

Classificação de revisões para construção de perfis em sistemas de recomendação

Thiago Fujisaka Tanaka, Marcelo G. Manzato
 Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo
 Av. Trabalhador Sancerlense, 400, Caixa Postal 668 – 13560-970
 São Carlos, SP – Brazil
 tftanaka@grad.icmc.usp.br, mmanzato@icmc.usp.br

RESUMO

Este artigo apresenta uma avaliação de diferentes algoritmos de análise de sentimentos baseados em revisões de filmes criadas por usuários. O principal objetivo é identificar as vantagens e limitações de cada técnica, buscando auxiliar a escolha de uma abordagem a ser utilizada em um projeto de maior escala, que, em síntese, busca construir um mecanismo de criação de perfis de usuários baseado em anotações, como comentários, etiquetas e revisões. Esses perfis irão conter as preferências de usuários sobre diferentes tópicos, e serão utilizados como base para permitir a recomendação de conteúdo adicional de acordo com seus principais interesses.

Categories and Subject Descriptors

H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing—*Indexing methods*

General Terms

Design, Algorithms

Keywords

classificação de sentimentos, construção de perfis, sistemas de recomendação.

1. INTRODUÇÃO

Devido à intensificação no uso da rede mundial de computadores durante os últimos anos, observou-se um grande aumento na quantidade de conteúdo multimídia disponível. Consequentemente, devido a essa sobrecarga de informações, tornou-se necessário um modo no qual usuários possam recuperar os dados conforme seus interesses.

Uma maneira de se realizar a recuperação de conteúdo multimídia é por meio de sistemas de recomendação. A recomendação seleciona automaticamente itens multimídia sem a intervenção do usuário, baseando-se apenas em dados apreciados anteriormente pelo indivíduo, oferecendo outros

conteúdos semelhantes. Isso é feito com base em um perfil de interesses que contém o histórico de itens avaliados anteriormente, de modo a se obter uma lista de preferências do usuário [7].

Dois desafios da recomendação multimídia são a indexação multimídia e a construção do perfil de interesses. No primeiro caso, a dificuldade está relacionada em manter todo o conteúdo indexado por palavras-chave relevantes, que não só representem seu título ou o descreva de maneira genérica, mas que envolva também elementos nele presentes, como locais, objetos, pessoas e situações. Essa indexação pode ser feita de maneira automática ou manual. A indexação automática, que possui a vantagem de ser executada de modo ágil e em grande escala, possui domínios restritos, pois detalhes presentes no conteúdo são inerentes à modalidade multimídia e em quais ambientes está presente. Já a indexação manual apresenta maior custo, pois especialistas são necessários para anotar grandes quantidades de itens multimídia disponíveis a cada dia.

Em relação à construção do perfil de interesses, duas metodologias são propostas pela literatura [10]: as coletas explícita e implícita de dados. Na coleta explícita, o próprio usuário fornece informações, como dados demográficos e avaliação de itens visitados anteriormente (notas e comentários). Uma desvantagem é a imposição de esforço adicional, que além de causar desinteresse, permite o fornecimento de informações incorretas. Já a coleta implícita obtém dados sem sua intervenção, exigindo desenvolvimento e manutenção de ferramentas que possam monitorar comportamentos, gerando altos custos e limitando o usuário à utilização de poucos tipos de dispositivos.

Por outro lado, recentes avanços relacionados à Web 2.0 [14] possibilitaram novas maneiras de se obter descrições sobre o conteúdo e usuário. O termo Web 2.0 define uma plataforma em que usuários produzem conteúdo e anotações, possibilitando a utilização de informações produzidas por eles na indexação desse conteúdo, e também, na construção do perfil de interesses. Assim, tem-se uma alternativa à indexação automática ou manual, eliminando problemas como a restrição de domínios e o alto custo, pois as anotações são colaborativas e não há necessidade de consultas a especialistas [12]. Adicionalmente, o ato de o usuário criar uma anotação relacionada a um conteúdo, como comentários, etiquetas, revisões e notas, indica o surgimento de uma reação (positiva ou negativa) nele [20]. Desse modo, o uso de anotações pode auxiliar a construção de seu perfil de interesses [7], e consequentemente, a recomendação de novos conteúdos.

Apesar de existirem vários trabalhos que exploram ano-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

WebMedia '12, October 15–18, 2012, São Paulo/SP, Brazil.

Copyright 2012 ACM 978-1-4503-1706-1/12/10 ...\$5.00.

tações para criar descrições sobre conteúdo multimídia [2, 9, 8], verifica-se ainda uma lacuna em como determinar a relevância das anotações em relação às preferências do usuário, uma vez que eles podem apresentar opiniões parciais e/ou diversas. Conseqüentemente, a classificação de sentimento do autor da anotação é um importante mecanismo para possibilitar a construção de perfis baseados em anotações mais eficientes. Como exemplo, dado um comentário razoavelmente completo do usuário sobre um filme, é possível extrair seus sentimentos em relação a diversos tópicos, como atuação do ator principal, assunto geral do filme relacionado ao gênero, cenas de interesse, etc. Essas informações, por sua vez, poderiam ser agregadas em um perfil para possibilitar o fornecimento de conteúdos relacionados.

O presente projeto de iniciação científica visa a extração de sentimentos relacionados a diversos tópicos a partir de anotações para possibilitar a construção de perfis de usuários. Entretanto, um passo anterior necessário para atingir esse objetivo é a classificação geral da anotação em positiva ou negativa. Este artigo tem como objetivo analisar diferentes técnicas de classificação de sentimento utilizando revisões de filmes criadas por usuários. Nesta análise, foram consideradas algumas estratégias de melhoria dos algoritmos existentes, de modo a quantificar o ganho obtido pela incorporação das funcionalidades. Uma dessas estratégias é a utilização da ferramenta SentiWordNet¹, que possibilita a mineração de opiniões pessoais a partir de termos de busca provenientes da base léxica WordNet².

Este artigo está organizado da seguinte maneira. Na Seção 2 são apresentados e discutidos alguns trabalhos relacionados. Na Seção 3 as abordagens selecionadas são detalhadas. A Seção 4 apresenta a avaliação dos métodos, comparando os resultados a partir de uma base de revisões única. Por fim, as Seções 5 e 6 descrevem as conclusões e agradecimentos deste trabalho.

2. TRABALHOS RELACIONADOS

Esta seção apresenta uma revisão dos trabalhos relacionados ao processo de classificação de sentimentos. Geralmente, as abordagens exploram diferentes características dos dados com o objetivo de melhorar os resultados de um classificador treinado a partir das informações extraídas.

Um dos trabalhos mais antigos da área foi proposto por Pang et al. [17], que desenvolveram um algoritmo de aprendizado supervisionado que utiliza *bag-of-words* em conjunto com máquinas de vetor de suporte (do Inglês, *support vector machines* – SVM) e unigramas.

Análises mais complexas que consideram partes do discurso (do Inglês, *part of speech* – POS) foram feitas em alguns trabalhos [22, 11, 6, 18], visando melhorar os resultados da classificação de sentimento em diferentes conjuntos de documentos. Além das informações contidas em partes de discurso, Turney [21] e Dave et al. [5] também propõem métodos para detectar e atribuir pesos a padrões de escrita a fim de se obter características dos dados a serem usados na classificação de sentimento em revisões de produtos.

Outras melhorias também foram obtidas por meio da identificação de sentenças objetivas e subjetivas [15], identificação de expressões coloquiais [23], e análise de aspectos do estilo da escrita [23, 1], tendo como objetivo a determinação

¹<http://sentiwordnet.isti.cnr.it/>

²<http://wordnet.princeton.edu/>

de opiniões específicas.

Ohana & Tierney [13] exploram a base SentiWordNet para extrair características do texto que são usadas em uma máquina de vetor de suporte para realizar a classificação de polaridade.

Alguns dos trabalhos mencionados e discutidos acima foram selecionados neste artigo para serem avaliados. A próxima seção descreve detalhadamente os algoritmos escolhidos, os quais serão posteriormente avaliados a partir de uma mesma base de dados para possibilitar a comparação das abordagens.

3. CLASSIFICAÇÃO DE REVISÕES

Um algoritmo de fácil entendimento e que apresenta um bom desempenho na área de classificação de revisões com análise de sentimentos é o classificador Naïve Bayes, construído com base na aplicação do teorema de Bayes. Uma abordagem didática deste algoritmo é apresentada pelo curso *online* de Processamento de Linguagem Natural da *Stanford University*³. Devido à sua simplicidade, é possível utilizá-lo como algoritmo de patamar para comparação com resultados obtidos de outras abordagens, inclusive aquelas que utilizam a base SentiWordNet.

Um classificador Naïve Bayes é alimentado com um conjunto de dados de treinamento para a realização de inferências em um conjunto de teste. Dado um documento a ser classificado e suas palavras, obtêm-se as probabilidades de esse documento pertencer às classes positiva e negativa. Assim, escolhe-se aquela classe cuja probabilidade é maior. Inicialmente, define-se a probabilidade condicional como:

$$P(c_j|d) \approx P(c_j) \prod_i P(w_i|c_j) \quad , \quad (1)$$

onde $P(c_j)$ é a probabilidade *a priori* da classe c_j (sentimento positivo ou negativo), d o documento a ser analisado e w_i a palavra de índice i do documento. A probabilidade *a priori* é definida como:

$$P(c_j) = \frac{\text{doccount}(C = c_j)}{N_{\text{doc}}} \quad , \quad (2)$$

onde $\text{doccount}(C = c_j)$ é o número de documentos pertencentes à classe c_j e N_{doc} é o número total de documentos.

Em seguida, são calculadas as probabilidades de cada palavra do documento dada uma classe. Tal probabilidade é definida como:

$$P(w_i|c_j) = \frac{\text{count}(w_i, c_j) + 1}{\sum_{w \in V} \text{count}(w, c_j) + |V|} \quad , \quad (3)$$

onde w_i representa a palavra de número i no documento, c_j a classe sendo avaliada e V o vocabulário de todo o conjunto de treinamento. Sobre esta fórmula é aplicada a técnica de suavização de Laplace [3], evitando-se que resultados desta probabilidade e da probabilidade condicional sejam zerados caso a palavra não ocorra no conjunto de treinamento.

Por fim, realiza-se o produto das probabilidades de cada classe, resultando em dois valores de probabilidade para um documento (probabilidade de o documento pertencer a uma inferência positiva ou negativa), sendo que o maior desses valores indica a provável classe a qual o documento pertence.

³<https://www.coursera.org/course/nlp/>

Buscando aumentar a eficiência do algoritmo, são eliminadas as *stop words*, que são palavras que não possuem nenhum valor sentimental atrelado a elas e que podem influenciar negativamente no algoritmo, caso apareçam demasiadamente em uma das classes do conjunto de treinamento. Além disso, negações são tratadas de modo que, sempre que um termo de negação aparecer (como “not” e “didn’t”), o prefixo “not_” seja adicionado a todas as palavras seguintes até a próxima pontuação [4, 17]. Desse modo, um documento que contém uma negação (como em “that was not good”) não terá seus aparentes termos positivos tratados de modo errôneo.

Uma outra melhoria na classificação [13] é utilizar a base léxica SentiWordNet, que contém uma mineração de opiniões pessoais a partir de termos de busca provenientes da base léxica WordNet. Em consultas a essa base léxica, além da própria palavra, é necessário saber sua classe gramatical (adjetivo, substantivo ou verbo). Assim, utiliza-se um etiquetador de partes de discurso [19], que adiciona a cada palavra um sufixo que descreve sua classe gramatical.

Tendo obtido as classes gramaticais é possível recuperar da base léxica os valores de positividade e negatividade associados aos termos. São desenvolvidas neste projeto abordagens que baseiam-se na soma destes valores, além da multiplicação dessa soma ao algoritmo de patamar (Naïve Bayes).

4. AVALIAÇÃO

Esta seção apresenta uma comparação dos algoritmos considerados, buscando-se, através da avaliação dos resultados, verificar como as abordagens se relacionam, inferindo possíveis vantagens e desvantagens de cada método.

Para esta avaliação, foi utilizada uma base de dados polarizada, introduzida por Pang & Lee [16], contendo 5331 revisões positivas sobre filmes e 5331 revisões negativas. Sobre esta base, foi aplicado o método de validação cruzada *10-fold*, evitando-se um possível sobreajuste de dados.

Através da contagem de verdadeiros positivos (VP), falsos positivos (FP), verdadeiros negativos (VN) e falsos negativos (FN) é possível calcular os índices de precisão e revocação (*precision e recall*). Em seguida, calcula-se a medida F1 que indica um valor único que combina ambas as métricas anteriores. Este processo é feito a cada particionamento da validação cruzada. Finalmente, as médias dos valores de precisão, revocação e F1 são calculadas para se obter os resultados finais.

Primeiramente, o algoritmo de patamar é aplicado por meio do classificador Naïve Bayes. Os resultados são apresentados na Tabela 1.

Tabela 1: Naïve Bayes.

Métrica/Classe	Positiva	Negativa
Precisão	0.7804	0.7790
Revocação	0.7786	0.7806
F1	0.7795	0.7798

Em seguida, são aplicadas a detecção de negações e a remoção de *stop words*. Em algumas iterações da validação cruzada este algoritmo se saiu melhor que o patamar, porém no geral foi influenciado por outras iterações nas quais apresentou resultados inferiores. Portanto, depois de calculada a média das métricas, a remoção de *stop words* e o tratamento de negações melhoraram somente a revocação da classe positiva. Isso indica que a estratégia é capaz apenas de reforçar aquelas palavras com conotação positiva que aparecem em

documentos classificados como positivos. Por exemplo, nas frases “that was good” (classe positiva) e “that was not good” (classe negativa), a palavra “good” teria a mesma probabilidade em ambas as classes no algoritmo de patamar, ao passo que nesse algoritmo, “good” terá uma probabilidade maior de pertencer à classe positiva. Consequentemente, novos documentos com inferência positiva contendo a palavra “good” serão mais facilmente detectados, aumentando, assim, a revocação.

Tabela 2: Algoritmo de patamar com tratamento de negações e remoção de *stop words*.

Métrica/Classe	Positiva	Negativa
Precisão	0.7534	0.7785
Revocação	0.7895	0.7410
F1	0.7710	0.7592

Na Tabela 3 são apresentados os resultados obtidos para o algoritmo que, para cada palavra pertencente a um documento, resgata um valor sentimental da base léxica SentiWordNet. As somas dos valores de positividade e negatividade são realizadas, seguidas de uma comparação entre elas. O maior valor será a classe inferida àquele documento.

Tabela 3: Algoritmo de soma de valores obtidos da SentiWordNet.

Métrica/Classe	Positiva	Negativa
Precisão	0.5823	0.5995
Revocação	0.6775	0.4982
F1	0.6262	0.5441

Finalmente, na Tabela 4, são apresentados os resultados do algoritmo que combina o algoritmo de patamar com os valores da SentiWordNet. Após o produtório de todas as probabilidades de termos dada uma classe no classificador Naïve Bayes, a soma de todos os valores de positividade ou negatividade desse documento é aplicada à probabilidade resultante do algoritmo de patamar.

Tabela 4: Algoritmo de patamar multiplicado por SentiWordNet.

Métrica/Classe	Positiva	Negativa
Precisão	0.6873	0.6980
Revocação	0.7335	0.6486
F1	0.7094	0.6723

A melhor eficiência foi observada no algoritmo de patamar. Apesar disso, sabe-se que esse algoritmo é dependente do conjunto de treinamento, dependendo portanto de seu tamanho e do domínio envolvido. No cenário de estudo deste projeto, um conjunto de teste envolvendo revisões de filmes apresenta bons resultados. Porém, caso o domínio do conjunto de teste fosse alterado, os resultados poderiam ser inferiores, visto que termos técnicos de outros domínios não seriam reconhecidos, pois não apareceriam no conjunto de treinamento.

Por outro lado, a utilização da base SentiWordNet é interessante por ser um conjunto de dados bem definido e fixo, que não depende do domínio a ser aplicado e de nenhum conjunto de dados de treinamento. Essa independência, entretanto, também apresenta algumas desvantagens, pois em certos domínios algumas palavras podem apresentar valores sentimentais mais fortes do que em outros.

Todas as abordagens apresentaram vantagens e desvantagens, portanto, uma análise mais aprofundada das técnicas será realizada em trabalhos futuros, com a intenção de se

elaborar uma melhor combinação dos algoritmos que resolve ou reduza seus problemas de uma maneira mais eficiente se comparado ao algoritmo de combinação apresentado neste trabalho.

5. CONCLUSÃO

Este artigo apresentou uma avaliação de diferentes algoritmos de classificação de polaridade de revisões em filmes. Uma abordagem baseada em Naïve Bayes foi considerada como patamar, e a partir dela, melhorias foram adicionadas para que fosse quantificado o ganho obtido com essas alterações.

Conforme mencionado na introdução deste artigo, o presente estudo será importante no passo seguinte do projeto de iniciação científica, que é a construção de perfis de interesse para aplicação em sistemas de recomendação. Desse modo, como trabalhos futuros pretende-se investigar meios de se extrair múltiplas inferências de diferentes tópicos a partir de uma única revisão ou comentário. Espera-se como resultado que um perfil de preferências seja construído a partir de usuários colaboradores, oferecendo um mecanismo diferenciado de personalização de conteúdo.

6. AGRADECIMENTOS

Os autores gostariam de agradecer o apoio financeiro das agências PIBIC/CNPq e FAPESP, número de processo 2011/17366-2.

7. REFERÊNCIAS

- [1] A. Abbasi, H. Chen, and A. Salem. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Trans. Inf. Syst.*, 26(3):12:1–12:34, 2008.
- [2] S. Angeletou, M. Sabou, and E. Motta. Folksonomy Enrichment and Search. In L. Aroyo, P. Traverso, F. Ciravegna, P. Cimiano, T. Heath, E. Hyvonen, R. Mizoguchi, E. Oren, M. Sabou, and E. P. B. Simperl, editors, *6th. European Semantic Web Conference*, volume 5554 of *Lecture Notes in Computer Science*, pages 801–805. Springer, 2009.
- [3] P. R. C. D. Manning and H. Schtze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.
- [4] S. Das and M. Chen. Yahoo! for amazon: Extracting market sentiment from stock message boards. In *Proceedings of the Asia Pacific Finance Association Annual Conference (APFA)*, 2001.
- [5] K. Dave, S. Lawrence, and D. M. Pennock. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web, WWW '03*, pages 519–528, New York, NY, USA, 2003. ACM.
- [6] M. Gamon. Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In *Proceedings of the 20th international conference on Computational Linguistics, COLING '04*, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.
- [7] S. Gauch, M. Speretta, A. Chandramouli, and A. Micarelli. User Profiles for Personalized Information Access. In P. Brusilovsky, A. Kobsa, and W. Nejdl, editors, *The Adaptive Web*, volume 4321 of *Lecture Notes in Computer Science*, pages 54–89, 2007.
- [8] H. Halpin, V. Robu, and H. Shepherd. The Complex Dynamics of Collaborative Tagging. In *Proceedings of the 16th International Conference on World Wide Web*, pages 211–220, Banff, Alberta, Canada, 2007.
- [9] M. Heckner, T. Neubauer, and C. Wolff. Tree, funny, to.read, google: What are Tags Supposed to Achieve? A Comparative Analysis of User Keywords for Different Digital Resource Types. In *Proceeding of the 2008 ACM Workshop on Search in Social Media*, pages 3–10, Napa Valley, California, USA, 2008.
- [10] D. Kelly and J. Teevan. Implicit feedback for inferring user preference: a bibliography. *SIGIR Forum*, 37(2):18–28, Sept. 2003.
- [11] A. Kennedy and D. Inkpen. Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence*, 22(2):110–125, 2006.
- [12] M. G. Manzato. *Uma arquitetura de personalização de conteúdo baseada em anotações do usuário*. PhD thesis, Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo, 2011.
- [13] B. Ohana and B. Tierney. Sentiment classification of reviews using SentiWordNet. In *9th. IT & T Conference, Dublin Institute of Technology*, Dublin, Ireland, 2009.
- [14] T. O'Reilly. What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. *Communications & Strategies*, 1:17, 2007.
- [15] B. Pang and L. Lee. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, ACL '04*, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.
- [16] B. Pang and L. Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of ACL*, pages 115–124, 2005.
- [17] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10, EMNLP '02*, pages 79–86, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.
- [18] F. Salvetti, S. Lewis, and C. Reichenbach. Automatic opinion polarity classification of movie reviews. *Colorado Research in Linguistics*, 17(1), 2004.
- [19] K. Toutanova and C. D. Manning. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics - Volume 13, EMNLP '00*, pages 63–70, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics.
- [20] K. H. L. Tso-Sutter, L. B. Marinho, and L. Schmidt-Thieme. Tag-aware Recommender Systems by Fusion of Collaborative Filtering Algorithms. In *Proceedings of the 2008 ACM Symposium on Applied Computing*, pages 1995–1999, Fortaleza, CE, Brazil, 2008.
- [21] P. D. Turney. 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, ACL '02*, pages 417–424, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.
- [22] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, HLT '05*, pages 347–354, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.
- [23] K. Yang, N. Yu, and H. Zhang. Widit in trec 2007 blog track: Combining lexicon-based methods to detect opinionated blogs. In E. M. Voorhees and L. P. Buckland, editors, *TREC*, volume Special Publication 500-274. National Institute of Standards and Technology (NIST), 2007.