Identificação de Pontos Perceptualmente Importantes (PIP) em séries temporais de tópicos extraídos de dados textuais

Lucas Santiago Rodrigues[1]
Roberta Akemi Sinoara[2]
Solange Oliveira Rezende[3]
Ricardo Marcondes Marcacini[4]
Maria Fernanda Moura[5]

Resumo: Neste trabalho é apresentado um módulo computacional denominado PIPC (PIP Classification) que permite identificar Pontos Perceptualmente Importantes (PIP) em séries temporais. O módulo foi desenvolvido para apoiar o projeto Compilação e Recuperação de Informações Técnico-científicas e Indução ao Conhecimento (CRITIC@), permitindo identificar os pontos relevantes da evolução temporal de um tópico extraído dos textos, identificar documentos textuais que possam auxiliar a interpretar tais pontos, bem como classificar a formação de próximos PIPs nas séries temporais. Foram realizados testes do módulo a partir de notícias sobre produção de milho no Brasil, e os resultados preliminares de avaliação do módulo são promissores.

Palavras-chave: séries temporais, extração de tópicos, classificação.

[1] Estudante de Sistemas de Informação da Universidade Federal de Mato Grosso do Sul - Campus de Três Lagoas (UFMS), estagiário da Embrapa Informática Agropecuária, Campinas, SP.

[2] Bacharel em Informática, doutoranda em Ciências da Computação e Matemática Computacional no Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo (ICMC-USP), São Carlos, SP.

[3] Graduada em Licenciatura em Ciências Habilitação Matemática, doutora em Engenharia Mecânica, professora associada da Universidade de São Paulo (ICMC-USP), São Carlos, SP.

[4] Bacharel em Informática, doutor em Ciências da Computação e Matemática Computacional, docente e pesquisador da Universidade Federal de Mato Grosso do Sul (UFMS), Três Lagoas, MS.

[5] Estatística, doutora em Ciências Matemáticas e da Computação, pesquisadora da Embrapa Informática Agropecuária, Campinas, SP.

Introdução

A extração automática de tópicos a partir de dados textuais tem recebido grande atenção nos últimos anos, uma vez que os tópicos permitem sumarizar, de forma inteligente, o conhecimento embutido em grandes conjuntos de dados textuais (REZENDE et al., 2003). Assim, é possível que usuários explorem coleções textuais, que são inerentemente não estruturadas, de uma forma organizada e intuitiva (MOURA, 2009). Uma aplicação promissora para esse tipo de tarefa é demonstrada no projeto CRITIC@, em que os resultados retornados por motores de busca sobre informação técnico-científicas são organizados por meio de tópicos. Além disso, um diferencial do projeto CRITIC@ é analisar a "evolução temporal" do tópico, ou seja, computar uma série temporal que permite visualizar os períodos em que um tópico é mais ou menos ativo (MARCACINI; REZENDE, 2010).

Nesse trabalho foi desenvolvido um módulo em linguagem de programação JAVA para identificação e classificação de PIP em séries temporais de tópicos extraídos a partir de dados textuais, denominada PIPC (PIP Classification). Os PIP's são definidos como pontos críticos de determinada série temporal, uma vez que representam pontos com valores muito diferentes de seus vizinhos, geralmente picos e vales de uma série. No módulo PIPC, foi implementado um algoritmo tradicional da literatura para identificação dos PIP's, baseada em distância Euclidiana para selecionar os pontos de nosso interesse. Além de identificar os PIP's, o módulo desenvolvido possui duas funções: a) identificar um documento textual mais relevante associado ao PIP de uma série; e b) predizer a probabilidade de ocorrência de um PIP nos próximos pontos da série por meio de algoritmos de classificação. Dessa forma, no contexto do projeto CRITIC@, os usuários podem analisar as séries temporais com um nível mais rico de detalhes e obter a tendência de formação de ponto perceptualmente importante no futuro. Os experimentos de avaliação da ferramenta foram realizados em tópicos extraídos de textos sobre a produção de Milho no Brasil e se mostraram promissores.

Materiais e Métodos

O módulo PIPC (PIP Classification) desenvolvido durante o estágio é ilustrado na Figura 1. Inicialmente, recebe, como entrada, uma série temporal de

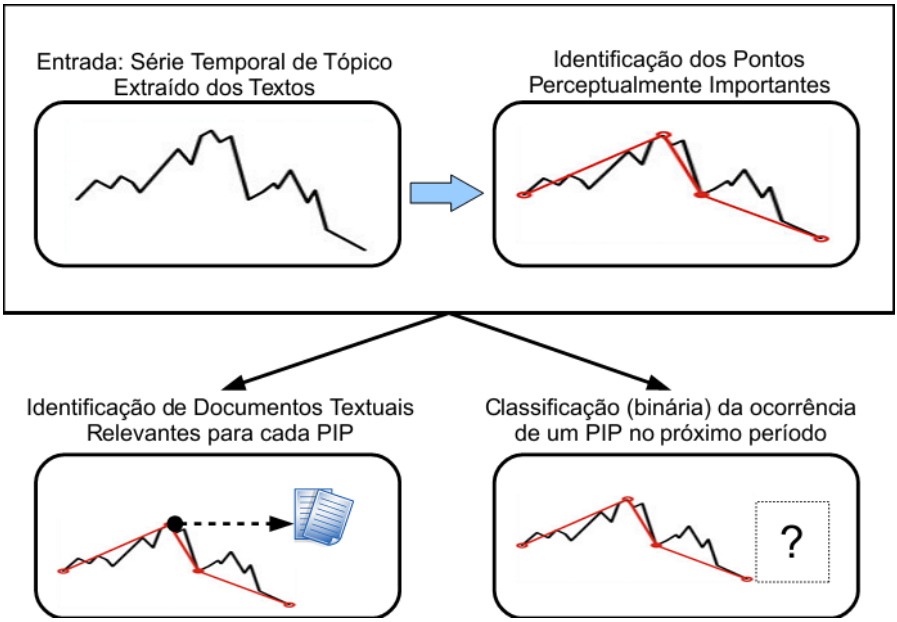

Figura 1. Fluxograma do módulo DATool.

tópico extraído dos textos. Esta série de entrada é um formato XML utilizado pelo CRITIC@, com os descritores dos tópicos, os documentos associados e a série temporal. No próximo passo é executado um algoritmo para identificação dos PIP's, conforme descrito em Sanches (2006). A ideia geral deste algoritmo utiliza a distância euclidiana entre os pontos da série temporal, inicializando-se com o primeiro e o último ponto da série. A partir desses dois pontos iniciais, calcula-se o ponto mais distante entre os dois e este é definido como um PIP. Este mesmo procedimento é executado recursivamente para cada dois PIPs anteriores da série. Um exemplo da execução desse algoritmo é apresentado na Figura 2.

Após a identificação dos PIPs da série temporal, o módulo PIPC permite a execução de duas funcionalidades, conforme descritas a seguir.

A primeira funcionalidade é a Identificação de documentos textuais relevantes para cada PIP, na qual o usuário pode selecionar um PIP na série temporal e visualizar qual documento textual representa aquele período. Para obter este resultado, foi utilizado o conceito de centroide, geralmente empregado na área de agrupamento de dados. Para tal, é obtido o conjunto

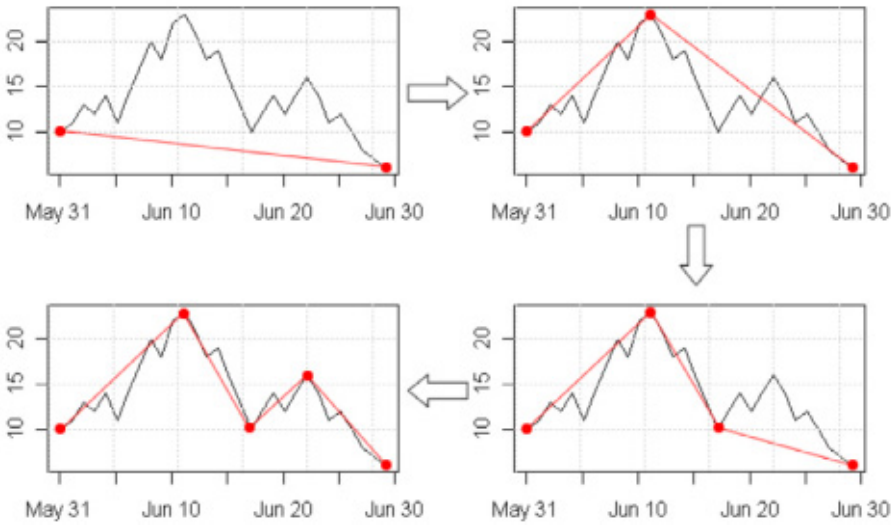

Figura 2. Exemplo de Identificação de Pontos Perceptualmente Importantes. No primeiro quadro são selecionados os pontos iniciais. Nas próximas iterações são selecionados os pontos mais distantes para cada dois pontos consecutivos.

Fonte: Sanches (2006).

de documentos publicados no período temporal de um PIP de interesse a partir do XML de entrada. A partir desse conjunto, é calculado o vetor centroide, que é um vetor médio que sumariza todos os documentos do conjunto. Em seguida, calcula-se a similaridade de cada documento do período a este centroide por meio da similaridade cosseno. O documento textual que apresenta maior similaridade ao centroide é então considerado o mais relevante do período. Desse modo, o usuário pode interagir com cada PIP e utilizar esses documentos mais relevantes como uma opção para interpretar e entender o motivo dos picos e vales existentes na série temporal.

A segunda funcionalidade é Classificação (binária) da ocorrência de um PIP no próximo período da série temporal do tópico. Nesse caso, a série temporal é modelada em uma nova tabela atributo-valor por meio de suas subsequências. Assim, definido o tamanho da subsequência, busca-se todas as subsequências da série temporal em que o último ponto da subsequência é um PIP. Essas subsequências são definidas como classe POSITIVA, ou seja, que representam a ocorrência de um PIP. Todas as outras subsequências da série temporal que não possuem PIP são definidas como classe

NEGATIVA. Por meio dessa estrutura, é possível treinar classificadores binários para que, dada uma nova subsequência, retorne à classe (positiva ou negativa) e ao respectivo valor de confiança da classificação. Essas informações são utilizadas para identificar uma possível tendência do próximo ponto para apoiar os usuários na análise das séries temporais de tópicos extraídos dos textos. Para a análise dos PIP's e dos documentos utilizamos a biblioteca Weka, que possui algoritmos de análises e foi incorporada aos códigos desenvolvidos durante a pesquisa.

Resultados e Discussão

A avaliação do módulo PIPC foi feita no domínio "Produtividade de Milho no Brasil". Este domínio foi escolhido devido à sua crescente demanda com o setor de avicultura, suinocultura e consumo industrial. Os meios de comunicação do Brasil divulgam com grande destaque o setor agrícola, dessa forma, temos incontáveis quantidades de dados que foram publicados durante décadas até a atualidade. Atualmente, com os avanços da tecnologia, esses dados podem ser usados por instituições e empresas para gerar conhecimento, sendo possível analisar vários indicadores de uma safra. No caso particular desse estudo, foi analisada a safra do milho.

Foram utilizados 5674 documentos textuais sobre produtividade de milho coletados de diversos portais especializados. Para os experimentos, foi selecionado um tópico em que os descritores "[aumento, produção, milho]" (e suas variações morfológicas) estão presentes com maior frequência, no período de 2008 à 2010, conforme ilustrado na Figura 3. Na figura também há um PIP selecionado para analisar a funcionalidade sobre identificação e notícias relevantes para o PIP. Neste período (entre maio e junho de 2009), o referido tópico contém notícias sobre expectativas e tendências de alta de produtividade no setor, e a notícia identificada mais próximo ao centroide é

Figura 3. Evolução temporal do tópico [aumento, producao, milho] no período de 2008 à 2010.

"Tendência é de alta nas exportações de milho"[6]. Esta análise é subjetiva e uma validação mais robusta depende de especialistas de domínio da área.

Para avaliar a classificação de ocorrência de um PIP no próximo período foi utilizado o algoritmo KNN, com a seguinte configuração: a) correlação de Spearman como critério de proximidade; b) número de vizinhos mais próximos (k) igual a 1; c) tamanho da subsequência (w) igual a 15 dias. O conjunto de treinamento foi construído utilizando 50 exemplos da classe positiva e 50 da classe negativa. A estratégia de avaliação selecionada foi a *leave-one-out*, obtendo uma taxa de acerto em torno de 78%. Para a série temporal em questão, a classe retornada foi negativa.

Considerações Finais

Neste trabalho foi apresentado o módulo PIPC (PIP Classification) desenvolvido durante o trabalho de estágio. É importante ressaltar que durante o estágio foram estudados e implementados algoritmos existentes na literatura e adaptados ao contexto do projeto CRITIC@, para fins de pesquisa e desenvolvimento. Outra observação é que os algoritmos foram empregados para a extração e análise computacional de séries de valores brutos, sem qualquer tratamento, a fim de que fosse possível utilizar o módulo para qualquer série temporal a ser escolhida.

Os trabalhos futuros incluem uma avaliação mais robusta considerando diferentes domínios de aplicação com apoio de especialistas de domínio.

Referências

MARCACINI, R. M.; REZENDE, S. O. Torch: a tool for building topic hierarchies from growing text collection. In: WORKSHOP ON TOOLS AND APPLICATIONS, 9.; BRAZILIAN SYMPOSIUM ON MULTIMEDIA AND THE WEB, 8., 2010, Belo Horizonte. **Proceedings**... Belo Horizonte: UFMG, 2010. p. 133-135. Webmedia.

[6] Disponível em: <http://www.abramilho.org.br/noticias.php?cod=463>.

MOURA, M. F. **Contribuições para a construção de taxonomias de tópicos em domínios restritos utilizando aprendizado estatístico**. 2009. 137 f. Tese (Doutorado em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, SP.

REZENDE, S. O.; PUGLIESI, J. B.; MELANDA, E. A.; PAULA, M. F. de. Mineracão de dados. In: REZENDE, S. O. (Ed.). **Sistemas inteligentes**: fundamentos e aplicações: Manole, 2003. p. 307-335.

SANCHES, R. A. **Redução de dimensionalidade em séries temporais**. 2006. 92 p. Dissertação (Mestrado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Carlos, SP.