Mecanismo de Encadeamento de Notícias por Reconhecimento de Implicação Textual

Phillipe S. Cavalcante, Wallace A. Pinheiro

Instituto Militar de Engenharia, Rio de Janeiro - RJ - Brasil
{cavalcante.phillipe, wallaceapinheiro}@gmail.com

Abstract. Atualmente, o acesso à informação relevante é feito com uma simples consulta a um sistema de busca na Web. Em se tratando de notícias, além de um sistema de consulta para o acesso à informação, sites especializados sugerem ao usuário notícias relacionadas para leitura complementar. No entanto, estas notícias relacionadas nem sempre complementam o conteúdo da primeira notícia lida pelo usuário. Os principais problemas desta abordagem são a existência de conteúdo redundante e a falta de ordenação segundo um critério que proporcione uma leitura textual coerente para o usuário, isto é, uma leitura que forneça notícias organizadas de tal modo que proporcione ao usuário uma progressão do conteúdo que está sendo lido. Este trabalho apresenta um mecanismo de encadeamento textual capaz de organizar notícias relacionadas segundo um critério de implicação textual, proporcionando ao usuário uma leitura textual coerente. Para isso, o mecanismo utiliza um sistema de reconhecimento de implicação textual para o encadeamento de notícias, e um sistema de similaridade para identificação de notícias com conteúdo semelhante. Os experimentos realizados mostram que o mecanismo produz encadeamentos textuais semelhantes aos encadeamentos organizados pelos usuários.

Categories and Subject Descriptors: H.3.3 [**Information Search and Retrieval**]: Information filtering, Retrieval models; I.7 [**Document and Text Processing**]: Miscellaneous; I.2.7 [**Natural Language Processing**]: Language parsing and understanding

Keywords: Encadeamento de Notícias, Reconhecimento de Implicação Textual, Notícias Relacionadas

1. INTRODUÇÃO

Sistemas de busca têm proporcionado ao usuário acesso à informação relevante e atualizada na *Web*. Contudo, estes sistemas recuperam tanta informação relacionada que escolher por onde começar ou continuar uma leitura tem se tornado uma tarefa trabalhosa para o usuário [Mulder et al. 2006; Kobayashi and Takeda 2000]. Essa tarefa é bastante comum em sites de notícias e tem sido tratada com o uso de sistemas de recomendação [Adomavicius and Tuzhilin 2005].

Sistemas de recomendação de notícias, em geral, sugerem notícias de acordo com um critério de similaridade textual [Lv et al. 2011]. No entanto, estas sugestões nem sempre complementam a notícia inicialmente lida pelo usuário. Os principais problemas encontrados nesta abordagem são: (*i*) redundância de conteúdo, uma vez que o critério de similaridade agrupa notícias semelhantes; e (*ii*) a falta de ordenação das notícias, de modo a proporcionar ao usuário uma progressão do conteúdo que está sendo lido.

O sistema proposto por [Rodrigues et al. 2010] aborda os dois problemas apresentados e utiliza o conceito de sabedoria das multidões para a sugestão de notícias em um encadeamento causal e temporal, mas, tem a necessidade de solicitar *feedback* do usuário para realizar o encadeamento. Outra solução é apresentada por [Shahaf and Guestrin 2011; 2012], que propõe um método de encadeamento textual entre duas notícias previamente selecionadas pelo usuário, porém, este método se torna inviável para a recomendação de notícias, já que a única informação diponível é a notícia selecionada pelo usuário.

Este artigo apresenta um mecanismo de encadeamento de notícias capaz de fornecer notícias organizadas segundo um critério de implicação textual, proporcionando ao usuário uma progressão do conteúdo que está sendo lido. Para isso, o mecanismo utiliza um sistema de reconhecimento de implicação textual, para o encadeamento das notícias, e um sistema de similiaridade textual, para identificação de notícias semelhantes.

Os resultados obtidos com este mecanismo foram comparados com o sistema proposto por [Rodrigues 2011; Rodrigues et al. 2010] e mostraram que o mecanismo de encadeamento de notícias por implicação textual tem uma acurácia maior do que a apresentada pelo encadeamento de notícias segundo um critério causal e temporal.

2. RECONHECIMENTO DE IMPLICAÇÃO TEXTUAL E SIMILARIDADE TEXTUAL

O reconhecimento de implicação textual identifica quando existe inferência semântica entre dois textos. Dados dois textos, t e h, t implica textualmente em h quando o significado de h é provável de ser inferido do significado de t [Tatar et al. 2009; Zanzotto et al. 2009].

Por exemplo, entre os textos t e h, a seguir, ocorre implicação textual:

t: O Brasil foi a sede da copa do mundo de 2014, e
h: O Brasil participou da copa do mundo de 2014.

Neste exemplo, h pode ser inferido semanticamente de t; se o Brasil foi a sede da copa do mundo de 2014, então o Brasil participou da copa do mundo de 2014. Portanto, dizemos que existe uma implicação textual entre t e h. Observa-se que não existe implicação textual no sentido contrário; a participação do Brasil na copa do mundo de 2014 não quer dizer que ele tenha sido o anfitrião da copa do mundo de 2014, por mais que isso seja verdade. A implicação textual não é tão formal quanto a implicação lógica, pois ela aceita casos em que a verdade de h seja plausível, em vez de certa.

O TF-IDF é um coeficiente de similaridade textual que indica o quão importante é uma palavra para um documento, em relação a um conjunto de documentos [Zhang et al. 2008]. Este coeficiente é utilizado no módulo de similaridade (ver figura 1) do mecanismo para a formação do índice de similaridade textual. O índice de similaridade é uma matriz de ordem $n \times n$, onde n é a quantidade de notícias do sistema. Neste trabalho, as notícias foram transformadas em vetores onde cada dimensão indica o TF-IDF da palavra que está presente no documento. Cada componente do índice de similaridade é obtida através do cosseno entre cada par de notícias [Zhang et al. 2008].

Outra medida de similaridade utilizada neste artigo é o coeficiente de Jaccard: $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$, onde A e B são dois conjuntos de termos, os documentos a serem comparados. O coeficiente de Jaccard avalia a similaridade e a diversidade das palavras presentes nos documentos [Rodrigues 2011]. Este coeficiente é utilizado no extrator de características do módulo de implicação textual, apresentado na figura 2.

3. MECANISMO DE ENCADEAMENTO NOTÍCIAS

O mecanismo de encadeamento de notícias é capaz de ordenar as notícias de acordo com um critério de implicação textual. Para isso, o mecanismo utiliza um sistema de reconhecimento de implicação textual e um sistema de identificação de similaridade textual. A figura 1 ilustra o mecanismo proposto.

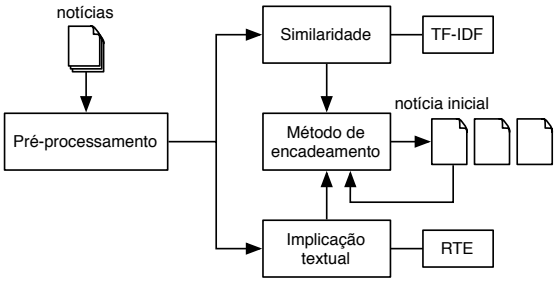

Fig. 1. Mecanismo de Encadeamento de Notícias.

O mecanismo é formado por dois módulos principais: (*i*) módulo de similaridade, responsável por gerar um índice de similaridade; e (*ii*) módulo de implicação textual, responsável por treinar a máquina de reconhecimento de implicação textual e gerar um índice de implicação textual (ver figura 2).

O módulo de *pré-processamento* é responsável pela limpeza textual das notícias do sistema. Algumas funções deste módulo são: remoção de *stopwords* e pontuações, conversão de caracteres para UTF-8 e transformação para caracteres minúsculos, e obtenção de lema das palavras.

O encadeamento textual é gerado por um algoritmo de estratégia gulosa, localizado no módulo *método de encadeamento*. Este algoritmo busca no módulo de *implicação textual* uma notícia que implique semanticamente na *notícia inicial*, escolhida pelo usuário, e a adiciona à lista encadeada de notícias. O primeiro elemento da lista encadeada é a própria notícia inicial.

Caso exista mais de uma notícia implicante, o algoritmo compara o valor de similaridade entre as notícias implicantes - esses valores são encontrados no módulo de *similaridade* - e escolhe aquela notícia com o menor valor de similaridade, adicionando-a à lista encadeada. Este processo se repete para a escolha da próxima notícia implicante: a notícia inicial passa a ser a notícia implicante até que não haja notícia implicante à notícia anterior.

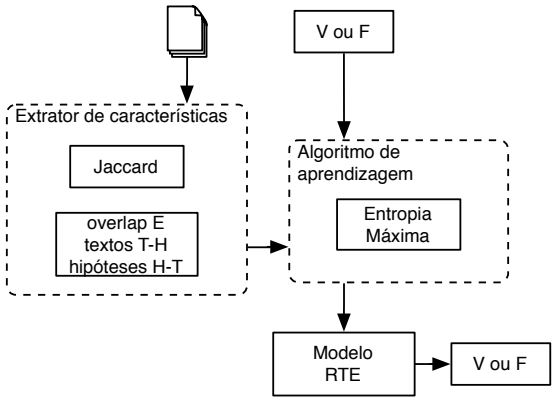

Fig. 2. Módulo de Implicação Textual.

A figura 2 apresenta a arquitetura do módulo de *implicação textual*. Este módulo contém um extrator de características da base de treinamento de Reconhecimento de Implicação Textual (RTE), esta base foi obtida de [Dagan et al. 2005] e traduzida para o português, e um algoritmo de aprendizagem supervisionada, baseado em Entropia Máxima [Malouf 2002]. O primeiro é responsável pela extração de características de similaridade baseadas no coeficiente de Jaccard. O segundo, gera um classificador que recebe como entrada pares de notícias e classifica o relacionamento entre elas como sendo de implicação textual ou não. Este componente identifica relações entre o fim de uma notícia e o início de outra, para a formação do encadeamento [Rodrigues 2011].

As características extraídas pelo módulo de implicação textual foram, dado duas notícias t e h:

(1) O conjunto de termos que existem em t, mas não existem em h;

(2) O conjunto de termos que existem em h, mas não existem em t;

(3) O conjunto de termos que t e h têm em comum;

3.1 Avaliação do Encadeamento

A avaliação do encadeamento gerado pelo mecanismo é feita através da observação do coeficiente de correlação de Spearman (ρ) [Rodrigues 2011]. Este coeficiente é um indicador da relação de ordem entre duas variáveis, X e Y. Para a obtenção do cálculo do coeficiente, dada uma amostra de tamanho n, primeiro as variáveis X e Y são ordenadas e transformadas em x e y. O coeficiente de Spearman é obtido pela equação: $\rho = 1 - \frac{6 \cdot \sum d_i^2}{n \cdot (n^2 - 1)}$, em que $d_i = x_i - y_i$ é a diferença entre as posições de cada observação das variáveis.

Os valores deste coeficiente variam entre -1 e 1. Quando o valor tende a 1, significa que a ordenação da variável Y segue a ordenação da variável X. Quando o valor tende a -1, significa que Y segue a ordenação invertida da variável X. Para este experimento, por exemplo, supondo que Y seja o encadeamento do mecanismo e que X seja o encadeamento feito pelo usuário, um valor de $\rho = 1$ (ou ρ tendendo a 1) diz que o encadeamento feito pelo mecanismo é similar ao encadeamento do usuário (ou tende ao encadeamento do usuário).

A escolha da avaliação usando o coeficiente de Spearman e a quantidade de usuários para o experimento é justificada em [Rodrigues 2011]. A principal razão para a escolha deste coeficiente é que ele avalia a ordenação entre duas variáveis. Com relação a amostragem de notícias feita para o experimento, a quantidade de notícias utilizadas foi de 5, todas relacionadas a um mesmo tópico. As notícias foram extraídas do Google News (news.google.com) sobre o tópico Crise na Coréia do Norte. Elas foram selecionadas considerando o tempo de publicação, a similaridade, e se juntas formavam um encadeamento textual. As notícias continham, em média, 300 palavras. Experimentos iniciais com 10 notícias causaram desconforto no usuário, quanto a ordenação das notícias segundo o critério de encadeamento textual. Isto é, a partir da sexta notícia o usuário já apresentava dificuldades de lembrar o conteúdo presente nas 5 notícias anteriores para poder ordenar as próximas 5 notícias.

A quantidade de participantes escolhida para o experimento foi de 10. Os participantes tinham entre 21 e 28 anos, e grau de instrução mínimo de Mestrado em Ciência da Computação. O artigo de [Rodrigues 2011] mostra que 9 é um número suficiente para a realização do experimento, baseando-se na proporção populacional do Brasil. A fórmula utilizada para o cálculo da quantidade de usuários é a seguinte: $n = \frac{Z_{\alpha/2}^2 \cdot p \cdot q}{E^2}$, onde: n é o número desejado de usuários (10); $Z_{\alpha/2}$ é o valor crítico que corresponde ao grau de confiança desejado (80%); p é a proporção populacional pertencente a categoria que se deseja estudar (50%); q é a proporção populacional que não pertence à categoria que se deseja estudar ($q = 1 - p$); e E é a margem de erro, que indica a diferença máxima entre a proporção da amostra utilizada e a verdadeira proporção populacional (20%).

4. RESULTADOS

Os resultados mostraram que o encadeamento de notícias, gerado pelo mecanismo, seguiu a ordenação das sequências encadeadas pelos usuários, conforme a figura 3. Na figura, o eixo horizontal representa o encadeamento realizado por cada usuário (de 0 a 9), e o eixo vertical representa o valor de Spearman entre o encadeamento gerado pelo mecanismo e o encadeamento de notícias de cada usuário. O coeficiente de Spearman obtido foi, em média, de $\rho = 0.84$, o que indica que as notícias geradas pelo mecanismo tiveram uma ordenação próxima da ordenação realizada pelo usuário.

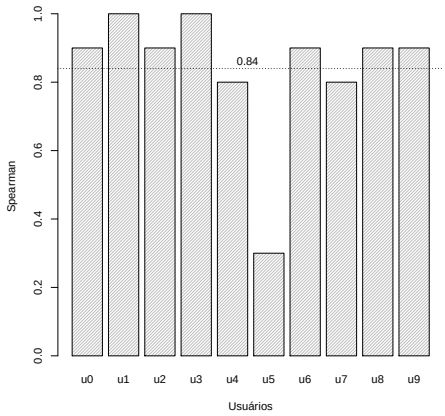

Fig. 3. Comparação do encadeamento do Mecanismo com o encadeamento dos usuários.

Além disso, o mecanismo produziu resultados melhores que os apresentados em [Rodrigues 2011] ($\rho = 0.3$) para um conjunto de notícias relacionadas a apenas um tópico.

Uma análise feita com os dados utilizados para o experimento, mostrou que a mediana do coeficiente de Spearman foi de 0.9 e que o valor mínimo foi de 0.3. Este valor mínimo é um *outlier*, por ser o único resultado fora do padrão encontrado.

Na tabela I é apresentado o encadeamento de notícias formado pelos usuários e pelo mecanismo. Na tabela, as notícias estão enumeradas de 1 a 5. Na primeira linha, encontra-se a lista de usuários submetidos ao experimento, enumerados de 0 a 9, e o mecanismo. Da primeira à décima coluna tem-se o encadeamento feito por cada usuário e a distância entre o encadeamento feito pelo mecanismo e por cada usuário, separados pelo símbolo /. Dessa forma, como o usuário 0 e o mecanismo escolheram a notícia 1 para ser a primeira notícia do encadeamento, a distância entre a posição da primeira notícia sugerida pelo mecanismo e a primeira notícia escolhida pelo usuário é igual a 0, ou seja, 1/0. Já no caso do usuário 5, a notícia escolhida para ser a primeira do encadeamento foi a notícia 4. Assim, a distância entre a posição da primeira notícia sugerida pelo mecanismo e a primeira notícia escolhida pelo usuário é igual a 3, ou seja, 4/3.

Table I. Encadeamento de Notícias: Usuário × Mecanismo

u_0	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	Mecanismo
1/0	1/0	1/0	1/0	2/1	4/3	1/0	2/1	1/0	1/0	1
2/0	2/0	2/0	2/0	1/1	1/1	2/0	1/1	2/0	2/0	2
3/0	3/0	4/1	3/0	3/0	3/0	3/0	3/0	3/0	4/1	3
5/1	4/0	3/1	4/0	5/1	2/2	5/1	5/1	5/1	3/1	4
4/1	5/0	5/0	5/0	4/1	5/0	4/1	4/1	4/1	5/0	5

Uma análise sobre os valores da tabela I mostra que o número de acertos feito pelo mecanismo é de aproximadamente 58%, considerando que a base de cálculo para isso seja $1 - \frac{e}{c}$, onde e é a quantidade total de erros (21) e c é a quantidade total de comparações realizadas para cada posição do encadeamento (50, número de usuários por quantidade de notícias).

5. CONCLUSÃO E TRABALHOS FUTUROS

O principal objetivo do mecanismo proposto é o encadeamento de notícias relacionadas segundo um critério de implicação textual. Para isso, o mecanismo utiliza técnicas de reconhecimento de implicação textual e de similaridade entre textos.

Os resultados obtidos mostraram que o mecanismo é capaz de fornecer encadeamentos de notícias mais próximos da necessidade do usuário, solucionando os problemas de redundância textual e ordenação, apresentados no início deste trabalho, de uma forma plausível. Além disso, a tarefa de encadeamento se mostrou capaz de ordenar as notícias no tempo, proporcionando uma leitura textual coerente. Sendo assim, um passo importante para a geração automática de textos para a contagem de uma história, do início ao fim.

O uso de técnicas para geração de árvores sintáticas sobre os textos durante o processo de implicação textual é uma abordagem que pode trazer melhorias à geração do encadeamento textual, bem como a utilização de lógica de primeira ordem, para a computação semântica das expressões de linguagem natural. Tais técnicas podem ser implementadas em trabalhos futuros.

REFERENCES

Adomavicius, G. and Tuzhilin, A. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. *IEEE Trans. on Knowl. and Data Eng.* 17 (6): 734–749, June, 2005.

Dagan, I., Glickman, O., and Magnini, B. The PASCAL Recognising Textual Entailment Challenge. In *Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment*, 2005.

Kobayashi, M. and Takeda, K. Information retrieval on the web. *ACM Comput. Surv.* 32 (2): 144–173, June, 2000.

Lv, Y., Moon, T., Kolari, P., Zheng, Z., Wang, X., and Chang, Y. Learning to model relatedness for news recommendation. In *Proceedings of the 20th international conference on World wide web*. WWW '11. ACM, New York, NY, USA, pp. 57–66, 2011.

Malouf, R. A comparison of algorithms for maximum entropy parameter estimation. In *proceedings of the 6th conference on Natural language learning - Volume 20*. COLING-02. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 1–7, 2002.

Mulder, I., de Poot, H., Verwij, C., Janssen, R., and Bijlsma, M. An information overload study: using design methods for understanding. In *OZCHI '06: Proceedings of the 20th conference of the computer-human interaction special interest group (CHISIG) of Australia on Computer-human interaction: design: activities, artefacts and environments*. ACM, New York, NY, USA, pp. 245–252, 2006.

Rodrigues, T. S. *WhySearch: Um Mecanismo Causal e Temporal de Encadeamento de Notícias*. M.S. thesis, Universidade Federal do Rio de Janeiro, 2011.

Rodrigues, T. S., Pinheiro, W. A., Souza, J. M., and Xexeo, G. Relacionando Notícias Web: Uma Abordagem Causal e Temporal. In *9th International Information and Telecommunication Technologies Symposium*, 2010.

Shahaf, D. and Guestrin, C. Connecting the dots between news articles. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence - Volume Volume Three*. IJCAI'11. AAAI Press, Barcelona, Catalonia, Spain, pp. 2734–2739, 2011.

Shahaf, D. and Guestrin, C. Connecting Two (or Less) Dots: Discovering Structure in News Articles. *ACM Trans. Knowl. Discov. Data* 5 (4): 24:1–24:31, Feb., 2012.

Tatar, D., Serban, G., and Andreea, M. Textual Entailment as a Directional Relation. *Journal of Research and Practice in Information Technology - ACJ*, 2009.

Zanzotto, F. M., Pennacchiotti, M., and Moschitti, A. A Machine Learning Approach to Textual Entailment Recognition. *Natural Language Engineering* vol. 15, pp. 551–582, 10, 2009.

Zhang, W., Yoshida, T., and Tang, X. TFIDF, LSI and Multi-word in Information Retrieval and Text Categorization. In *Systems, Man and Cybernetics, 2008. SMC 2008. IEEE International Conference on*. pp. 108–113, 2008.