Torch-ETS: análise exploratória de tópicos emergentes com apoio de agrupamento hierárquico de textos*

Bruna Z. Panaggio
Instituto de Ciências Matemáticas e de Computação - ICMC
Universidade de São Paulo - USP
brunazp@grad.icmc.usp.br

Ricardo M. Marcacini
Instituto de Ciências Matemáticas e de Computação - ICMC
Universidade de São Paulo - USP
rmm@icmc.usp.br

Solange O. Rezende
Instituto de Ciências Matemáticas e de Computação - ICMC
Universidade de São Paulo - USP
solange@icmc.usp.br

RESUMO

A análise exploratória de tópicos emergentes é uma tarefa relevante para diversas aplicações, pois permite monitorar a evolução de tópicos extraídos de notícias, artigos científicos e redes sociais. Neste artigo é apresentada a ferramenta Torch-ETS, desenvolvida para suprir a necessidade de analisar tópicos emergentes de forma contextualizada, uma vez que um tópico pode ser importante para determinado segmento, mas menos importante para outros. Para tal, a Torch-ETS permite integrar a análise de tópicos emergentes com métodos de agrupamento hierárquico, na qual a evolução de um tópico pode ser explorada considerando os diferentes temas existentes na coleção textual.

Palavras-chave

Mineração de Textos, Tópicos Emergentes, Clustering

1. INTRODUÇÃO

A análise exploratória de informações publicadas na web é uma tarefa relevante para diversas aplicações. Em especial, o monitoramento de tópicos emergentes a partir de textos têm recebido grande atenção na literatura, como a análise de tendências de tópicos extraídos de notícias, artigos científicos e redes sociais [4].

A análise de tópicos emergentes pode ser visto como um processo de Mineração de Textos (MT) com quatro etapas bem definidas: (1) pré-processamento dos dados textuais; (2) descoberta de tópicos relevantes implícitos em uma coleção textual; (3) construção de gráficos com evolução temporal dos tópicos; e (4) análise de tendências dos tópicos [2]. Nesta última etapa, os usuários geralmente estão interessados em detectar novas tendências, medir a correlação entre tópicos distintos e, também, na identificação de documentos que auxiliam a interpretação de determinados períodos da evolução de um tópico.

A maioria dos trabalhos existentes na literatura adotam uma abordagem "global" para a descoberta de tópicos, ou seja, a análise da tendência dos tópicos é realizada considerando todos os documentos coletados em um período de tempo. No entanto, é de grande utilidade explorar a evolução dos tópicos de forma "local", focando-se em determinados grupos de documentos da coleção textual. Assim, o usuário pode analisar a evolução de um tópico como "*Rising Oil Prices*" em documentos relacionados à *economia* e comparar esta evolução em documentos relacionados à *política*; permitindo verificar se o tópico tem recebido a mesma importância em contextos diferentes.

Métodos de agrupamento hierárquico de documentos é uma forma promissora para auxiliar a análise exploratória de tópicos emergentes de forma local (contextualizada). Esses métodos organizam os documentos em grupos e subgrupos, no qual documentos de um mesmo grupo são similares entre si, mas dissimilares aos documentos de outros grupos. Desta maneira, o usuário pode explorar a coleção textual de forma intuitiva e em diversos níveis de granularidade. O grupo raiz contém todos os documentos da coleção, e a análise dos tópicos emergentes nesse nível é equivalente a uma abordagem tradicional "global". Conforme o usuário navega nos grupos e subgrupos, a evolução dos tópicos é ajustada em relação ao contexto representado no grupo de interesse.

Apesar dos recentes avanços nesse tema, as ferramentas computacionais existentes na literatura não permitem uma análise exploratória contextualizada da evolução dos tópicos. Assim, neste trabalho é apresentada a ferramenta Torch-ETS (*Emerging Topic System*) que disponibiliza um processo de MT para detecção e monitoramento de tópicos emergentes, com o diferencial de utilizar agrupamento hierárquico durante a análise exploratória da evolução dos tópicos. A ferramenta aqui descrita é uma extensão de um trabalho anterior, denominado *Torch - Topic Hierarchies* [1], que realiza agrupamento hierárquico e incremental de textos. Os novos módulos desenvolvidos para detecção e monitoramento de tópicos emergentes foram incorporados à ferramenta Torch, bem como uma nova interface para visualização e exploração dos resultados.

2. A FERRAMENTA TORCH-ETS

A ferramenta Torch-ETS foi desenvolvida para apoiar a análise exploratória de tópicos emergentes provenientes de dados textuais. O desenvolvimento da Torch-ETS é baseado na Linguagem Java e sua arquitetura está organizada em 3 módulos: (1) Representação de Dados Textuais, (2) Mineração Temporal de Textos e (3) Interface Visual para Análise Exploratória. Nas próximas seções, cada módulo da arquitetura serão descritos em mais detalhes, além das principais funcionalidades existentes na ferramenta.

2.1 Representação de Dados Textuais

Documentos textuais coletados na web, por exemplo, notí-

*Trabalho realizado com apoio da FAPESP e CNPq.

WebMedia'11: Proceedings of the 17th Brazilian Symposium on Multimedia and the Web. X Workshop on Tools and Applications.
October 3 -6, 2011, Florianópolis, SC, Brazil.
ISSN 2175-9650.
SBC - Brazilian Computer Society

cias, artigos científicos, registros em blogs e redes sociais, são essencialmente não estruturados (ou semi-estruturados). Esses documentos precisam ser transformados para um formato estruturado, adequado para algoritmos de extração de padrões.

A Torch-ETS realiza a estruturação dos textos considerando dois modelos: *Vector Space Model* (VSM) e *Temporal Space Model* (TSM). No VSM, cada documento é representado por um vetor $\vec{d_i} = \{w_{i1}, w_{i2}, ..., w_{iM}\}$, em que w_{ij} indica o peso do j-ésimo atributo (termo) no documento d_i, por exemplo, um valor de frequência. Para a construção do VSM, são utilizadas técnicas tradicionais da literatura, como a remoção de stopwords, stemming e seleção de atributos. Já no TSM (*Temporal Space Model*), os documentos são divididos em t janelas temporais, em que t é um parâmetro de granularidade temporal definido pelo usuário (i.e. diário, semanal, quinzenal, mensal, etc). Assim, para cada documento é extraída uma série temporal $TSM(d_i) = \{Y(t), t \in T\}$, no qual $Y(t)$ é um valor de correlação entre o conteúdo do documento d_i observado no período t. Ao final, a série temporal indica a cobertura temporal do documento no período de interesse.

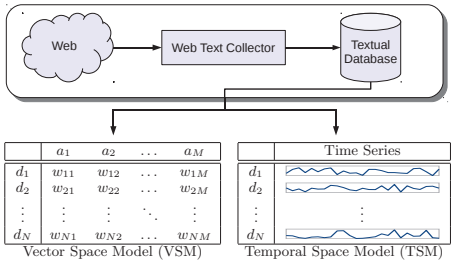

Figura 1: Representação estruturada dos textos

A saída do módulo é composta por duas representações (VSM e TSM), conforme ilustrado na Figura 1. Na próxima etapa, essas representações são utilizadas para a mineração temporal de textos.

2.2 Mineração Temporal de Textos

O módulo de Mineração Temporal de Textos é responsável pelas tarefas de extração de padrões da Torch-ETS. Na Figura 2 é ilustrada uma visão geral das tarefas utilizadas. As representações dos dados textuais obtidas anteriormente são empregadas em duas tarefas: Hierarchical Clustering e BIGRAM Topic Model.

2.2.1 Hierarchical Clustering

O algoritmo de agrupamento hierárquico disponibilizado na Torch-ETS é baseado no Bisecting-Kmeans [5]. Para a formação dos grupos de documentos, é empregada uma medida de similaridade apropriada para dados temporais, definida na Equação 1.

$$sim(d_i, d_j) = \alpha \cdot S_V(d_i, d_j) + (1 - \alpha) \cdot S_T(d_i, d_j) \qquad (1)$$

Nessa medida, dado dois documentos d_i e d_j, a similaridade entre eles é baseada na combinação entre a similaridade do conteúdo $S_V(d_i, d_j)$ e a similaridade temporal $S_T(d_i, d_j)$. A similaridade de conteúdo S_V é calculada por meio da medida Cosseno usando a representação VSM (*Vector Space Model*) dos documentos. Já a similaridade temporal é calculada por meio da correlação de Pearson entre as séries temporais de cada documento na representação TSM (*Temporal Space Model*). O parâmetro α define o peso de cada representação para a similaridade entre os documentos.

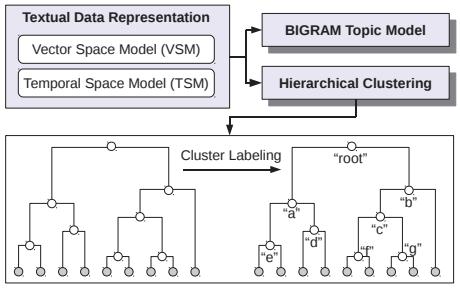

Figura 2: Módulo de mineração temporal dos textos

A saída do algoritmo de agrupamento hierárquico é uma árvore binária conhecida como dendrograma. Para finalizar a etapa de agrupamento, um processo denominado *Cluster Labeling* [3] é aplicado no dendrograma, no qual são identificados descritores em cada grupo da hierarquia, visando facilitar a interpretação do agrupamento.

2.2.2 BIGRAM Topic Model

Para a identificação de tópicos a partir dos textos, adotou-se um conceito similar ao BIGRAM Topic Model [6]. Na Torch-ETS, a construção desse modelo possui duas fases: (1) extração de bigramas (2) extração de uma série temporal para cada bigrama extraído. Cada bigrama selecionado forma um tópico na coleção textual, e sua respectiva série temporal indica a evolução desse tópico no período analisado.

• **Extração de Bigramas:** A extração de bigramas é apoiada por meio da medida Mutual Information (MI), que calcula a dependência entre duas variáveis, conforme a Equação 2. Assim, dados dois termos (atributos) x e y, a probabilidade conjunta $p(x, y)$ é definida como o número de documentos em que x e y ocorrem juntos dividido pelo total de documentos. A probabilidade $p(x)$ (ou $p(y)$) indica o número de documentos em que x (ou y) ocorre dividido pelo total de documentos. Nesse caso, quanto maior o valor de MI, mais relevante é o bigrama para formação de um tópico.

$$MI(x, y) = p(x, y) \cdot log\left(\frac{p(x, y)}{p(x) \cdot p(y)}\right) \qquad (2)$$

Após computar os valores de Mutual Information para um número suficiente de bigramas, os k bigramas com maiores valores de MI são selecionados para a formação dos tópicos.

• **Extração da Série Temporal:** Nesta fase são extraídas as séries temporais que caracterizam a evolução dos tópicos identificados na fase anterior. A partir de um grupo de documentos G, um parâmetro de granularidade temporal t e um bigrama b, a série temporal com a evolução do tópico representado por b é definida na Equação 3.

$$Y(b, G) = \{S_V(b, g_1), S_V(b, g_2), ..., S_V(b, g_t)\} \qquad (3)$$

Nesta equação, $b = \sum_{i \in S} \frac{d_i}{|S|}$, no qual S é o subconjunto de documentos em que b ocorre no modelo VSM. Para cada período temporal $t \in T$, é obtido um representante $g_k = \sum_{i \in L} \frac{d_i}{|L|}$, em que L são os documentos no modelo VSM publicados no período t. Por fim, os valores da série são computados por meio da similaridade S_V (Cosseno) entre o tópico b e g_k, em cada período t. Desta forma, se um

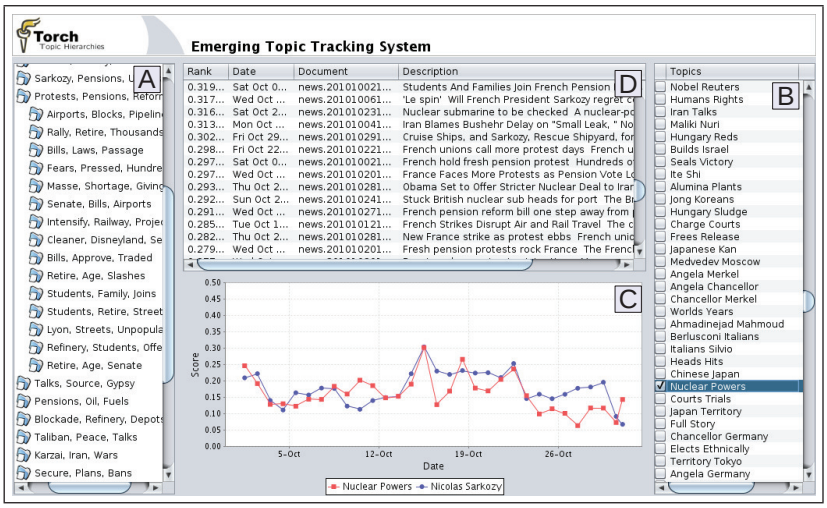

Figura 3: Interface de análise exploratória da Torch-ETS

determinado bigrama b ocorre com alta frequência em um período, o valor da medida se aproxima de 1; caso contrário é proximo de 0.

2.3 Interface Visual para Análise Exploratória de Tópicos Emergentes

Após a representação estruturada dos dados textuais e a mineração temporal dos textos, é possível explorar os resultados obtidos a partir da Interface de Visualização. Na Figura 3 é ilustrado um exemplo da interface a partir de uma base de notícias da web. Nesse módulo, o usuário tem acesso às principais funcionalidades da ferramenta, conforme descritas a seguir.

• **Organização Hierárquica dos Textos**: permite selecionar grupos de documentos relacionados a um mesmo contexto, de acordo com o resultado do agrupamento hierárquico, ilustrado na Figura 3(A).

• **Análise de Tópicos Emergentes**: na Figura 3(B), são listados os tópicos emergentes identificados a partir de um grupo de documentos. Os tópicos são ordenados de acordo com sua relevância para o contexto de interesse. Assim, a lista de tópicos emergentes é atualizada conforme o usuário navega na organização hierárquica, permitindo realizar a análise contextualizada dos tópicos emergentes.

• **Evolução dos Tópicos**: ao selecionar os tópicos emergentes de interesse, o usuário pode visualizar a evolução desses tópicos no tempo. Na Figura 3(C) é apresentada uma comparação de dois tópicos selecionados. O valor "*Score*" representa a importância do tópico no período.

• **Recomendação de Documentos**: uma vez que o usuário identificou um tópico emergente de interesse, a ferramenta Torch-ETS recomenda uma lista ordenada dos documentos mais relacionados ao tópico. Essa opção é útil em análise exploratória, uma vez que facilita a interpretação de eventos relatados nos textos. Na Figura 3(D) é ilustrado um exemplo dessa funcionalidade, no qual os primeiros documentos da lista são identificados como os mais relevantes para os tópicos selecionados.

3. CONSIDERAÇÕES FINAIS

Neste trabalho foi apresentada a ferramenta Torch-ETS, com uma breve descrição da sua arquitetura e as principais funcionalidades. A partir de um conjunto de documentos, a ferramenta obtém uma representação estruturada para os dados textuais, realiza um processo de mineração temporal dos textos e, por fim, apresenta uma interface de visualização para que o usuário explore tópicos emergentes nos textos. Além disso, a Torch-ETS disponibiliza a análise contextualizada dos tópicos emergentes, na qual a evolução de um tópico pode ser analisada em diferentes níveis e temas existentes na coleção textual. A ferramenta está disponível em `http://sites.labic.icmc.usp.br/torch/`, incluindo a base textual para os exemplos aqui ilustrados.

Como trabalhos futuros, espera-se estender o modelo de identificação de tópicos baseado em bigramas para n-gramas, além das respectivas medidas para calcular a relevância de n-gramas. Novas técnicas para visualização dos resultados também podem ser incorporadas, principalmente técnicas que integram representações de conteúdo e temporal em uma mesma projeção visual dos dados.

4. REFERENCES

[1] R. M. Marcacini and S. O. Rezende. Torch: a tool for building topic hierarchies from growing text collection. In *WFA'2010: IX Workshop de Ferramentas e Aplicações (Webmedia 2010)*, pages 1–3, 2010.

[2] Q. Mei and C. Zhai. Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In *SIGKDD'05: Proceedings of the 11th Int. Conf. on Knowledge Discovery and Data Mining*, pages 198–207. ACM, 2005.

[3] M. Moura, R. Marcacini, and S. Rezende. Easily labeling hierarchical document clusters. In *WAAMD'08: Workshop de Algoritmos e Aplicações em Mineração de Dados*, pages 37–45, 2008.

[4] R. Schult and M. Spiliopoulou. Discovering emerging topics in unlabelled text collections. In *Advances in Databases and Information Systems*, pages 353–366. Springer, 2006.

[5] M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques. In *KDD'2000: Workshop on TM*, pages 1–20, 2000.

[6] H. Wallach. Topic modeling: beyond bag-of-words. In *ICML'06: Proceedings of the 23rd Int. Conf. on Machine learning*, pages 977–984. ACM, 2006.