Aplicação da lógica nebulosa para mineração de opiniões

Matheus Cardoso de A. Silva
matheus.mcas@gmail.com

Orientador: Prof. Dr. Angelo Loula
angelocl@ecomp.uefs.br

Co-orientador: Prof. Dr. Matheus Pires
mgpires@ecomp.uefs.br

Mestrado

Mestrado em Ciência da Computação - UFBA/UEFS

Ano/semestre de ingresso: 2012.2

Época esperada de conclusão: 2014.1

Etapas concluídas: Exame de qualificação

Etapas futuras: Defesa em agosto de 2014

Abstract. *Opinions are central in almost all human activities, because they are a relevant influence on people's behavior. The internet and the web have created mechanisms that made possible for people to share their opinions and for other people and organizations to find out more about opinions and experiences from individuals and help in decision making. The quantity and diversity of these sources is increasing and the average reader has difficulty to extract and summarize the existing opinions on these sources. Furthermore, opinions involve sentiments that are vague and inaccurate textual descriptions. A methodology to handle, computationally, vague and inaccurate data is the Fuzzy Logic. Automated opinion mining systems based on fuzzy logic might be useful to handle the opinion mining task of this great amount and diversity of sources. This project aims to evaluate the fuzzy logic application to these systems and proposes a fuzzy approach to opinion mining.*

Palavras-chave: mineração de opinião; sentimento, lógica nebulosa

Feira de Santana

Abril 2014

1. Problema de pesquisa e caracterização da contribuição

As opiniões são as principais influenciadoras do comportamento humano e permeiam quase todas as atividades executadas no dia-a-dia pelas pessoas [Liu 2012]. É comum as pessoas pedirem opiniões a familiares ou amigos, por exemplo, sobre qual marca de carro escolher numa compra ou sobre hotéis em que querem se hospedar. E saber a opinião de outrem é importante também para empresas [Liu 2012, Pang and Lee 2008]. O surgimento da internet e o advento da web criaram um novo espaço para que pessoas e organizações pudessem descobrir mais sobre as opiniões e experiências de outras pessoas, sejam elas do próprio círculo social ou indivíduos completamente desconhecidos.

Minerar opiniões na web, contudo, não é uma tarefa simples. A quantidade e a diversidade de fontes é muito grande [Kim et al. 2006], e cada uma delas possui muitas informações opinativas, com formatos diferentes e problemas de sintaxe (e.g. erros de grafia, concordância verbal, nominal). Com isso, o leitor comum da internet tem dificuldades de extrair e resumir as opiniões existentes nessas fontes. O trabalho realizado por [Horrigan 2008] corrobora essas dificuldades, relatando que 58% das pessoas que acessaram a web para procurar opiniões acharam que as informações estavam perdidas, algumas impossíveis de serem encontradas, confusas e/ou numerosas.

Além desses problemas, a tarefa de minerar opiniões se torna mais difícil quando as opiniões são acompanhadas de sentimentos que são, por sua vez, subjetivos e imprecisos. Para identificar sentimentos em frases e documentos, frequentemente é necessário lidar com termos imprecisos e vagos, como ”bom”, ”ruim”, ”péssimo”, ”muito bom”, ”pouco ruim”. Uma metodologia da Inteligência Computacional proposta para tratar computacionalmente dados imprecisos e vagos é a Lógica Nebulosa (*Fuzzy Logic*), introduzida por [Zadeh 1988]. Enquanto na lógica clássica, um elemento pode ser classificado somente como relacionado ou não relacionado a uma categoria, na lógica nebulosa, um elemento pode ser classificado como parte de um ou mais conjuntos ao mesmo tempo, com diferentes graus de pertinência [Zadeh 1988]. Assim, um sentimento de uma opinião de um dado documento, por exemplo, em vez de ser classificado somente como positivo ou negativo, poderá ser classificado, pela lógica nebulosa, como ambos, com diferentes graus de pertinência.

Considerando as dificuldades apresentadas que são enfrentadas pelos usuários e empresas, quando estes necessitam minerar opiniões da web, fica evidente que o uso de sistemas automatizados de mineração de opiniões e a Lógica Nebulosa podem ser úteis em processos de tomada de decisão. Portanto, este trabalho tem como objetivo propor e avaliar o desempenho de um sistema automatizado de mineração de opiniões baseado na lógica nebulosa.

2. Fundamentação teórica e trabalhos relacionados na área

2.1. Fundamentação teórica

Segundo [Liu 2012], mineração de opinião é o campo de estudo que analisa as opiniões, sentimentos, avaliações, atitudes e emoções de pessoas direcionadas a entidades ou alvos, como produtos, serviços, organizações, indivíduos, problemas, eventos, tópicos e seus atributos. É uma área que vem sendo investigada,principalmente, em três níveis: i) análise de documento, ii) sentenças e iii) entidades e seus aspectos. O primeiro nível foca em

classificar opiniões de um documento como positivas ou negativas. O segundo, classifica as opiniões de cada sentença separadamente e o último nível foca em descobrir todos os alvos existentes em sentenças e classificar as opiniões direcionadas a aqueles alvos [Liu 2012].

Seja para minerar opiniões em documentos ou sentenças, a mineração de opinião precisa lidar com termos vagos e imprecisos. A Lógica Nebulosa, proposta por [Zadeh 1965], é uma metodologia da Inteligência Computacional criada para tratar esses tipos de informações. Estas são modeladas por meio de conjuntos chamados Conjuntos Nebulosos [Zadeh 1965]. Tais conjuntos permitem que um objeto pertença a um ou mais conjuntos ao mesmo tempo, mas com graus de pertinência para cada um deles. Assim, uma opinião de um documento ou sentença pode ser, ao mesmo tempo, positiva e negativa, mas com graus de positividade e negatividade diferentes.

2.2. Trabalhos relacionados

Um dos trabalhos mais citados sobre a aplicação de lógica nebulosa em mineração de opinião é o de [Andreevskaia and Bergler 2006]. Foi um dos primeiros estudos a utilizarem o conceito de conjuntos nebulosos para a mineração de opinião. Neste estudo, os autores apresentam um algoritmo de extração de adjetivos (STEP, do inglês *Sentiment Tag Extraction Program*) do Wordnet [1] e os reclassifica (positivamente, negativamente ou de maneira neutra) utilizando o conceito de conjuntos nebulosos. Por fim, o trabalho discute as contribuições realizadas pela pesquisa e, sobre lógica nebulosa, aponta a importância do conceito de centralidade derivado do grau de pertinência dos conjuntos nebulosos. Não houveram, todavia, resultados apresentados que remetessem os ganhos ao uso de lógica nebulosa.

Em [Ohana and Tierney 2009] é abordado o problema da anotação de valores para palavras que denotam opiniões e sentimentos, como adjetivos e advérbios. Trabalhos como [Pang et al. 2002] e [Kennedy and Inkpen 2006] utilizaram anotação manual para esses e outros tipos de palavras (e.g. verbos e substantivos). [Ohana and Tierney 2009] apresenta um comparativo de abordagens nebulosas de mineração de opinião utilizando um dicionário com termos anotados, o SentiWordNet [Esuli and Sebastiani 2006]. Este dicionário contém informação opinativa (e.g. "good" tem 0.85 grau de positividade e 0.15 de negatividade) sobre termos extraídos do WordNet. Os resultados mostraram que o uso do SentiWordNet, com as abordagens usadas, se aproximam e ultrapassam os resultados obtidos pela anotação manual realizada em [Pang et al. 2002] e [Kennedy and Inkpen 2006].

O trabalho realizado em [Khan 2011] também utiliza o SentiWordNet para atribuir valores numéricos para termos opinativos. Contudo, neste trabalho, a analise é feita sobre as sentenças dos documentos. Neste estudo, as sentenças são (i) classificadas em objetivas e subjetivas; (ii) o valor semântico das palavras são extraídas do SentiWordNet e (iii) cada sentença é classificada baseando-se em regras de estruturas contextuais de sentenças e utilizando as informações das palavras extraídas do SentiWordNet. Os resultados obtidos (83% de precisão média) superaram outros trabalhos já realizados na literatura como em [Go et al. 2009], [Andreevskaia and Bergler 2008] e [Hu and Liu 2004].

[1] Um dicionário de palavras que as relaciona em conjuntos de sinônimos. Disponível em: http://wordnet.princeton.edu/

3. Estado atual do trabalho

As etapas de revisão da literatura e levantamento sobre lógica nebulosa e classificação nebulosa já foram realizadas. O processo de mineração de opinião também já foi definido, com base em [Mouthami et al. 2013], [Subasic and Huettner 2001] e [Moraes et al. 2012], o qual é constituído pelas seguintes etapas: definição do domínio, pré-processamento, transformação, seleção de características, classificação, e análise. Em relação ao domínio de aplicação, está sendo utilizada a base de filmes da Universidade de Cornell, introduzida em [Pang and Lee 2004] e largamente utilizada na área [2].

A pesquisa se encontra na fase de avaliação de sistemas nebulosos para mineração de opinião e também na definição de seus componentes. O SentiWordNet, utilizado em [Ohana and Tierney 2009] e [Khan 2011], está sendo avaliado para ser utilizado na etapa de transformação do processo de mineração de opinião. O SentiWordNet é um dicionário que associa três pontuações numéricas a cada conjunto de sinônimos de outro dicionário, o WordNet. Essas pontuações descrevem os graus de objetividade, positividade e negatividade dos termos de um conjunto de sinônimos [Esuli and Sebastiani 2006]. Por exemplo, o conjunto de sinônimos para ”estimable”, quando esta palavra corresponde à medição, possui pontuação 1 para objetividade e 0 para positividade e negatividade. Por outro lado, quando ”estimable” refere-se a respeito ou consideração, possui grau 0.75 de positividade, 0 de negatividade e 0.25 de objetividade.

Esta avaliação está sendo feita, devido ao problema de se determinar a polaridade de um termo textual (positivo ou negativo) e quão forte é essa polaridade (e.g. fracamente positivo, fortemente negativo). Há diferentes abordagens nos trabalhos relacionados a esta pesquisa, como em [Turney 2002], [Nadali et al. 2010], [Poria et al. 2012] e [Jusoh and Alfawareh 2013]. Contudo, estas pesquisas são dependentes do domínio (os percentuais de precisão são diferentes para filmes e automóveis, por exemplo) e utilizam anotações humanas, sujeitas a erros de julgamento, para atribuir graus de positividade, negatividade e objetividade para termos textuais, além do tempo maior despendido por usar mão-de-obra humana. O uso do SentiWordNet perpassa esses problemas.

Os resultados obtidos até agora, corroboram com os relatados em [Ohana and Tierney 2009] e [Khan 2011], os quais comprovaram que o uso do SentiWordNet pode produzir resultados iguais ou melhores que a utilização de anotações humanas ou outras abordagens da literatura e independente do domínio escolhido.

4. Desenvolvimento necessário para a conclusão

O próximo passo dessa pesquisa é avaliar os termos textuais mais relevantes num texto opinativo e como utiliza-los. É sabido que adjetivos possuem maior carga opinativa num documento [Hu and Liu 2004] [Esuli and Sebastiani 2006] [Jusoh and Alfawareh 2013] [Nadali et al. 2010], mas que o uso isolado destes não produz bons resultados [Wiebe and Mihalcea 2006]. Essa definição impacta desde a fase de pré-processamento até a classificação.

Após isso, é necessário definir como o SentiWordNet (ou outro dicionário) será utilizado na fase de transformação. [Ohana and Tierney 2009], por exemplo, utiliza somente os graus associados aos termos dos conjuntos de sinônimos. [Khan 2011], por

[2]Trabalhos que utilizaram a base da Cornell: http://www.cs.cornell.edu/people/pabo/movie-review-data/otherexperiments.html

outro lado, utiliza as definições dos conjuntos de sinônimos na tentativa de resolver problemas de ambiguidade dos termos.

Outra tarefa é pesquisar e avaliar a viabilidade do uso de técnicas redução de dimensionalidade na etapa de seleção de características do processo de mineração de opinião. Essa etapa pode fazer como que o sistema de mineração torne-se mais eficiente, reduzindo a quantidade de dados a serem analisados e selecionando os de mais relevância para a análise [Moraes et al. 2012].

Por fim, definir os componentes do sistema nebuloso (conjuntos nebulosos, regras, método de inferência, etc.) que serão usados na etapa de classificação e avaliar os resultados frente a outros trabalhos que utilizaram a mesma base de dados.

5. Avaliação dos resultados

Como a pesquisa se encontra na fase de avaliação de sistemas nebulosos para mineração de opinião e na definição de seus componentes, os resultados finais deste trabalho ainda não foram obtidos.

References

Andreevskaia, A. and Bergler, S. (2006). Mining wordnet for a fuzzy sentiment: Sentiment tag extraction from wordnet glosses. In *EACL*, pages 209–216.

Andreevskaia, A. and Bergler, S. (2008). When specialists and generalists work together: Overcoming domain dependence in sentiment tagging. In *ACL*, pages 290–298.

Esuli, A. and Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, volume 6, pages 417–422.

Go, A., Bhayani, R., and Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, pages 1–12.

Horrigan, J. A. (2008). Online shopping. *Pew Internet & American Life Project Report*, 36.

Hu, M. and Liu, B. (2004). Mining opinion features in customer reviews. In *Proceedings of the 19th national conference on Artifical intelligence*, AAAI'04, pages 755–760. AAAI Press.

Jusoh, S. and Alfawareh, H. M. (2013). Applying fuzzy sets for opinion mining. In *Computer Applications Technology (ICCAT), 2013 International Conference on*, pages 1–5. IEEE.

Kennedy, A. and Inkpen, D. (2006). Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence*, 22(2):110–125.

Khan, A. (2011). Sentiment classification by sentence level semantic orientation using sentiwordnet from online reviews and blogs. *International Journal of Computer Science & Emerging Technologies*, 2(4).

Kim, P., Anderson, E., and Joseph, J. (2006). The forrester wave: Brand monitoring. *Cambridge: Forrester Research*.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167.

Moraes, R., Valiati, J. F., and GaviãO Neto, W. P. (2012). Document-level sentiment classification: An empirical comparison between svm and ann. *Expert Systems with Applications*.

Mouthami, K., Devi, K. N., and Bhaskaran, V. M. (2013). Sentiment analysis and classification based on textual reviews. In *Information Communication and Embedded Systems (ICICES), 2013 International Conference on*, pages 271–276. IEEE.

Nadali, S., Murad, M., and Kadir, R. (2010). Sentiment classification of customer reviews based on fuzzy logic. In *Information Technology (ITSim), 2010 International Symposium in*, volume 2, pages 1037–1044. IEEE.

Ohana, B. and Tierney, B. (2009). Sentiment classification of reviews using sentiwordnet. In *9th. IT & T Conference*, page 13.

Pang, B. and Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, page 271. Association for Computational Linguistics.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.

Poria, S., Gelbukh, A., Cambria, E., Das, D., and Bandyopadhyay, S. (2012). Enriching senticnet polarity scores through semi-supervised fuzzy clustering. In *Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on*, pages 709–716. IEEE.

Subasic, P. and Huettner, A. (2001). Affect analysis of text using fuzzy semantic typing. *Fuzzy Systems, IEEE Transactions on*, 9(4):483–496.

Turney, P. D. (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 417–424. Association for Computational Linguistics.

Wiebe, J. and Mihalcea, R. (2006). Word sense and subjectivity. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 1065–1072. Association for Computational Linguistics.

Zadeh, L. (1988). Fuzzy logic. *Computer*, 21(4):83–93.

Zadeh, L. A. (1965). Fuzzy sets. *Information and control*, 8(3):338–353.