP, E. R. Matrices, Vector Spaces, and Information Retrieval. **SIAM Rev.**, Philadelphia, PA, USA, v.41, n.2, p.335–362, 1999.

BODENDORF, F.; KAISER, C. Mining Customer Opinions on the Internet - A Case Study in the Automotive Industry. **International Workshop on Knowledge Discovery and Data Mining**, Los Alamitos, CA, USA, v.0, p.24–27, 2010.

DEERWESTER, S. et al. Indexing by latent semantic analysis. **JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE**, [S.l.], v.41, n.6, p.391–407, 1990.

GANAPATHIBHOTLA, M.; LIU, B. Mining opinions in comparative sentences. In: COLING '08: PROCEEDINGS OF THE 22ND INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS, Morristown, NJ, USA. **Anais...** Association for Computational Linguistics, 2008. p.241–248.

GONG, Y.; LIU, X. Creating Generic Text Summaries. **Document Analysis and Recognition, International Conference on**, Los Alamitos, CA, USA, v.0,

p.0903, 2001.

HU, M.; LIU, B. Mining and summarizing customer reviews. In: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, New York, NY, USA. **Proceedings...** ACM, 2004. p.168–177.

ORENGO, V.; HUYCK, C. A Stemming Algorithm for the Portuguese Language. **String Processing and Information Retrieval, International Symposium on**, Los Alamitos, CA, USA, v.0, p.0186, 2001.

PANG, B.; LEE, L. Opinion Mining and Sentiment Analysis. **Found. Trends Inf. Retr.**, Hanover, MA, USA, v.2, n.1-2, p.1–135, 2008.

POPESCU, A.-M.; ETZIONI, O. Extracting product features and opinions from reviews. In: HUMAN LANGUAGE TECHNOLOGY AND EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, Morristown, NJ, USA. **Proceedings...** Association for Computational Linguistics, 2005. p.339–346.

TURNEY, P. D. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In: ANNUAL MEETING ON ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 40., Morristown, NJ, USA. **Proceedings...** Association for Computational Linguistics, 2002. p.417–424.