Identificando o Assunto dos Documentos em Coleções Textuais Utilizando Termos Compostos

Fabiano Fernandes dos Santos
Solange Oliveira Rezende
Instituto de Ciências Matemáticas e de Computação
USP - Universidade de São Paulo - São Carlos, Brasil
Email: {fabianof,solange}@icmc.usp.br

Veronica Oliveira de Carvalho
Instituto de Geociências e Ciências Exatas
UNESP - Univ Estadual Paulista - Rio Claro, Brasil
Email: veronica@rc.unesp.br

***Resumo*—Diferentemente dos problemas de recuperação de informação, nos quais o usuário conhece o que ele está procurando, às vezes o usuário precisa *compreender* de forma mais geral os assuntos abordados na coleção para *explorar* os documentos de interesse. Para cada grupo ou tópico obtido, um conjunto de descritores é selecionado entre os termos da coleção e cabe ao usuário identificar o assunto de cada grupo a partir da lista de descritores apresentada. Normalmente, o conjunto de descritores é composto por termos simples. Entretanto, muitos termos possuem significado próprio quando combinados entre si. Produzir uma lista de termos que já considere na sua construção o uso de termos compostos pode diminuir o esforço necessário para a compreensão dos assuntos identificados. Neste artigo é proposta uma abordagem para identificação de assuntos em coleções de documentos que combina técnicas de regras de associação e de agrupamento de dados. As regras de associação são aplicadas para extrair termos compostos formando o contexto local da relação entre os termos. Essas regras são representadas em uma estrutura *bag-of-words* cujas dimensões são as mesmas da *bag-of-words* produzida pela coleção de documentos e são agrupadas, formando o contexto geral das relações. A ideia é que a informação da vizinhança dos termos compostos extraídos ajudam a identificar (a) termos diferentes utilizados em um mesmo contexto ou com mesmo sentido e (b) termos idênticos mas que são usados em contextos diferentes ou com significados diferentes. Os resultados da avaliação indicam que o uso de termos compostos com a abordagem proposta melhora a identificação de assuntos nas coleções de documentos avaliadas.**

I. Introdução

Cada vez mais as atividades de organização e recuperação de informações são mediadas por sistemas informatizados destinados a esse fim [1]. Compreender a estrutura e o conteúdo de grandes coleções de documentos ainda é um grande problema de pesquisa. Muitas vezes o usuário precisa *compreender* de forma mais geral os assuntos abordados na coleção para *explorar* os documentos de interesse, diferentemente dos problemas de recuperação de informação, nos quais o usuário tem conhecimento sobre o que ele está procurando [2]. O agrupamento de documentos e a extração de tópicos são algumas das técnicas interessantes para apoiar o usuário nessa tarefa, pois descrevem a coleção de documentos em uma forma que revela sua estrutura interna e as interrelações, ainda que de formas distintas. Nessas técnicas, são obtidos grupos de termos ou documentos relacionados de acordo com um critério estabelecido. Para cada grupo ou tópico obtido, um conjunto de descritores é selecionado entre os termos da coleção. Por exemplo, o grupo descrito pelos termos "computador, processador, *internet*, *smartphone*, sistema" provavelmente trata do assunto "*tecnologia*". Cabe ao usuário identificar o assunto de cada grupo a partir da lista de descritores apresentada.

A identificação de uma lista de descritores relevante é fundamental para o bom entendimento do assunto apresentado. Normalmente, a lista de descritores é composta por termos simples, uma vez que grande parte das técnicas utiliza como entrada do processo a representação dos documentos extraída pelo modelo *bag-of-words*, em que cada documento é representado por um vetor de pesos correspondentes aos atributos do texto. Os atributos são formados por termos simples extraídos da coleção de documentos. Entretanto, muitos termos possuem significado próprio quando combinados entre si, como "*inteligência*" e "*artificial*". O modelo *bag-of-words* tem a limitação de assumir que os termos são independentes e ignora totalmente a relação entre eles [1], [3]. Produzir uma lista de termos que já considere na sua construção o uso de termos compostos pode diminuir o esforço do usuário necessário para a compreensão de cada grupo ou tópico extraído.

Extrair termos compostos de forma eficiente para incorporar na representação dos documentos ainda é um grande desafio de pesquisa, e tem atraído grande atenção recentemente [3]–[5]. Diversos modelos para representação de documentos foram propostos na literatura para capturar a dependência entre os termos, destacando-se os modelos baseados em frases ou termos compostos [4], [6], [7] e os modelos baseados em dimensões latentes, como o LSI [8] e os de extração de tópicos [9], [10] e suas extensões [5], [11]–[14]. Entre estes, os modelos de extração de tópicos apresentam os resultados mais interessantes. A revisão da literatura aponta para ganhos significativos de interpretabilidade dos tópicos extraídos com a adição da informação de dependência dos termos. As regras de associação podem ser utilizadas para aproximar os resultados de técnicas para identificar os assuntos da coleção de documentos que utilizam termos compostos e evitar algumas de suas deficiências [1], [6], [7]. Regras de associação representam correlações ou coocorrências entre itens, e possuem duas vantagens para esse contexto. O algoritmo para geração de regras de associação é muito eficiente. Como, em geral, são necessárias apenas regras com 2 ou 3 termos, o que é suficiente para objetivos práticos, o algoritmo somente percorre a coleção de documentos 2 ou 3 vezes. Apesar disso, pouca pesquisa tem sido feita nessa direção [1].

Neste trabalho, é proposta uma abordagem não-supervisionada para identificação de assuntos em coleções de documentos que combina técnicas de regras de associação e

de agrupamento de dados. Na abordagem proposta, as regras de associação são aplicadas para extrair termos compostos formando o contexto local da relação entre os termos. As regras de associação de toda a coleção de documentos são agrupadas, formando o contexto geral das relações. A ideia é que a informação da vizinhança dos termos compostos extraídos ajudam a identificar (a) termos diferentes utilizados em um mesmo contexto ou com mesmo sentido e (b) termos idênticos mas que são usados em contextos diferentes ou com significados diferentes. Para cada grupo, algumas regras são selecionadas para formar o conjunto de descritores, que é composto por termos simples e compostos. Esses grupos com descritores representam os assuntos identificados na coleção e fornecem uma estrutura para apoiar o usuário na compreensão e exploração da coleção de documentos.

As principais contribuições dessa proposta são: (i) Um novo método que apóia a exploração de grandes coleções de documentos combinando contexto local e geral das relações entre os termos para extrair novos atributos representativos que permite a descrição dos documentos utilizando poucas dimensões significativas; (ii) Uma abordagem que permite representar as regras de associação no mesmo espaço de atributos dos documentos. As regras de associação extraídas são representadas em uma *bag-of-words* cujas dimensões são as mesmas da *bag-of-words* produzida pela coleção de documentos, e o peso de cada termo é dado pela sua frequência nas transações cobertas pela regra de associação. Essa representação permite agrupar dependências compostas por termos distintos mas que possuem um significado semelhante. Esses grupos formam o contexto geral das relações; (iii) A construção de um conjunto de descritores para cada grupo composto por termos simples e compostos combinados melhora o entendimento dos assuntos gerais da coleção.

II. Trabalhos relacionados

Grande parte das técnicas para identificação dos assuntos da coleção utilizam a representação produzida pelo modelo *bag-of-words* como entrada do processo. Esse modelo representa cada documento como um vetor de termos distintos que aparecem na coleção. Cada componente do vetor representa o peso de cada termo em cada documento da coleção. Seja $D = \{d_1, d_2, ..., d_m\}$ uma coleção com m documentos, e $T = \{t_1, t_2, ..., t_n\}$ o conjunto de n termos distintos da coleção. Cada documento d_j é representado por um vetor de termos $\vec{d_j} = \{w_{1j}, w_{2j}, ..., w_{nj}\}$, no qual cada peso w_{ij} quantifica a importância do termo $t_i \in T$ para o documento $d_j \in D$. Para os termos da coleção que não estão presentes no documento d_j, $w_{ij} = 0$. Nesse modelo, tradicionalmente o peso w_{ij} representa a medida da frequência do termo no documento ou uma variação desse esquema [1]. A coleção de documentos é então representada pela matriz W com dimensões $m \times n$, conhecida na literatura como matriz documento-termo. Cada linha de W corresponde a um documento em D, isto é, o vetor $\vec{d_j}$, e cada coluna descreve a distribuição de cada termo na coleção de documentos.

Dentre as técnicas encontradas na literatura para identificar os assuntos de uma coleção de documentos, as propostas baseadas em extração de dimensões latentes, como a Análise da Semântica Latente [8] (*Latent Semantic Analysis* - LSA) e as de extração de tópicos [9], [10], se destacam pela qualidade dos resultados obtidos e pela boa interpretabilidade das dimensões extraídas. O modelo *Latent Dirichlet Allocation* (LDA) [10] é uma das técnicas mais proeminentes para extração de tópicos. Em [10], propõe-se um modelo generativo que descreve uma coleção de documentos à partir de um conjunto reduzido de tópicos. Esse modelo se torna atrativo por descobrir grupos de termos que aparecem frequentemente juntos nos documentos [2], [13]. O trabalho apresentado em [11] é um dos primeiros a incorporar termos compostos no processo de extração de tópicos com LDA. Utilizou-se em [11] modelos de linguagem hierárquica de Dirichlet para extender o algoritmo original do LDA e considerar, durante a inferência dos tópicos, o peso do termo avaliado condicionalmente ao peso do termo anterior a ele. Os autores de [14] argumentam que ainda que os resultados dessa e de outras propostas baseadas em modificações do algoritmo de inferência sejam teoricamente interessantes, o algoritmo de inferência fica computacionalmente mais complexo e pode inviabilizar sua aplicação para o usuário final. Os autores de [14] avaliam a contribuição do uso de termos compostos no modelo de tópicos obtidos aplicando o modelo LDA clássico. Foi proposto o uso de um processo de extração de termos compostos antes da etapa de extração de tópicos. Os termos compostos são incluídos na representação dos documentos e os tópicos são obtidos com o processo tradicional do LDA. Propostas mais recentes integram técnicas de extração de *itemsets* frequentes para explorar a informação da dependência dos termos na extração dos tópicos sem aumentar demais a complexidade do processo [5], [12], [13], mas pouca avaliação foi feita nestes trabalhos dos ganhos qualitativos dos tópicos obtidos.

O uso de regras de associação para identificar dependência entre termos tem se mostrado uma estratégia interessante na literatura, como apresentado nos trabalhos [5]–[7], [12], [13], levando a bons resultados nas aplicações de mineração de textos e gerando atributos que possuem uma boa interpretabilidade. Em geral, as regras de associação são utilizadas para obter uma representação semelhante a *bag-of-words*. Essa representação é utilizada, então, como entrada para um algoritmo LDA tradicional, que não está adaptado para considerar as características de atributos compostos. Dessa forma, ele não explora de forma significativa a informação adicionada com os novos atributos obtidos.

III. Abordagem proposta

A abordagem proposta explora termos compostos em um contexto local e geral da relação entre termos para identificar assuntos em coleções de documentos. Ela explora a relação existente entre os termos obtidas no contexto local, o que permite atacar problemas importantes como a polissemia mesmo quando o termo avaliado é composto. Por exemplo, os termos compostos "mineração de dados" e "extração de conhecimento" podem ser empregados nos textos com significado semelhante, mas são formados por termos distintos. Ainda, um deles pode ocorrer com frequência muito maior que o outro, tornando difícil identificar esse tipo de relação pela comparação estatística das distribuições. A ideia é que a informação da vizinhança dos termos, obtida pela análise do conjunto de transações que cada regra cobre, ajudam a identificar termos diferentes utilizados em um mesmo contexto ou com mesmo sentido e termos idênticos mas que são usados em contextos diferentes ou com significados diferentes.

Na Figura 1 são apresentados os principais passos da

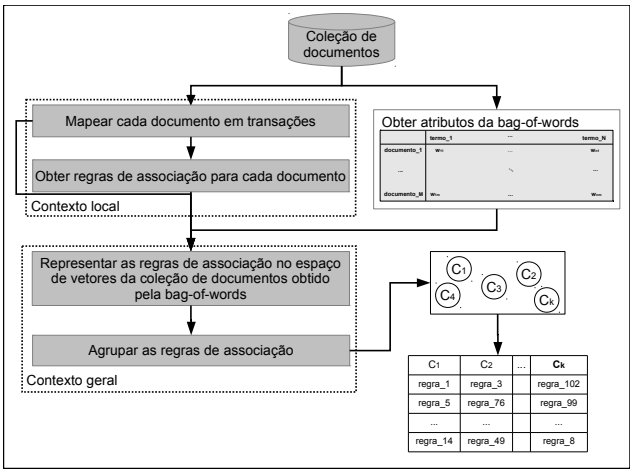

Figura 1. Visão geral da abordagem proposta.

abordagem proposta. O contexto local captura a correlação entre os termos em cada documento pela coocorrência obtida com a extração de regras de associação. Para capturar o contexto geral, foi proposta uma representação intermediária para as regras de associação que permite representá-las no espaço de vetores original dos documentos. Cada regra de associação extraída é representada por um vetor de termos distintos que aparecem na coleção de documentos, e o peso de cada termo é dado pela sua frequência nas transações cobertas pela regra de associação. Isso permite aplicar medidas de similaridade como a medida de cosseno, para determinar o contexto geral de cada relação obtida considerando todas as outras relações da coleção. Os grupos obtidos formam, então, o conjunto de assuntos identificados dos documentos.

Antes de iniciar o processo, recomenda-se o pré-processamento da coleção de documentos com a simplificação de palavras, remoção de *stopwords*, e a limpeza e padronização dos documentos como descritos em [15].

A. Contexto Local da Relação

O contexto local da relação entre os termos é obtido aplicando-se as ideias propostas no modelo *bag-of-related-words* [7], que oferece um processo eficiente para extração de termos compostos utilizando regras de associação. Nesse modelo, as regras de associação são extraídas para cada documento e, após selecionadas, são utilizadas para construir os termos compostos da coleção. Como cada documento é processado de forma independente, os termos compostos são obtidos para o contexto local da relação. Dos quatro passos propostos pelo modelo *bag-of-related-words*, foram utilizados dois deles na abordagem proposta: (1) Mapear os documentos em transações; (2) Extrair as regras de associação das transações de cada documento. Ainda, optou-se na abordagem proposta pelo mapeamento em janelas deslizantes pois foi aquele que apresentou melhores resultados segundo [7].

No mapeamento dos documentos em transações utilizando janelas deslizantes, a primeira transação contém apenas a primeira palavra do documento, a segunda contém as duas primeiras palavras, e assim por diante, até que a janela contenha o número de palavras igual ao tamanho definido "$tamanho\_janela$". Após isso, a janela desliza uma palavra e considera as próximas "$tamanho\_janela$" palavras do documento. O resultado desse processo é a coleção $D_C = \{d_{C1}, d_{C2}, ..., d_{Cm}\}$ em que d_{Cj} corresponde as transações obtidas pelo mapeamento do documento $d_j \in D$. Cada conjunto de transações $d_{Cj} \in D_C$ é processado por um algoritmo de geração de regras de associação (como o Apriori [1]). Define-se os valores $supmin$ e $confmin$ de suporte mínimo e confiança mínima, respectivamente, utilizados pelo algoritmo de geração de regras de associação para cada d_{Cj}. Os autores de [7] propuseram o cálculo automático do valor $supmin$ para cada documento, cuja fórmula é:

$$supmin(d_{Cj}) = \frac{\left(\sum_{\forall t_i \in A} w_{ij}\right) / |A|}{|d_{Cj}|} \quad (1)$$

na qual A é o conjunto de termos $t_i \in T$ para o documento d_j tal que $w_{ij} > 0$, $|A|$ é o número total de termos de d_j e $|d_{Cj}|$ é o número total de transações do documento d_j. Segundo os autores de [7], esse cálculo apresenta resultados comparáveis aos obtidos com a definição manual do valor de $supmin$, e possibilita tirar do usuário a responsabilidade de definir esse parâmetro. Utilizou-se o valor de $confmin = 0$, permitindo gerar todas as regras de associação possíveis. As regras de associação que serão exploradas ao longo do processo são selecionadas utilizando um valor de corte para uma medida objetiva. O resultado desse processo é a coleção $D_R = \{d_{R1}, d_{R2}, ..., d_{Rm}\}$ em que d_{Rj} corresponde as regras de associação obtidas pelo processamento das transações d_{Cj} do documento d_j.

Por fim, calcula-se o valor de uma medida objetiva da literatura para filtrar as regras de associação. A seleção da medida objetiva é um passo importante, já que cada uma delas possui uma semântica própria e tem influência no tipo de termo composto que será obtido. Em [7], também foi proposto o cálculo automático do valor de corte para uma dada medida objetiva em cada documento, cuja fórmula é:

$$min\_medida(d_{Rj}) = \frac{\sum_{\forall r_{jz} \in d_{Rj}} valor\_da\_medida(r_{jz})}{|d_{Rj}|} \quad (2)$$

na qual r_{jz} é a z-ésima regra de associação em d_{Rj}, $valor\_da\_medida(r_{jz})$ é o valor calculado da medida objetiva para a regra de associação r_{jz} e $|d_{Rj}|$ é o número de regras de associação em d_{Rj}. Com o valor de corte definido, todas as regras de associação com o valor de medida objetiva maior ou igual ao valor de corte são selecionadas. O resultado desse processo é a coleção $D_S = \{d_{S1}, d_{S2}, ..., d_{Sm}\}$ em que d_{Sj} corresponde as regras de associação selecionadas de d_{Rj} com valor de medida objetiva maior do que min_medida_j.

Vale ressaltar que o modelo *bag-of-related-words* em sua proposta original também gera atributos compostos por um único termo. Isso é possível ao processar regras obtidas do tipo $\varnothing \Rightarrow atributo$, que é formada por apenas um item. Esse formato também foi mantido na proposta apresentada aqui, uma vez que grande parte dos termos utilizados nos documentos possuem sentido próprio como termo simples. Como é sugerido em [1], em geral, são necessárias apenas regras com 2 ou 3 termos. Assim, optou-se pela geração de regras de associação com, no máximo, três itens. Por fim, as regras de associação avaliadas aqui podem apresentar vários itens no antecedente mas apenas um item no consequente devido a otimizações de implementação do algoritmo para geração de regras de associação.

B. Contexto Geral da Relação

Uma das dificuldades ao agrupar regras de associação está no cálculo da similaridade entre duas regras. Inspirados na proposta de [16], que representa o conjunto de *itemsets* frequentes extraídos no mesmo espaço de representação dos documentos, propõe-se aqui uma representação intermediária das regras de associação obtidas utilizando os atributos extraídos para formar a representação *bag-of-words* da coleção, com o conjunto de termos $T = \{t_1, t_2, ..., t_n\}$ dos documentos. Entretanto, em nossa proposta, o peso de cada atributo para cada regra de associação pode ser obtido ao explorar o conjunto de transações cobertas pela regra de associação, ou seja, o conjunto de transações que gerou aquela regra. Assim, considera-se no processo o contexto local da relação e sua vizinhança para encontrar o contexto geral das relações. De acordo com a hipótese distribucional [17], termos que ocorrem em contextos similares tendem a ter significados semelhantes. Assim, cada regra r_{jz} é representada por um vetor de termo $\vec{r_{jz}} = \{p_{1jz}, p_{2jz}, ..., p_{njz}\}$, no qual cada pelo p_{ijz} quantifica a importância do termo $t_i \in T$ para a regra de associação r_{jz}, de forma que p_{ijz} representa a frequência do termo t_i nas transações em $d_{Tj} \in D_T$ cobertas pela regra r_{jz}. Uma vez obtida a representação, aplica-se um algoritmo de agrupamento de textos utilizando uma medida de similaridade. Assim como em outros métodos, o número K de grupos gerados deve ser informado pelo usuário.

IV. Experimentos e resultados

Nesta seção apresentamos e discutimos a avaliação experimental da abordagem proposta para identificação de assuntos em coleções textuais. Para comparação, foi escolhido o modelo LDA tradicional. Os tópicos do modelo LDA foram produzidos à partir de uma *bag-of-words* formada por termos simples e também à partir da representação obtida pelo modelo *bag-of-related-words* [7], formada por termos simples e compostos.

A. Configuração experimental

Coleções de documentos: Para os experimentos, utilizou-se duas coleções de documentos em inglês[1], como apresentado na Tabela I. A coleção Re8 é composta por textos jornalísticos, os quais representam uma porção significativa dos tipos de documentos digitais disponíveis. A coleção ACM é composta por textos de artigos científicos, que também representam documentos digitais de grande interesse.

Tabela I. Descrição das coleções de textos. A quantidade de termos corresponde ao valor na coleção pré-processada.

Coleção	# classes	# docs	# termos	Descrição
Re8	8	7674	17336	Coleção de notícias das 8 classes mais frequentes da coleção Reuters-21578
ACM	40	3491	222713	Coleção de artigos científicos de conferências de diferentes áreas da computação extraídos do repositório digital da ACM

Pré-processamento: Todas as bases de textos foram pré-processadas utilizando o mesmo método. Foi realizada uma padronização dos textos, removendo-se números, símbolos e também as *stopwords*. Os termos foram obtidos pela redução das palavras ao seu *stem* aplicando-se o algoritmo de Porter [15]. Os termos foram selecionados removendo-se 5% dos termos mais frequentes e 5% dos termos menos frequentes.

Técnicas utilizadas: Comparou-se a abordagem proposta com diferentes medidas objetivas de regras de associação, o modelo produzido pelo LDA considerando a *bag-of-words* tradicional como entrada e o modelo produzido pelo LDA considerando a *bag-of-related-words* como entrada. Este último modelo permite comparar o impacto do uso de termos compostos no processo tradicional do LDA. A extração dos termos compostos antes da execução do LDA é apontado por [14] como uma alternativa viável para inclusão de dependência de termos no processo e que obtém bons resultados.

Configuração de parâmetros O modelo *bag-of-related-words* foi obtido utilizando a ferramenta FEATuRE[2]. Com base nos resultados apresentados em [7], optou-se pelo uso do suporte automático e pelo valor médio da medida objetiva para selecionar os atributos. Ainda de acordo com os resultados em [7], foi utilizada a medida objetiva *Kappa* para construir a representação *bag-of-related-words* para essa avaliação. Para a construção das transações, adotou-se a janela de tamanho 5, que apresentou os melhores resultados segundo [7].

Tabela II. Valores dos parâmetros dos métodos avaliados.

LDA	
Representação de entrada	*bag-of-words*, *bag-of-related-words*
Hiperparâmetros α e β	estimados automaticamente pela ferramenta
Iterações para a amostragem de Gibbs	10000
Número de tópicos (k)	50, 100, 150
Abordagem proposta	
Tamanho da janela	5
Suporte/confiança mínimos por documento	supmin Automático (Equação 1) / confmin 0
Medidas objetivas	Added Value, Certainty Factor, Collective Strength, Confiança, Conviction, ϕ-Coefficient, Gini Index, IS, J-Measure, Kappa, Klosgen, Lambda, Laplace, Lift, Mutual Information LHS, Novelty, Odds Ratio
Valor de corte para a medida objetiva por documento	Automático (Equação 2)
Agrupamento das regras	Bi-Secting K-Means com medida de similaridade de cosseno
Número de grupos (k)	50, 100, 150

Conforme apresentado na Tabela II, para avaliar as descrições obtidas, foram utilizados os valores de $k = 50$, $k = 100$ e $k = 150$ tanto para o LDA quanto para a abordagem proposta, como utilizado em [18]. Ainda, os modelos LDA foram obtidos utilizando a ferramenta MALLET[3]. Para cada tópico, foram selecionados os dez termos com maior probabilidade para formar o conjunto de descritores. Para a abordagem proposta, as transações foram obtidas utilizando uma janela de tamanho 5, mesmo valor utilizado para obter a representação *bag-of-related-words*. As regras de associação foram obtidas e, para selecioná-las, avaliou-se as medidas objetivas apresentadas na Tabela II. Detalhes sobre as características das medidas objetivas utilizadas e suas propriedades podem ser encontrados em [19], [20]. Para o agrupamento das regras, foi aplicado o algoritmo *Bi-Secting K-Means*[4], e a medida de similaridade utilizada foi a cosseno. Para cada grupo, foram selecionadas as dez melhores regras, de acordo com a medida objetiva utilizada no passo anterior, para formar o conjunto de descritores. Para

[1]Estas coleções podem ser acessadas em http://sites.labic.icmc.usp.br/text_collections/ .

[2]Disponível em http://sites.labic.icmc.usp.br/feature/ .

[3]Disponível em http://mallet.cs.umass.edu/ .

[4]Implementado na ferramenta Cluto. Disponível em http://glaros.dtc.umn.edu/gkhome/views/cluto .

efeito de comparação, os itens da regra foram unidos de forma a apresentar um formato semelhante aos de outros modelos. Por exemplo, a regra $inteligencia \Rightarrow artificial$ forma o termo "inteligencia_artificial".

B. Medidas de avaliação

Avaliação do termo intruso: Na detecção de termos intrusos [18], uma lista com $N+1$ termos[5] é apresentada, sendo os N termos com maior probabilidade para o tópico e um "termo intruso" escolhido aleatoriamente entre aqueles com baixa probabilidade no tópico de interesse, mas alta probabilidade em outro tópico. O objetivo é identificar na lista qual é o termo intruso que não pertence ao tópico em questão. Na proposta original, a tarefa de detecção de termos intrusos é realizada por um grupo de especialistas [18]. Neste trabalho, utilizou-se a avaliação automática para detecção de termo intruso implementada na ferramenta *topic_interpretability*[6]. Para cada termo descritor do tópico, incluindo o termo intruso, calcula-se um peso baseado na sua coocorrência com todos os outros termos descritores deste tópico. Esses pesos são utilizados como atributos para treinar um classificador que produz um modelo de regressão. Este classificador é então utilizado para identificar o termo intruso em todos os casos apresentados. Nessa avaliação, tópicos com maior interpretabilidade ou mais coerentes são aqueles em que é fácil identificar o termo intruso. Quanto maior é a quantidade de termos intrusos detectados, melhor é o modelo produzido. Para calcular o peso de cada coocorrência, utilizou-se como base a medida Informação Mútua Pontual (*Pointwise Mutual Information - PMI*). O valor de PMI de cada termo t_i é obtida pela soma dos valores de PMI entre t_i e todos os outros termos da lista de descritores, incluindo o termo intrudo. A TI_PMI para cada termo t_i do conjunto de descritores em relação a todos os outros descritores t_j do mesmo grupo ou tópico é:

$$TI\_PMI(t_i) = \sum_{j=1}^{N} log(\frac{p(t_i, t_j)}{p(t_i) * p(t_j)}),\ i \neq j$$

Foram selecionados como descritores para cada tópico ou grupo 10 termos, e foi adicionado a lista 1 termo intruso. O termo intruso é selecionado aleatoriamente em uma lista formada pelos 10 termos com menor peso para o tópico de interesse e que também tenham sido selecionados como descritores de algum outro tópico ou grupo. Para as descrições obtidas pelo modelo LDA, os 10 termos com menor peso são aqueles com menor probabilidade no tópico. Para a abordagem proposta, os termos com menor peso são aqueles com menor valor na medida objetiva utilizada para selecionar as regras.

Avaliação da coerência observada: Os autores de [21] definiram a interpretabilidade de um tópico baseado na avaliação feita por usuários especialistas da coerência observada dos N termos selecionados como descritores do tópico. Na proposta original, os tópicos extraídos são apresentados para um grupo de especialistas que, seguindo um conjunto de instruções padronizadas, devem avaliar a qualidade do tópico quanto a sua utilidade em uma escala de 3 pontos onde a nota 3 indica um tópico útil (coerente) e a nota 1 indica um tópico pouco útil (menos coerente). Entre os métodos automáticos avaliados, medida PMI foi apontada como a que mais se aproxima da avaliação feita pelos especialistas, e pode ser utilizada para automatizar a avaliação da coerência do tópico considerando os termos selecionados como descritores e sua coocorrência em relação a uma coleção de referência. O processo automatizado de avaliação da coerência observada implementado na ferramenta *topic-interpretability* também foi utilizado neste trabalho. Nessa implementação, a coerência observada do k-ésimo grupo ou tópico, considerando seu conjunto de descritores, é dada pela soma do valor de PMI de todas as combinações de pares de termos da lista de descritores. Assim, medida CO_PMI:

$$CO\_PMI(C_k) = \sum_{j=2}^{N} \sum_{i=1}^{j-1} log(\frac{p(t_j, t_i)}{p(t_i) * p(t_j)})$$

Uma vez que algumas dimensões produzidas pelos métodos podem não ser representativas, já que não foi utilizada nenhuma heurística para determinar um valor ideal de K para os métodos comparados, considerou-se que os tópicos no quartil superior são os mais interessantes. Assim, foram selecionados para comparação um total de 25% dos grupos ou tópicos que foram melhores avaliados pela medida de Coerência Observada.

Configuração da ferramenta de avaliação: As avaliações foram realizadas de forma automática utilizando a ferramenta *topic-interpretability*. Tanto para o caso da detecção de termos intrusos quanto para o caso da coerência observada, a ferramenta utiliza uma coleção de referência para calcular a coocorrência entre os termos selecionados como descritores necessária para o cálculo da PMI. Neste artigo, o *corpus* de referência utilizado foi a Wikipédia em inglês[7]. Todos os artigos foram pré-processados seguindo o mesmo procedimento utilizado para os documentos das coleções avaliadas. Para contabilizar as coocorrências, a ferramenta foi configurada para considerar termos que ocorreram em um mesmo artigo, independente da distância entre eles.

C. Resultados e discussão

Avaliação do termo intruso: Os resultados obtidos para a coleção Re8 e para a coleção coleção ACM são apresentados na Tabela III. Considerando todos os casos, a abordagem proposta foi superior aos modelos gerados pelo LDA. Para a coleção ACM, o classificador utilizado acertou todos ou quase todos os casos para a maioria das medidas objetivas. Na coleção Re8, os melhores valores obtidos pela abordagem proposta foram, aproximadamente, o dobro dos valores obtidos pelo melhor modelo LDA, ainda que não tenham acertado a totalidade dos casos. Destaca-se em ambas as coleções os resultados obtidos pela medida ϕ-*Coefficient*, que foi a melhor medida em 5 dos 6 casos, sendo a medida objetiva com resultados mais estáveis. Essa medida objetiva indica o grau de associação (ou correlação) entre os itens do antecedente e do consequente da regra de associação e, de acordo com [19], sua semântica é muito semelhante a da medida estatística χ^2. É importante observar que, no caso das medidas objetivas, muitas delas são propostas visando identificar esse mesmo tipo de relação, ou seja, possuem semântica semelhante, porém possuem propriedades bem distintas. Por exemplo, as medidas *Mutual Information LHS* e *Collective Strenght* possuem semântica semelhante a ϕ-*Coefficient*, porém seus resultados foram bem distintos nessa avaliação. Destaca-se também a

[5] N é o número de palavras do tópico apresentadas para o usuário. Normalmente utiliza-se $N = 10$.

[6] Disponível em https://github.com/jhlau/topic_interpretability/.

[7] Versão extraída em 5 de agosto de 2013.

Tabela III. RESULTADOS PARA AS COLEÇÕES RE8 E ACM DA TAREFA DE DETECÇÃO DE TERMOS INTRUSOS (TI) ORDENADOS PELA QUANTIDADE ABSOLUTA DE TERMOS ENCONTRADOS CORRETAMENTE.

Coleção Re8

K=50		K=100		K=150	
Configuração	TI	Configuração	TI	Configuração	TI
ϕ-Coefficient	38	ϕ-Coefficient	52	Laplace	71
Odds Ratio	28	Laplace	42	Confiança	66
Novelty	24	Odds Ratio	40	ϕ-Coefficient	64
Gini Index	24	Confiança	39	Odds Ratio	57
Kappa	24	Novelty	32	LDA + *bag-of-words*	38
Klosgen	22	Added Value	30	Klosgen	37
Added Value	22	Gini Index	27	Kappa	36
Certainty Factor	19	Kappa	27	Gini Index	34
Lift	17	LDA + *bag-of-words*	25	LDA + *bag-of-related-words*	34
Laplace	17	Collective Strength	24	Collective Strength	33
Confiança	15	Certainty Factor	24	Novelty	33
Collective Strength	13	LDA + *bag-of-related-words*	24	Conviction	33
Conviction	12	Lambda	24	Added Value	31
IS	12	Lift	22	Lambda	31
Mutual Information LHS	10	Klosgen	22	Certainty Factor	30
LDA + *bag-of-related-words*	9	Conviction	22	J-Measure	27
Lambda	9	J-Measure	17	Lift	26
J-Measure	7	Mutual Information LHS	13	IS	26
LDA + *bag-of-words*	3	IS	11	Mutual Information LHS	21

Coleção ACM

K=50		K=100		K=150	
Configuração	TI	Configuração	TI	Configuração	TI
Confiança	50	Odds Ratio	100	ϕ-Coefficient	150
Novelty	50	Kappa	100	Odds Ratio	149
ϕ-Coefficient	50	ϕ-Coefficient	100	Laplace	148
Klosgen	50	Conviction	100	Gini Index	148
Kappa	50	Added Value	100	Novelty	148
Odds Ratio	50	Laplace	99	Certainty Factor	148
Conviction	50	Confiança	99	Kappa	147
Added Value	50	Gini Index	99	Collective Strength	147
Laplace	49	Novelty	99	Klosgen	147
Gini Index	49	Collective Strength	99	Added Value	147
Collective Strength	49	Klosgen	99	Confiança	146
Lift	49	Lift	99	Lift	146
Certainty Factor	49	Certainty Factor	99	Conviction	143
Lambda	45	Lambda	94	Lambda	142
LDA + *bag-of-words*	23	LDA + *bag-of-related-words*	50	LDA + *bag-of-words*	63
LDA + *bag-of-related-words*	22	LDA + *bag-of-words*	48	LDA + *bag-of-related-words*	51
Mutual Information LHS	12	J-Measure	19	J-Measure	31
J-Measure	7	Mutual Information LHS	16	Mutual Information LHS	28
IS	4	IS	15	IS	16

medida *Odds Ratio*, que também tem semântica semelhante, e apresentou bons resultados, estando entre as 3 melhores em 5 dos 6 casos. De acordo com [19], ela também compartilha muitas propriedades com a medida ϕ-*Coefficient*.

Observando apenas os resultados obtidos pelo modelo LDA, não é possível apontar uma diferença significativa entre os resultados obtidos. O uso de termos compostos nesse caso não trouxe benefícios diretos para o modelo, sendo que o uso de termos simples foi ligeiramente melhor em alguns casos.

Avaliação da coerência observada: Na Tabela IV são apresentados os valores de Coerência Observada para o melhor tópico ou grupo e o menor valor da medida para o tópico ou grupo do quartil superior para as coleções Re8 e ACM. Nessa avaliação não foi possível apontar uma medida como sendo a mais estável para todos os casos. Destaca-se aqui o bom desempenho da medida *Collective Strength* para a coleção Re8. Os autores de [19] apontam que essa medida apresenta propriedades similares a medida ϕ-Coefficient, que apresentou bons resultados na avaliação de termo intruso. A medida *Odds Ratio* também apresentou bons resultados nessa base para os valores de k igual a 50 e 100, mas não teve um comportamento estável em todos os casos. Ainda, para o valor de coerência observada em 25%, o modelo LDA foi ligeiramente superior aos resultados obtidos pelas melhores medidas objetivas com a abordagem proposta, mas não são suficientes para compensar a diferença de resultados para o melhor tópico avaliado.

Para a coleção ACM, a abordagem proposta foi superior ao modelo LDA em um dos casos. Para a abordagem proposta, as medidas *Confiança*, *Laplace* e *IS* se destacam. As duas primeiras medidas visam avaliar a cobertura de uma regra de associação em relação ao conjunto de transações condicionalmente a cobertura do antecedente da regra de associação. A medida *Laplace* é uma variação da medida *Confiança* na qual aplica-se uma correção visando penalizar regras de associação muito específicas, ou seja, que cobrem poucas transações. A medida *IS* pode ser interpretada como a medida cosseno, muito utilizada para determinar a similaridade entre documentos representados por modelos VSM, e é equivalente a média geométrica da *Confiança* das regras geradas a partir de um par de itens [19]. De uma forma geral, a abordagem proposta tem melhores resultados para medidas que avaliam a cobertura das regras de associação em relação as transações obtidas. Diferentemente dos casos anteriores, em que as medidas de destaque são aquelas que avaliam o grau de associação entre antecedente e consequente. Uma possível causa para esse efeito é o tamanho médio dos documentos dessa coleção, bem como a diferença do estilo de escrita em relação aos textos de notícia da coleção Re8. Por fim, considerando o valor de coerência observada em 25%, o modelo LDA também obteve melhores resultados nesse caso, enquanto o desempenho da abordagem proposta foi significativamente inferior em muitos dos casos. Destaca-se o bom resultado da medida *IS* para os valores de k 100 e 150, que se aproximaram do desempenho do modelo LDA considerando tanto o valor de coerência observada em 25% quanto no melhor valor obtido.

Também na avaliação de coerência observada, considerando apenas os resultados obtidos pelo modelo LDA, não é possível apontar uma diferença significativa entre os resultados obtidos. O uso de termos compostos nesse caso também não trouxe benefícios diretos para o modelo, sendo que o uso de termos simples obteve resultados ligeiramente melhores em alguns casos. Esse resultado fornece alguns indícios de que, para que as técnicas tenham ganhos significativos com o uso de termos compostos, ela deve utilizar de uma forma explícita a informação de que existe dependência entre os termos durante a identificação dos assuntos da coleção.

Um exemplo com a lista de termos selecionados para os melhores casos da abordagem proposta e pelo modelo LDA considerando a avaliação da coerência observada é apresentada na Tabela V. Para a coleção Re8, considerando a abordagem proposta, a medida objetiva com melhores resultados foi a *Collective Strength*, que tem como propriedade selecionar regras de associação que possuam forte grau de associação entre o antecedente e o consequente. Uma inspeção visual dos grupos com melhor valor de coerência observada indica que esse comportamente se reflete nos termos compostos selecionados, destacando o termo "*bank*" que apareceu em todos os termos compostos selecionados. Ainda para a coleção Re8, o modelo LDA combinado com *bag-of-words* e *bag-of-related-words* foi melhor avaliado nos tópicos relacionados com "petróleo". Para a coleção ACM, a abordagem proposta foi melhor avaliada para a medida objetiva *Laplace*. Essa medida tem como propriedade selecionar regras de associação que possuam uma maior cobertura das transações. Ainda sim, pode-se observar um domínio do termo "*model*" para o valor de k 50 e "*cell*" para os outros casos. Apesar disso, em todos os valores de k, um dos termos compostos selecionados não possui esses termos, possivelmente por influência da seleção realizada com a medida objetiva. Para o modelo LDA combinado com *bag-of-words* para todos os valores de k e *bag-of-related-words* para o valor de k 50, houve um domínio de tópicos com termos pouco significativos. Possivelmente, os termos selecionados são resultado de problemas no pré-processamento, que separou alguns termos importantes na coleção ACM. Esses

Tabela IV. RESULTADOS PARA AS COLEÇÕES RE8 E ACM DO VALOR DE COERÊNCIA OBSERVADA (CO) DO TÓPICO COM MELHOR AVALIAÇÃO E DO TÓPICO COM MENOR VALOR DE MEDIDA ENTRE OS 25% MELHORES TÓPICOS. OS RESULTADOS ESTÃO ORDENADOS PELO MELHOR VALOR DE CO OBTIDO.

Coleção Re8

K=50			K=100			K=150		
Configuração	Melhor CO	CO em 25%	Configuração	Melhor CO	CO em 25%	Configuração	Melhor CO	CO em 25%
Collective Strength	1,57	0,56	Collective Strength	1,68	0,58	Collective Strength	1,86	0,59
Odds Ratio	1,50	0,37	Odds Ratio	1,50	0,50	Klosgen	1,70	0,58
Novelty	1,28	0,57	Lift	1,49	0,60	Certainty Factor	1,70	0,59
Gini Index	1,28	0,54	LDA + *bag-of-words*	1,39	0,69	Lift	1,58	0,62
Kappa	1,27	0,56	Certainty Factor	1,27	0,59	Kappa	1,54	0,6
Certainty Factor	1,27	0,55	ϕ-Coefficient	1,24	0,48	Gini Index	1,54	0,59
Klosgen	1,22	0,55	Added Value	1,24	0,57	Novelty	1,54	0,61
Lift	1,18	0,52	Novelty	1,22	0,61	Mutual Information LHS	1,52	0,63
Added Value	1,09	0,55	Gini Index	1,22	0,60	Odds Ratio	1,50	0,55
Conviction	1,05	0,57	LDA + *bag-of-related-words*	1,21	0,67	J-Measure	1,42	0,62
LDA + *bag-of-related-words*	0,94	0,67	IS	1,16	0,62	Confiança	1,40	0,45
ϕ-Coefficient	0,93	0,36	Klosgen	1,11	0,56	Laplace	1,39	0,52
LDA + *bag-of-words*	0,92	0,69	Lambda	1,07	0,49	LDA + *bag-of-words*	1,38	0,71
Laplace	0,89	0,31	Laplace	1,05	0,50	Added Value	1,38	0,59
Confiança	0,89	0,40	Kappa	1,05	0,60	IS	1,28	0,61
IS	0,83	0,61	Confiança	0,95	0,44	ϕ-Coefficient	1,24	0,48
Lambda	0,83	0,41	J-Measure	0,95	0,62	Lambda	1,07	0,49
Mutual Information LHS	0,76	0,60	Mutual Information LHS	0,90	0,63	LDA + *bag-of-related-words*	1,03	0,70
J-Measure	0,76	0,60	Conviction	0,86	0,60	Conviction	1,02	0,59

Coleção ACM

K=50			K=100			K=150		
Configuração	Melhor CO	CO em 25%	Configuração	Melhor CO	CO em 25%	Configuração	Melhor CO	CO em 25%
LDA + *bag-of-words*	1,54	0,80	Laplace	1,40	0,17	LDA + *bag-of-words*	1,52	0,81
Laplace	1,37	0,13	IS	1,34	0,63	Laplace	1,40	0,18
Confiança	1,30	0,18	LDA + *bag-of-words*	1,30	0,81	LDA + *bag-of-related-words*	1,39	0,77
Gini Index	1,06	0,32	Confiança	1,30	0,19	IS	1,34	0,63
Novelty	1,02	0,33	LDA + *bag-of-related-words*	1,21	0,77	Confiança	1,30	0,19
Mutual Information LHS	0,95	0,64	Odds Ratio	1,19	0,23	Odds Ratio	1,19	0,24
LDA + *bag-of-related-words*	0,89	0,77	Kappa	1,16	0,25	J-Measure	1,18	0,65
ϕ-Coefficient	0,89	0,24	J-Measure	1,14	0,68	Kappa	1,16	0,25
Klosgen	0,87	0,14	Gini Index	1,06	0,35	Gini Index	1,06	0,31
J-Measure	0,86	0,62	Novelty	1,02	0,29	Novelty	1,04	0,28
IS	0,84	0,58	Collective Strength	0,98	0,37	Collective Strength	0,98	0,31
Kappa	0,82	0,24	Klosgen	0,93	0,18	ϕ-Coefficient	0,95	0,24
Collective Strength	0,79	0,28	ϕ-Coefficient	0,89	0,26	Conviction	0,93	0,18
Odds Ratio	0,63	0,17	Mutual Information LHS	0,87	0,71	Klosgen	0,93	0,17
Conviction	0,63	0,10	Lift	0,79	0,30	Mutual Information LHS	0,87	0,71
Lift	0,55	0,28	Conviction	0,72	0,16	Lift	0,79	0,28
Added Value	0,55	0,10	Lambda	0,56	0,12	Added Value	0,67	0,17
Certainty Factor	0,49	0,15	Added Value	0,55	0,13	Lambda	0,56	0,13
Lambda	0,33	0,10	Certainty Factor	0,54	0,14	Certainty Factor	0,54	0,18

problemas são inerentes dos processos automáticos de extração de termos, uma vez que a conversão dos textos em um formato mais adequado para as ferramentas pode gerar erros que são acumulados no processo. Apesar disso, pode-se observar que os tópicos com menor valor entre os 25% melhores tópicos são significativos, e apresentam termos completos e tópicos visualmente coerentes. O modelo LDA combinado com *bag-of-related-words* foi capaz de tratar melhor esses erros para os tópicos avaliados de k 50 e 100. Ainda, modelo LDA com *bag-of-related-words* selecionou poucos termos compostos.

V. CONCLUSÃO

Neste trabalho foi proposta uma abordagem não-supervisionada para identificação de assuntos em coleções de documentos que combinam técnicas de regras de associação e de agrupamento de dados, explorando explicitamente a dependência entre os termos para extrair termos compostos, melhorando a interpretabilidade dos assuntos identificados. Os assuntos são obtidos ao combinar o contexto local e o contexto geral da relação entre os termos. A abordagem proposta foi comparada com o modelo LDA tradicional e o modelo LDA utilizando uma representação que inclui termos compostos (*bag-of-related-words*), ambas estado da arte na área. Os experimentos indicam que a abordagem proposta produz uma lista de descritores para cada grupo significativamente melhor do que a produzida aplicando o modelo LDA.

Considerando as diferentes características entre as coleções avaliadas, o uso de termos compostos com a abordagem proposta apresentou bons resultados na coleção Re8, formada por textos de notícias, para a avaliação de termo intruso e de coerência observada. Na coleção ACM, formada por artigos científicos da área de computação, a abordagem proposta apresentou um resultado superior ao LDA na avaliação do termo intruso, mas para a avaliação de coerência observada não foi possível identificar uma diferença significativa entre os modelos comparados. Considerando apenas os assuntos identificados pelo modelo LDA, a avaliação não apontou um ganho no uso de termos compostos, dando indícios de que os benefícios do uso de termos compostos será maior se o modelo considerar essa informação explicitamente durante a identificação dos assuntos da coleção.

Ainda, o modelo proposto processa cada documento de forma independente, tornando a abordagem interessante para aplicação em cenários incrementais. Como trabalhos futuros, pretende-se aproveitar a abordagem proposta para construir uma representação com dimensionalidade reduzida da coleção. Para isso, cada descrição obtida será tratada como uma dimensão latente, e diferentes formas de mapeamento dos documentos da coleção nessas novas dimensões serão avaliadas.

AGRADECIMENTOS

Os autores agradecem a CAPES (processo DS-6345378/D) e a FAPESP (processo número 2014/08996-0) pelo apoio financeiro concedido.

REFERÊNCIAS

[1] B. Liu, *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications)*, 2nd ed. Springer, 2011.

[2] Y. Hu, J. Boyd-Graber, B. Satinoff, and A. Smith, "Interactive topic modeling," *Machine Learning*, vol. 95, no. 3, pp. 423–469, 2014.

[3] X. Cheng, D. Miao, C. Wang, and L. Cao, "Coupled term-term relation analysis for document clustering," in *IJCNN*. IEEE, 2013, pp. 1–8.

[4] F. Figueiredo, L. Rocha, T. Couto, T. Salles, M. A. Gonçalves, and W. Meira Jr., "Word co-occurrence features for text classification," *Inf. Syst.*, vol. 36, no. 5, pp. 843–858, Jul. 2011.

[5] Y. Gao, Y. Xu, Y. Li, and B. Liu, "A two-stage approach for generating topic models," in *PAKDD*, vol. 7819. Springer, 2013, pp. 221–232.

[6] B. Pôssas, N. Ziviani, W. Meira, Jr., and B. Ribeiro-Neto, "Set-based model: A new approach for information retrieval," in *SIGIR*. New York, NY, USA: ACM, 2002, pp. 230–237.

Tabela V. LISTA COM OS 10 TERMOS SELECIONADOS DA MELHOR MEDIDA OBJETIVA SELECIONADA PARA A ABORDAGEM PROPOSTA E PELO LDA DO TÓPICO COM MELHOR AVALIAÇÃO E DO TÓPICO COM MENOR VALOR DE MEDIDA ENTRE OS 25% MELHORES TÓPICOS.

Coleção Re8 - k=50			
Método proposto + *Collective Strenght*	Melhor CO	1,57	commerce_bank pct_bank money_bank prime_bank bank_money bank_rate company_bank rate_bank bank_holdings save_bank
	CO em 25%	0,56	mead westinghouse echolab shareholders strong growers manufacturing pennwalt valley waste
LDA + *bag-of-words*	Melhor CO	0,92	energy price dlrs ecuador pipeline crude petroleum barrel gas oil
	CO em 25%	0,69	minister government plan told states spokesman official reuter meeting talks
LDA + *bag-of-related-words*	Melhor CO	0,94	products price opec mln dlrs bpd crude saudi barrel oil
	CO em 25%	0,67	congress bill trade reagan bill_trade house legislation year senate foreign
Coleção Re8 - k=100			
Método proposto + *Collective Strenght*	Melhor CO	1,68	commerce_bank offer_bank pct_bank prime_bank bank_rate bank_oper company_bank rate_bank bank_holdings save_bank
	CO em 25%	0,58	opec_saudi saudi_opec saudi_economy opec_arabia arabia_opec saudi day field_august tampico economy_saudi
LDA + *bag-of-words*	Melhor CO	1,39	refinery texaco shell oper pipeline bpd crude refining petroleum oil
	CO em 25%	0,69	petrobras guaranteed accept card branch bank chase manhattan credit bankamerica
LDA + *bag-of-related-words*	Melhor CO	1,21	gcc mln petrobras crowns arabia saudi nazer riyal oil arabia_saudi
	CO em 25%	0,67	law commission securities market crazy sec eddie company entertainment crazy_eddie
Coleção Re8 - k=150			
Método proposto + *Collective Strenght*	Melhor CO	1,86	commerce_bank offer_bank texas_bank bank_oper bank_rate company_bank rate_bank sumitomo_bank bank_holdings bank_sumitomo
	CO em 25%	0,59	semiconductor analysts european future rumours slate entertainment won market_dollar money
LDA + *bag-of-words*	Melhor CO	1,38	refinery mln stock pipeline bpd crude refining petroleum barrel oil
	CO em 25%	0,71	filing company rules suit federal bankruptcy appeal court settlement seeking
LDA + *bag-of-related-words*	Melhor CO	1,03	deficit_trade budget exports deficit trade yeutter concern volcker year countries
	CO em 25%	0,70	increase record share stock april declared dividend split_stock split date
Coleção ACM - k=50			
Método proposto + *Laplace*	Melhor CO	1,37	model_multiple event_stream model_object model_problem model_level model_existing model_simple event_model model_account model_proposed
	CO em 25%	0,13	transactions_detection transactions_dastm transactions_serializability transactions_programmer transactions_tmobject transactions_atomic transactions_hardware transactions_dstm transactions_stm transactions_techniques
LDA + *bag-of-words*	Melhor CO	1,54	con erent cient nite rst speci nition signi cation ned
	CO em 25%	0,80	sensor system node mote data deployment tiny sense network application
LDA + *bag-of-related-words*	Melhor CO	0,89	decoding input general set random computer problem code polynomial function
	CO em 25%	0,77	information phone system location mobile context data device user application
Coleção ACM - k=100			
Método proposto + *Laplace*	Melhor CO	1,40	cell_maximum type_referenced cell_current type_cell cell_space cell_list cell_finally cell_complexity cell_detection cell_called
	CO em 25%	0,17	topology_vector allocation_top region_vector party_vector haplotype_vector transactions_vector vector simd_vector score_atypical texture_field
LDA + *bag-of-words*	Melhor CO	1,30	con erent haplotype cient rst speci nition signi cation ned
	CO em 25%	0,81	sensor system node mote data deployment tiny sense network application
LDA + *bag-of-related-words*	Melhor CO	1,21	compression bit decoding encoding scheme error quantization data code wavelet
	CO em 25%	0,77	information security anonymity privacy data attacks protection user mix adversary
Coleção ACM - k=150			
Método proposto + *Laplace*	Melhor CO	1,40	cell_maximum type_referenced cell_current type_cell cell_space cell_list cell_finally cell_complexity cell_detection cell_called
	CO em 25%	0,18	agent_information communication_visualization adaptive_information resource_information tag_information cache_leakage domain_acquire robot_crawler database_protection media
LDA + *bag-of-words*	Melhor CO	1,52	con erent cient rst speci ect nition signi cation ned
	CO em 25%	0,81	process center dataset partition cluster means computer dryad iteration algorithm
LDA + *bag-of-related-words*	Melhor CO	1,39	con erent cient rst speci nition signi cation ned algorithm
	CO em 25%	0,77	point shortest connected vehicle table intersection path network routing path_shortest

[7] R. G. Rossi and S. O. Rezende, "Building a topic hierarchy using the bag-of-related-words representation," in *DocEng*. New York, NY, USA: ACM, 2011, pp. 195–204.

[8] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *J. Am. Soc. Info. Sci.*, vol. 41, no. 6, pp. 391–407, 1990.

[9] T. Hofmann, "Probabilistic latent semantic indexing," in *SIGIR*. New York, NY, USA: ACM, 1999, pp. 50–57.

[10] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[11] H. M. Wallach, "Topic modeling: Beyond bag-of-words," in *ICML*. New York, NY, USA: ACM, 2006, pp. 977–984.

[12] H. D. Kim, D. H. Park, Y. Lu, and C. Zhai, "Enriching text representation with frequent pattern mining for probabilistic topic modeling," *Proc. Am. Soc. Info. Sci. Tech.*, vol. 49, no. 1, pp. 1–10, 2012.

[13] D. Zhu, Y. Fukazawa, E. Karapetsas, and J. Ota, "Intuitive topic discovery by incorporating word-pair's connection into lda," in *Web Intelligence*. IEEE, 2012, pp. 303–310.

[14] J. H. Lau, T. Baldwin, and D. Newman, "On collocations and topic models," *ACM Trans. Speech Lang. Process.*, vol. 10, no. 3, pp. 10:1–10:14, Jul. 2013.

[15] B. M. Nogueira, M. F. Moura, M. S. Conrado, R. G. Rossi, R. M. Marcacini, and S. O. Rezende, "Winning some of the document preprocessing challenges in a text mining process," in *SBBD*, Porto Alegre : SBC. Porto Alegre : SBC, 2008, p. 10–18.

[16] R. M. Marcacini, G. N. Correa, and S. O. Rezende, "An active learning approach to frequent itemset-based text clustering," in *ICPR*. IEEE, 2012, pp. 3529–3532.

[17] P. D. Turney and P. Pantel, "From frequency to meaning: vector space models of semantics," *J. Artif. Int. Res.*, vol. 37, no. 1, pp. 141–188, 2010.

[18] J. Chang, J. L. Boyd-Graber, S. Gerrish, C. Wang, and D. M. Blei, "Reading tea leaves: How humans interpret topic models," in *NIPS*. Curran Associates, Inc., 2009, pp. 288–296.

[19] P.-N. Tan, V. Kumar, and J. Srivastava, "Selecting the right objective measure for association analysis," *Inf. Syst.*, vol. 29, no. 4, pp. 293–313, 2004.

[20] M. Steinbach, P. Tan, H. Xiong, and V. Kumar, "Objective Measures for Association Pattern Analysis," *Contemporary Mathematics*, no. 443, pp. 205 – 226, 2007.

[21] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, "Automatic evaluation of topic coherence," in *HLT*. Stroudsburg, PA, USA: ACL, 2010, pp. 100–108.