

Agrupamento hierárquico como instrumento de apoio ao usuário no Mapeamento Sistemático

João Antunes¹, Danilo M. Eler², Solange O. Rezende¹

¹Instituto de Ciências Matemáticas e de Computação (ICMC)
Universidade de São Paulo (USP)
Caixa Postal 668 – 13.560-970 – São Carlos – SP – Brasil

²Departamento de Matemática e Computação (DMC)
Universidade Estadual Paulista “Júlio de Mesquita Filho” (UNESP)
19.060-900 – Presidente Prudente – SP – Brasil

joao4ntunes@gmail.com, daniloeler@fct.unesp.br, solange@icmc.usp.br

Abstract. *Bibliographical reviews are essential for the development of all scientific work, since it presents an indicative of the area to be researched and of its viability, it avoids reworking and favors the evolution of research approached topics. However, there are limitations in the method of performing these revisions, either by the enormous amount of research time, by the selection of non-repeated works, and by the subjectivity in categorizing works as accepted. To alleviate these problems, which require a great deal of mental effort on the part of the researcher, a methodology is proposed in this work, with the objective of visually supporting the identification of related works, using hierarchical grouping.*

Resumo. *Revisões bibliográficas são essenciais para o desenvolvimento de todo trabalho científico, pois apresentam um indicativo da área a ser pesquisada e de sua viabilidade, evita retrabalhos e favorece a evolução dos tópicos de pesquisa abordados. Entretanto, existem limitações no método de realização dessas revisões, seja pelo enorme tempo de pesquisa, da seleção de trabalhos não repetidos e da subjetividade em se categorizar trabalhos como aceitos. Para amenizar esses problemas, que demandam alto esforço mental por parte do pesquisador, uma metodologia é proposta neste trabalho, com o objetivo de apoiar visualmente na identificação de trabalhos relacionados, utilizando agrupamento hierárquico.*

1. Introdução

A organização de informações digitais é uma tarefa custosa e não trivial, que demanda muito tempo e recursos computacionais, como memória e processamento. Entretanto, essa organização é fundamental para adquirir-se conhecimento, pois favorece a absorção e compreensão de relacionamentos e referências entre objetos, principalmente quando se trata de um volume grande de dados. Em dados textuais, essa necessidade se torna mais evidente devido à dificuldade de se reconhecer automaticamente os assuntos tratados nos documentos, bem como o contexto em que cada documento está inserido, devido à não estruturação desse tipo de dado. Nesse sentido, técnicas de Mineração de Textos são utilizadas para ajudar na extração de padrões ocultos entre os textos.

Para Rezende (2003), a Mineração de Textos é uma subárea da Descoberta de Conhecimento em Bases de Dados (do inglês, *Knowledge Discovery in Databases* (KDD)), definida por Fayyad (1996) como um processo para prover significado em grandes volumes de dados, permitindo que um analista extraia informações ocultas e difíceis de serem obtidas apenas por uma análise manual [Fayyad et al. 1996]. Na Mineração de Textos, o foco está direcionado na busca por padrões textuais, que podem ser empregados em tarefas de agrupamento de documentos, classificação, sumarização e recuperação de informações [Rezende 2003].

Por ser uma linguagem natural às pessoas, dados textuais são abundantes em qualquer domínio de aplicação, sendo uma fonte importante de conhecimento em potencial. No meio científico, a análise de publicações pode ser utilizada para descoberta de trabalhos relacionados e tendências de pesquisa, que favorecem o embasamento teórico de projetos, a padronização de conceitos, e a possibilidade de comparações para validação de novos métodos. Neste trabalho, foram utilizadas técnicas de Mineração de Textos para agrupamento de publicações similares, considerando a hipótese de que é possível apoiar a identificação de trabalhos relacionados por meio de agrupamentos hierárquicos e utilizando múltiplas visualizações coordenadas. A justificativa é que a análise não-supervisionada dos dados contribui para um acompanhamento regular e periódico de novas publicações, inviabilizado pelas dificuldades na realização de uma revisão bibliográfica manual, que pode ser cansativa e demorar meses para ser concluída.

Para validação da hipótese proposta, foi realizado um estudo de caso com os resultados obtidos de dois mapeamentos sistemáticos distintos na área de Mineração de Textos, ambos realizados de forma manual por meio de consultas às bases *ACM Digital Library*, *IEEE Xplore*, *Science Direct*, *Scopus* e *Web of Science*. O primeiro trata-se de um levantamento de trabalhos relacionados à “Análise de Sentimentos Baseada em Aspectos”, publicados entre 2005 e 2015. Após a fase de Seleção (remoção de (i) trabalhos duplicados, (ii) com apenas uma página (posters, apresentações, resumos e editoriais), (iii) indisponíveis ou não-acessíveis, (iv) escritos em idiomas diferentes do inglês e português (todos os resumos e títulos estão escritos no idioma inglês) e (v) não relacionados ao tema), 441 trabalhos foram aceitos do total de 599. No segundo mapeamento foram investigadas as publicações relacionadas à “semântica na Mineração de Textos”, publicados até fevereiro de 2016 [Sinoara et al. 2017]. Nesse trabalho, foram encontrados 3984 trabalhos relacionados ao tema, dos quais 1693 foram aceitos na fase de Seleção.

O objetivo deste trabalho é agrupar hierarquicamente os artigos aceitos na fase de Seleção de ambos os mapeamentos descritos, considerando seus títulos e resumos como fonte de dados e a classificação na fase de Extração (responsável por verificar se um trabalho está relacionado ao tema buscado) para comparação dos resultados entre as abordagens manuais e automáticas. Para apoiar essa tarefa, foi utilizado o protótipo da ferramenta Pinda [Antunes and Eler 2017], devido à presença de múltiplas visualizações coordenadas que contribuem para uma exploração visualmente mais rica do mesmo conjunto de dados [North and Shneiderman 2000, Boukhelifa et al. 2003].

2. Trabalhos relacionados

A Mineração de Textos têm importância considerável em tarefas de agrupamento de documentos, pois provê meios de se obter uma análise não-linear e possivelmente mais

objetiva de dados não-estruturados. Sua aplicação em Mapeamentos Sistemáticos possibilita a análise de documentos quanto às questões de evolução de tópicos, tendências e relacionamentos entre áreas, por exemplo. Ferramentas como o StArt (do inglês, *State of the Art through Systematic Review*) podem ser aplicadas para auxiliar no processo manual de Revisão Sistemática [Kitchenham 2004], permitindo seguir um protocolo de pesquisa com critérios bem definidos de Seleção e Extração de trabalhos relacionados à um determinado tópico de pesquisa [Zamboni et al. 2010]. Apesar de efetivo, esse processo manual é demorado e específico para cada tópico, inviabilizando um acompanhamento temporal contínuo dos trabalhos publicados, que pode ser empregado para identificação de tendências temáticas e relacionamentos entre diferentes áreas da ciência [Boyack et al. 2005, Leydesdorff et al. 2013, van Eck and Waltman 2014].

Em [Silva et al. 2016], é proposto um método para mapeamento hierárquico de publicações científicas em diferentes áreas de pesquisa, utilizando uma visualização tridimensional de rede para apoiar o Pós-processamento da Mineração de Textos, caracterizado pela análise manual dos resultados gerados, neste caso, pelo agrupamento automático dos documentos. Embora apresente recursos interativos para manipulação dos resultados, a distinção dos grupos no ambiente da ferramenta é baseada apenas em características de cor, tamanho e distância relativa dos documentos. Acreditamos que isso seja uma lacuna a ser preenchida, já que diferentes visualizações podem agregar diferentes pontos de vista a respeito de uma base de dados, principalmente quando há coordenação entre as visualizações [North and Shneiderman 2000].

Visualizações de rede permitem que se obtenha uma visão geral dos resultados, mas conforme o volume de dados aumenta os relacionamentos entre documentos pode se tornar difícil de ser compreendido, devido às singularidades e detalhes que podem ser ocultados pela densidade de informações visuais. Além disso, o uso de múltiplas visualizações pode apresentar maior acessibilidade ao usuário, por explorar diferentes formas, cores e estruturas visuais para representar o mesmo conjunto de dados. Para realizar esse estudo, o protótipo da ferramenta Pinda foi adaptado para poder realizar agrupamentos de bases de dados locais, já que sua primeira versão era específica para agrupamento de dados advindos da *Web* por meio de APIs de busca [Antunes and Eler 2017].

3. Trabalho proposto

A principal contribuição que este trabalho procura trazer é validar sua hipótese evidenciando os benefícios do agrupamento hierárquico não-supervisionado em pesquisas científicas, em especial na fase de Revisão Bibliográfica, quando é feito um levantamento do estado da arte de determinado tema. Em relação à hierarquia dos agrupamentos obtidos, espera-se que o apoio visual de múltiplas visualizações coordenadas providencie um entendimento menos subjetivo dos relacionamentos entre os grupos de documentos e apoie o processo de validação dos resultados.

Conforme mencionado na Seção 1, os dados utilizados para este estudo de caso foram obtidos por meio de acesso *online* aos repositórios de 5 bases de dados (*ACM Digital Library*, *IEEE Xplore*, *Science Direct*, *Scopus* e *Web of Science*) de resumos de publicações científicas sobre os tema “Análise de Sentimentos Baseada em Aspectos” e “semântica na Mineração de Textos”. Para realização das pesquisas, foram consideradas respectivamente as expressões ((*feature** OR *aspect**

OR entity OR entities OR target*) AND (based opinion mining) OR (based sentiment analysis)) OR (''feature extract'' OR ''feature extraction'' OR ''aspect extract'', OR ''aspect extraction'')) AND (''opinion mining'' OR ''sentiment analysis'' OR ''sentiment detection'' OR ''sentiment orientation'' OR ''sentiment extract*'' OR ''opinion analysis'') (441 trabalhos aceitos, dos quais 157 foram selecionados) e semantic* AND text* AND (mining OR representation OR clustering OR classification OR association rules) (1693 trabalhos aceitos, dos quais 801 foram selecionados), aplicadas nos títulos e palavras-chave sempre que possível.

Para a fase de Seleção, os dados sobre as publicações foram exportados para o formato BibTex (modelo de referência bibliográfica utilizado pelo LaTeX), contendo o resumo embutido em cada referência. Todos os resumos e títulos dos trabalhos aceitos estavam escritos no idioma inglês, com predominância de linguagem em forma culta por se tratar de publicações científicas. Além disso, outra característica importante é o tamanho conciso desse conjunto de dados, dado o fato de serem resumos de publicações, que variam de 50 a 250 palavras, aproximadamente.

Para a avaliação dos grupos obtidos foi utilizado o Coeficiente de Silhueta, um método proposto por Rousseeuw (1986) para avaliar a consistência interna de objetos agrupados, medindo a similaridade entre eles. Supondo três grupos distintos de objetos, A , B e C , resultantes de algum algoritmo de agrupamento, deseja-se calcular uma medida $s(i)$ para cada objeto de forma a permitir um comparativo da similaridade dos grupos. Para isso, dado que todos os grupos contém mais de um objeto, $a(i)$ é definida como a dissimilaridade (distância) média do objeto i para todos os outros objetos do mesmo grupo A . Dessa forma, $d(i, C)$ é definido como a dissimilaridade média de i para todos os objetos do grupo C , sendo $b(i) = \min_{C \neq A} d(i, C)$ o menor de todos esses valores, tal que $b(i) = d(i, B)$ (denominado vizinho do objeto i) seja a dissimilaridade média de i para o cluster B . Essas informações permitem calcular a silhueta por meio da fórmula $s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$, $-1 \leq s(i) \leq 1$. Quanto mais próximo de -1, pior a similaridade do objeto i em relação ao grupo atual a que pertence; analogamente, quanto mais próximo de 1, melhor a similaridade do objeto a seu grupo atual [Rousseeuw 1987].

Nesse trabalho é proposta a utilização de agrupamento hierárquico não-supervisionado para auxiliar no processo de Revisão Bibliográfica, utilizando o protótipo da ferramenta Pinda para visualização coordenada dos resultados. Para a análise estatística dos textos, foi utilizada a medida TF-IDF (*Term Frequency - Inverse Document Frequency*), por considerar a relevância dos termos em relação aos documentos em que ocorre [Sparck Jones 1988]. Para a representação dos dados foi utilizado o modelo *Bag-Of-Words*, por ser uma das técnicas mais utilizadas em Mineração de Textos e pelo escopo deste trabalho não considerar a semântica contida entre os termos. Os agrupamentos foram efetuados utilizando o algoritmo de particionamento *K-means*, por ser um algoritmo consagrado, apresentar bons resultados a medida que o volume de dados aumenta, e necessitar de apenas um parâmetro (quantidade k de grupos) [Abbas 2008], permitindo que a hierarquia de grupos fosse construída com sucessivas aplicações do mesmo algoritmo para cada nível. Para Projeção Multidimensional de Dados (projeção de um conjunto de dados

de alta dimensão (número de atributos ou termos) para outro de baixa dimensão (nesse caso, bidimensional)) foi utilizada a técnica de Redução de Dimensionalidade *Principal Component Analysis* (PCA), por apresentar bons resultados em aplicações textuais devido à característica de considerar a maior variabilidade dos dados, tendendo à identificar os padrões mais relevantes do conjunto [Paulovich 2008], e medida comum para o cálculo da similaridade entre os documentos, devido à sua aplicabilidade em representações vetoriais de textos [Tan et al. 2005].

4. Pré-processamento

Com o propósito de preparar os dados para a fase de Representação, foi realizada uma série de tratamentos para adequação dos textos, sintetizados na Figura 1. Por ser um requisito da ferramenta Pinda, houve a necessidade de conversão das referências BibTex para uma estrutura padronizada do tipo JSON (*JavaScript Object Notation*), contendo o título, a URL (utilizado apenas para informação), o resumo e a categoria (se foi aceita ou não na fase de Extração) de cada publicação.

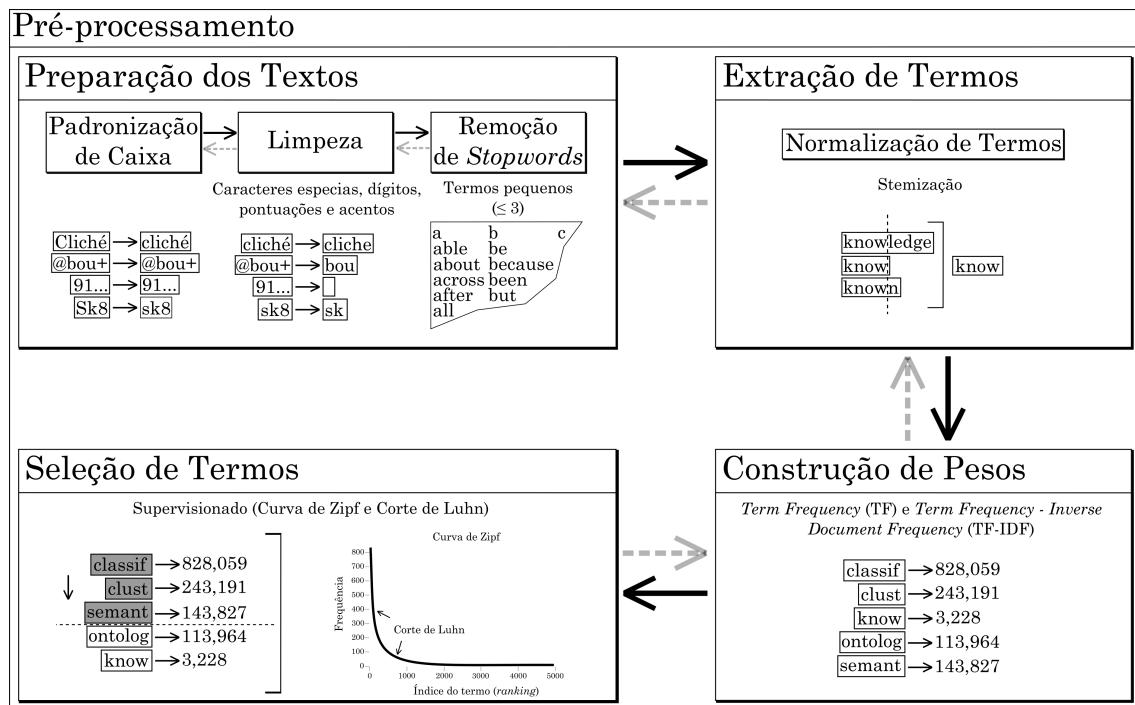


Figura 1. Tratamentos textuais realizados na fase de Pré-processamento.

Na etapa de Padronização de Caixa, foi padronizada a caixa dos títulos e resumos, convertendo todos os caracteres textuais para letra minúscula. Isso foi necessário para facilitar a fase de Remoção de Stopwords, que compara cada termo com uma lista de palavras consideradas irrelevantes ao escopo do problema, como preposições e artigos, que não agregam muita informação em análises sintáticas. Entre essas duas fases, foi realizada uma Limpeza dos dados, substituindo caracteres especiais (&, \$, % etc.) e dígitos por espaços em branco, e caracteres acentuados por suas correspondentes letras sem acento. Similar à Padronização de Caixa, a Limpeza é importante para potencializar o processo de identificação de stopwords, já que um mesmo termo pode ter inúmeras variações de escrita.

Para reduzir o conjunto de termos relevantes, foi aplicado o Algoritmo de Porter (do idioma inglês) para stemização de termos, que transforma cada termo em seu respetivo radical [Porter 1997]. Como ilustrado na Figura 1, na fase de Extração de Termos e etapa de Normalização, os termos *knowledge*, *know* e *known* foram convertidos para o mesmo radical *know*, o que favorece na identificação de termos análogos e intensifica o processo de agrupamento. Assim, técnicas de Combinação de Termos não foram empregadas, devido à subjetividade nas etapas de Enriquecimento, uso de Expressões do Domínio e N-gramas.

Sabe-se que, dependendo do escopo, um mesmo termo pode ter importâncias diferentes e até contrárias. Em um documento sobre Mineração de Textos, por exemplo, o termo “conhecimento” pode ocorrer muito mais vezes do que em um documento sobre Engenharia de Software, em que o termo “requisito” pode ocorrer em maior quantidade. Baseado nesse fato, o uso da frequência total de um termo em um conjunto de documentos, denominado *Term Frequency* (TF), pode ocultar a relevância relativa do termo considerando todos os documentos da base de dados. Por esse motivo, nesse trabalho foi utilizada a medida *Term Frequency - Inverse Document Frequency* (TF-IDF), que calcula o peso dos termos considerando a ocorrência dos mesmos em todos os documentos analisados. Assim, um termo muito frequente em poucos documentos pode ter menor valor do que um termo com baixa frequência mas presente em muitos documentos [Sparck Jones 1988].

Tabela 1. Valores das métricas TF e TF-IDF sem e com stemização.

		TF		TF-IDF	
		Sem stemming	Com Stemming	Sem stemming	Com Stemming
Base de Dados 1	#termos	4132	2496	4132	2496
	mín	1	1	2,644	2,644
	méd	9,179	15,195	11,173	15,069
	máx	1078	1192	203,109	226,095
	Var	1305,265	3396,247	291,669	523,807
	DP	36,128	58,277	17,078	22,886
Base de Dados 2	#termos	8342	4953	8342	4953
	mín	1	1	3,229	3,229
	méd	16,569	27,906	23,292	32,134
	máx	3610	4118	702,648	831,245
	Var	5611,839	14517,352	2304,533	4639,459
	DP	74,912	120,488	48,005	68,113

Após as fases de Preparação dos Textos, Extração de Termos e Construção de Pessoas, para a Base de Dados 1 (Análise de Sentimentos Baseada em Aspectos) foram selecionados 4132 termos, reduzidos para 2496 termos dado o processo de stemização, e para a Base de Dados 2 (semântica na Mineração de Textos) foram selecionados 8342 termos, reduzidos para 4953 termos pelo processo de stemização. Na Tabela 1 é possível observar o impacto da stemização nos conjuntos de dados, bem como comparar as métricas TF e TF-IDF. Nesse caso, o processo de stemização reduziu a quantidade de termos em 60,4% para a Base de Dados 1, e em 59,37% para a Base de Dados 2. Em relação à

frequência dos termos, é normal existir uma grande variância de valores, pois existem mais termos que aparecem poucas vezes do que termos que ocorrem em grande quantidade. Nos testes realizados, a stemização e a métrica TF-IDF obtiveram os melhores resultados, destacados em negrito na Tabela 1. Além do número de termos, também foi possível reduzir a Variância (comparando TF e TF-IDF com *stemming*) e por conseguinte o Desvio Padrão, em razão do aumento das frequências mínima e média e diminuição da frequência máxima.

Dado a predominância de termos com frequências muito baixas, mesmo após a stemização e uso da métrica TF-IDF, foi necessário selecionar os termos considerados mais representativos para o conjunto de dados. Para isso, foi aplicada a Lei de Zipf para construção dos gráficos de frequências ilustrados na Figura 2. Os valores das frequências estão ordenados em ordem decrescente, o que permite denominar as curvas como Curvas de Zipf [Zipf 1949]. É possível perceber que tanto para TF quanto para TF-IDF há uma predominância de termos com baixa frequência e uma parcela pequena de termos com frequência alta.

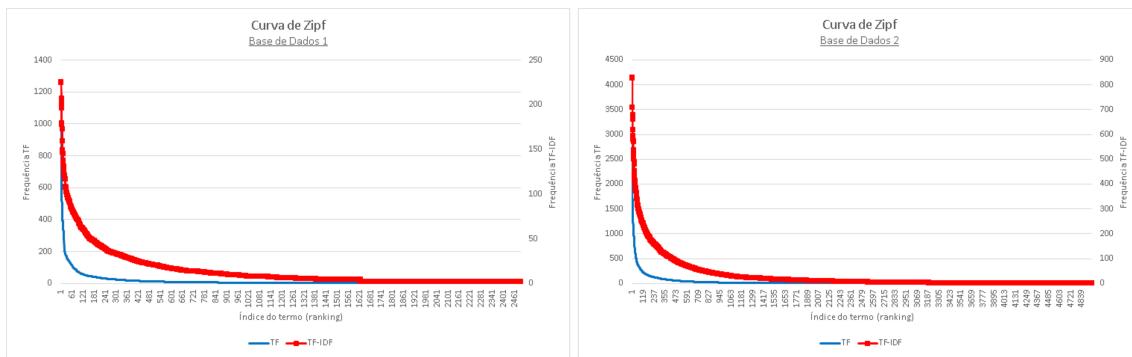


Figura 2. Curvas de Zipf para a Base de Dados 1 e Base de Dados 2, respectivamente, ambas com termos stemizados.

Para a métrica TF, a seleção de termos pode ser realizada “cortando-se” o número de termos com base em um intervalo intermediário de frequências, de 50 à 500, por exemplo, o que pode implicar na remoção de termos muito relevantes. No caso da métrica TF-IDF, entretanto, esse corte (denominado Corte de Luhn [Luhn 1958]) pode ser realizado de forma menos subjetiva, eliminando-se apenas os termos com frequência menor à um limite inferior definido. Para definição do corte inferior deste último caso, para cada base de dados foi definida a metade da quantidade total de documentos como sendo um número representativo de termos, ou seja, considerou-se que um par de documentos pode ser distinguido por um único termo. Nos testes realizados, para a Base de Dados 1 foram obtidos 219 termos eliminando-se os termos com frequência menor do que 42,311, e para a Base de Dados 2, 846 termos foram obtidos eliminando-se os termos com frequência inferior à 44,362. Por motivos de facilidade, esses limites foram definidos como 40 e 50, respectivamente, resultando em um total de 233 termos para a Base de Dados 1 e 774 termos para a Base de Dados 2.

5. Extração de padrões

Algoritmos de agrupamento se baseiam em características comuns entre os dados para criar grupos bem definidos e com pouca semelhança entre si. Em dados textuais, essas

características podem ser extraídas dos termos, do assunto, do contexto ou da semântica dos documentos, por exemplo. Isso implica em relacionamentos espúrios que podem levar à interpretação equivocada dos resultados, já que um mesmo atributo pode estar muito presente em um documento e não estar presente em outro, e mesmo assim existir grande similaridade entre esses documentos. Por esse motivo, normalmente é necessário projetar o espaço de características dos dados para aplicar os algoritmos de agrupamento, muitas vezes utilizando planos bidimensionais para essa tarefa.

Para projetar os dados textuais, foi utilizado o método PCA, por reduzir as dimensões dos dados a partir de combinações lineares com o objetivo de selecionar suas características mais significativas. Os pontos das coordenadas bidimensionais de cada documento (resultantes do processo de Redução de Dimensionalidade) foram agrupados utilizando o algoritmo K-means, por ser um dos principais algoritmos de agrupamento [Wu et al. 2008], ter complexidade computacional linear [Tan et al. 2005], apresentar interpretação natural de similaridade baseada em distância (nesse caso, foi utilizada a distância cosseno, por calcular o ângulo entre os pares de pontos e com isso transmitir de forma mais relativa a similaridade entre os documentos) e ter bons resultados em diversas aplicações. O número de grupos (valor de k) foi definido como 5, imposto pela ferramenta Pinda por ser um indicativo de usabilidade baseado no contexto de Mecanismos de Busca [Antunes and Eler 2017].

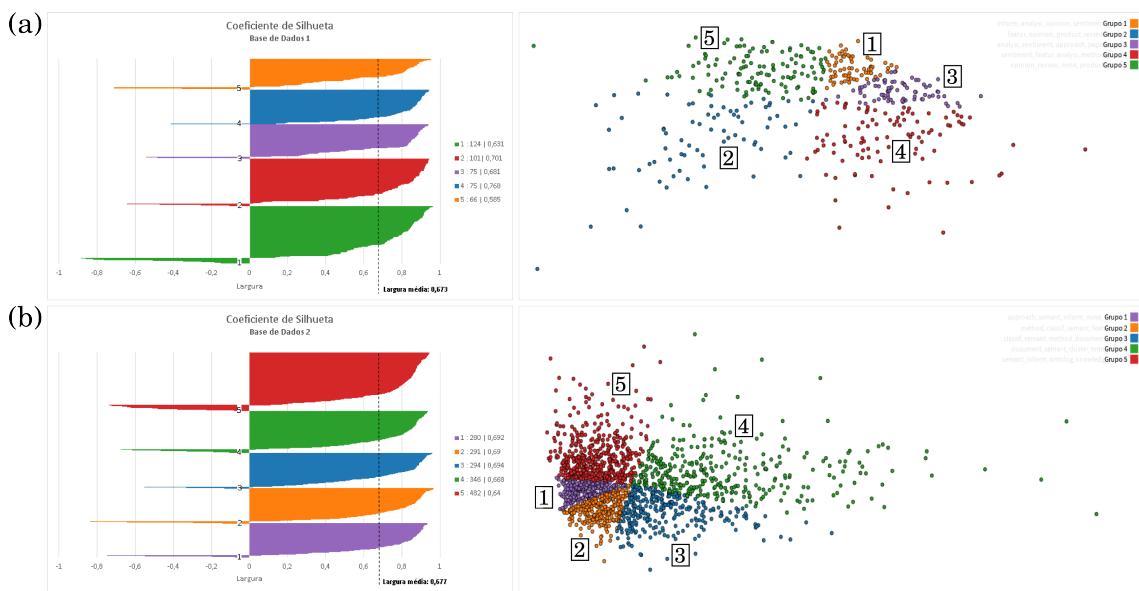


Figura 3. Coeficientes de Silhueta dos agrupamentos da Base de Dados 1 (a) e da Base de Dados 2 (b).

Como é possível notar na Figura 3 (a), o agrupamento resultante da Base de Dados 1 (contendo 441 documentos) possui menor densidade do que o agrupamento resultante da Base de Dados 2 (contendo 1693 documentos), ilustrado na Figura 3 (b). Isso pode ser explicado pela diferença na quantidade de documentos presentes nas duas bases, e pelo fato da Base de Dados 2 possuir quase quatro vezes o tamanho da Base de Dados 1. Em ambos os agrupamentos, entretanto, foram gerados grupos bem definidos e distintos entre si, com larguras médias de silhueta similares (0,673 para a Base de Dados 1 e 0,677 para a Base de Dados 2) e próximas de 1, indicando boa qualidade. Nas legendas dos gráficos

de Coeficiente de Silhueta, também é possível notar o índice de cada grupo, seguido da quantidade de elementos presentes e o valor de seu coeficiente.

Com o objetivo de caracterizar os grupos gerados e favorecer no entendimento dos resultados, foram atribuídos descritores (termos com alta representatividade no grupo de documentos) a todos os grupos obtidos, permitindo verificar os “principais” (5 termos mais frequentes) termos de cada conjunto. Baseado no primeiro descritor de cada grupo ilustrado na Figura 3 (a), é possível concluir que a Base de Dados 1 é formada pelos conceitos de “informação”, “característica”, “análise”, “sentimento” e “opinião”, respectivamente derivados dos termos stemizados “inform” (*information*), “featur” (*feature*), “analysi” (*analysis*), “sentiment” (*sentiment*) e “opinion” (*opinio*). De forma análoga, é possível concluir que a Base de Dados 2 é formada pelos conceitos de “abordagem”, “método”, “classificação”, “documento” e “semântica”, respectivamente derivados dos termos “approach” (*approach*), “method” (*method*), “classif” (*classification*), “document” (*document*) e “semant” (*semantic*). Como a Base de Dados 1 trata de “Análise de Sentimentos Baseada em Aspectos” e a Base de Dados 2 trata da “semântica na Mineração de Textos”, é válido concluir que o uso de descritores pode ser útil na análise textual de documentos.

6. Pós-processamento

Talvez o principal diferencial da metodologia empregada, o uso de múltiplas visualizações coordenadas contribuem para uma análise menos subjetiva e mais ampla do mesmo conjunto de dados. As funcionalidades de interação da ferramenta possibilitam entender os relacionamentos entre os documentos sob diferentes perspectivas, implicando em um aumento da percepção do usuário sobre os padrões identificados, como pode ser observado na Figura 4.

A ferramenta Pinda possui seis técnicas coordenadas de Visualização: Miniatura (representa a página do conteúdo original da publicação, permitindo navegar por ela sem sair da ferramenta), Diretórios (exibe a organização hierárquica dos grupos formados, similar à organização de arquivos dos sistemas operacionais), *Scatterplot* (representa a similaridade dos principais grupos por meio de distância e cor, plotados em um gráfico bidimensional), *Treemap* (exibe a organização hierárquica dos grupos formados, distinguídos por atributos de cor e grandeza), *Sunburst* (similar à técnica *Treemap*, com a diferença de exibir todos os níveis hierárquicos de um grupo específico) e *Snippets* (exibe o conteúdo (título, URL e resumo) original das bases de dados por meio da representação textual comumente utilizada em Mecanismos de Busca) [Antunes and Eler 2017].

Para diferenciar os trabalhos aceitos e rejeitados, foi adicionada uma nova funcionalidade à ferramenta, que destaca os documentos aceitos/rejeitados por meio de transparência dos pontos da visualização *Scatterplot*. Nos testes realizados, não foi possível identificar alguma relação entre os trabalhos aceitos e rejeitados com base nessa informação, devido à proximidade dos pontos (*Scatterplot*) e provável ausência de padrões que caracterizam subjetivamente publicações como aceitas ou rejeitadas. Na Figura 4 (a), o documento em destaque (*ASPECT-BASED OPINION MINING FROM PRODUCT REVIEWS*) trata-se de um dos principais artigos de Análise de Sentimentos Baseada em Aspectos que aborda o modelo *Latent Dirichlet Allocation* (LDA). Com o auxílio da visualização *Scatterplot*, é possível notar que, embora o grupo desse docu-

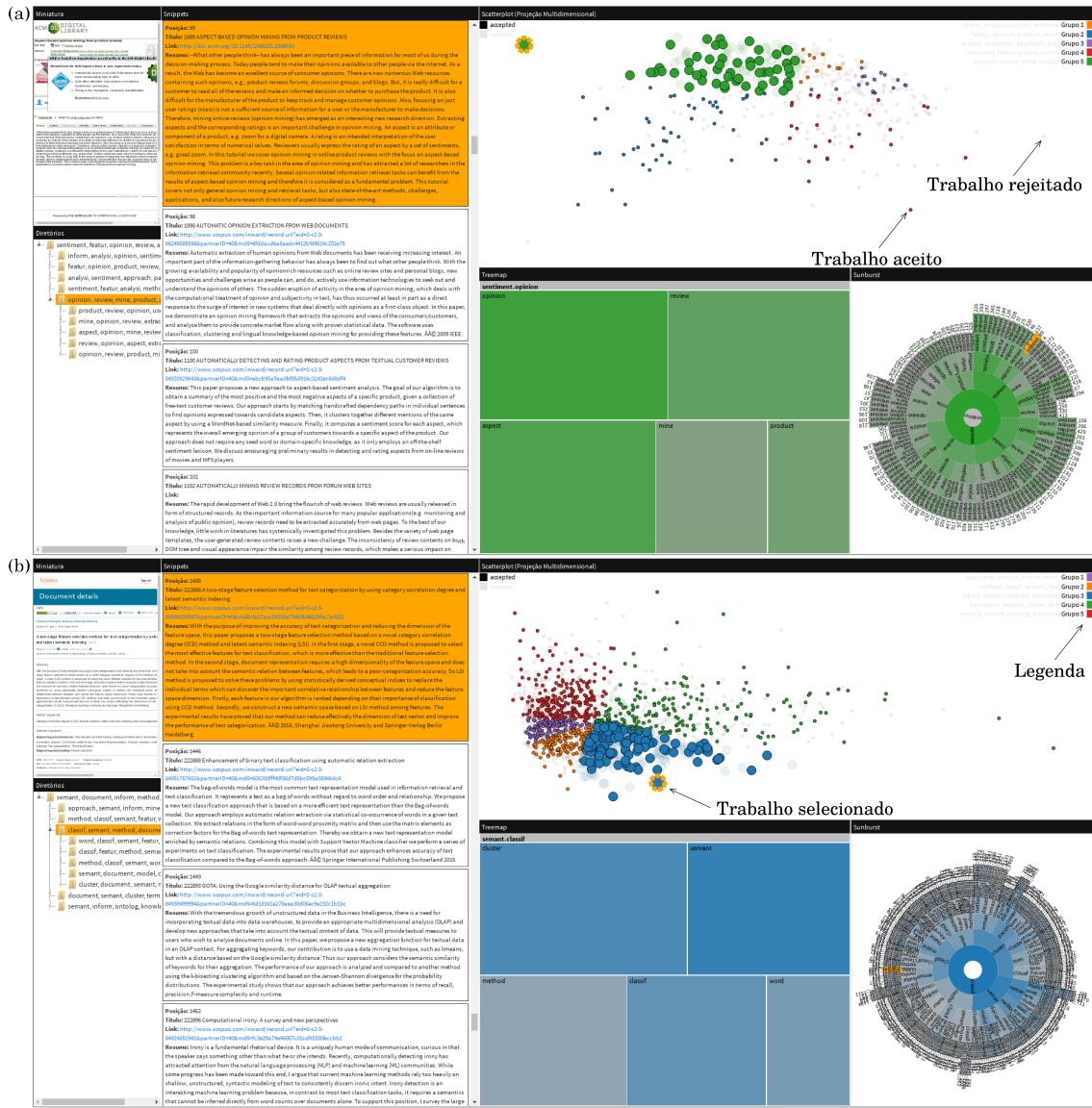


Figura 4. Exemplo de utilização da ferramenta Pinda para a Base de Dados 1 (a) e Base de Dados 2 (b).

mento (caracterizado com os descritores *opinion*, *review*, *mine*, *product* e *aspect*) seja bem compacto, o mesmo é um *outlier* ou ponto atípico, já que está longe do centro de seu grupo.

Sobre o conteúdo dos grupos, foi possível constatar sobre a Base de Dados 1 (ver Figura 3 (a)) que o Grupo 1 (laranja) é composto por artigos que usam técnicas de Aprendizado de Máquina para Análise de Sentimentos; o Grupo 2 (azul) apresenta artigos que utilizam o termo “feature” ao invés de “aspect”; o Grupo 3 (roxo) reúne artigos que utilizam técnicas de PLN, identificação de tópicos, aplicações e revisões da literatura; o Grupo 4 (vermelho) trata de técnicas de Análise de Sentimentos em idiomas diferentes do inglês; o Grupo 5 (verde) reúne publicações sobre técnicas de Análise de Sentimentos que usam PLN. Em relação à Base de Dados 2 (ver Figura 3 (b)), foi possível constatar que o Grupo 1 (roxo) reúne o segundo maior grupo de *surveys*, com métodos e algoritmos

baseados em verbos e domínio de aplicação predominantemente baseado em textos da Web; o Grupo 2 (laranja) é o único que não possui *surveys*, embora pareça estar relacionado à trabalhos sobre doenças e medicamentos; o Grupo 3 (azul) reúne trabalhos sobre classificação e agrupamento de textos, Análise de Sentimentos, PLN e disambiguação de palavras baseado em conhecimento externo (ver trabalho destacado na Figura 4 (b)); o Grupo 4 (verde) trata de agrupamento e métodos ou algoritmos baseados uso de conhecimento externo, principalmente; o Grupo 5 (vermelho) reúne publicações sobre uso de ontologias, extração de conhecimento e genes, indicando similaridade com a área médica. Além disso, trata-se do grupo com maior número de *surveys* e com predominância de aplicações sobre saúde e vida.

7. Conclusões

Com base nos resultados da Seção 5, e nas análises da Seção 6, é possível concluir que a metodologia empregada nesse trabalho permitiu agrupar de forma eficiente os documentos de ambas as bases de dados. As múltiplas visualizações foram essenciais para a descoberta de artigos relevantes, mesmo sendo pontos “fora da curva”.

A coordenação das visualizações mostrou-se de fundamental importância para o entendimento menos subjetivo das relações entre os documentos, possibilitando diferentes perspectivas do mesmo conjunto de dados. De forma geral, isso tende a apresentar conclusões menos subjetivas e resultados mais precisos.

Como trabalhos futuros, podem ser desenvolvidos sistemas de monitoramento contínuo de publicações científicas ou aplicações em outras áreas de atuação, como análise de notícias, redes sociais ou qualquer outra que possua dados textuais e necessite de agrupamento, até mesmo a organização de documentos pessoais.

Referências

- Abbas, O. A. (2008). Comparisons between data clustering algorithms. *Int. Arab J. Inf. Technol.*, 5(3):320–325.
- Antunes, J. and Eler, D. M. (2017). Pinda: visualização hierárquica para o agrupamento de resultados de mecanismos de busca. *Revista de Informática Teórica e Aplicada (RITA)*.
- Boukhelifa, N., Roberts, J., and Rodgers, P. (2003). A coordination model for exploratory multiview visualization. In *Coordinated and Multiple Views in Exploratory Visualization, 2003. Proceedings. International Conference on*, pages 76–85.
- Boyack, K. W., Klavans, R., and Börner, K. (2005). Mapping the backbone of science. *Scientometrics*, 64(3):351–374.
- Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *American Association for Artificial Intelligence*.
- Kitchenham, B. (2004). Procedures for performing systematic reviews. Keele university. technical report tr/se-0401, Keele University and NICTA, Department of Computer Science, Keele University, UK.
- Leydesdorff, L., Rafols, I., and Chen, C. (2013). Interactive overlays of journals and the measurement of interdisciplinarity on the basis of aggregated journal-journal ci-

- tations. *Journal of the American Society for Information Science and Technology*, 64(12):2573–2586.
- Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM J. Res. Dev.*, 2(2):159–165.
- North, C. and Shneiderman, B. (2000). Snap-together visualization: can users construct and operate coordinated visualizations? *International Journal of Human-Computer Studies*, 53(5):715 – 739.
- Paulovich, F. V. (2008). *Mapeamento de dados multi-dimensionais - integrando mineração e visualização*. PhD thesis, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos - SP, Brasil.
- Porter, M. F. (1997). Readings in information retrieval. In Sparck Jones, K. and Willett, P., editors, *Readings in Information Retrieval*, chapter An Algorithm for Suffix Stripping, pages 313–316. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.
- Rezende, S. (2003). *Sistemas inteligentes: fundamentos e aplicações*. Manole.
- Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53 – 65.
- Silva, F. N., Amancio, D. R., Bardosova, M., da F. Costa, L., and Jr., O. N. O. (2016). Using network science and text analytics to produce surveys in a scientific topic. *Journal of Informetrics*, 10(2):487 – 502.
- Sinoara, R. A., Antunes, J., and Rezende, S. O. (2017). Text mining and semantics: a systematic mapping study. *Journal of the Brazilian Computer Society (JBCS)*.
- Sparck Jones, K. (1988). Document retrieval systems. In Willett, P., editor, *Document Retrieval Systems*, chapter A Statistical Interpretation of Term Specificity and Its Application in Retrieval, pages 132–142. Taylor Graham Publishing, London, UK, UK.
- Tan, P.-N., Steinbach, M., and Kumar, V. (2005). *Introduction to Data Mining, (First Edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.
- van Eck, N. J. and Waltman, L. (2014). Citnetexplorer: A new software tool for analyzing and visualizing citation networks. *Journal of Informetrics*, 8(4):802 – 823.
- Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z.-H., Steinbach, M., Hand, D. J., and Steinberg, D. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1):1–37.
- Zamboni, A., Thommazo, A., Hernandes, E., and Fabbri, S. (2010). Start uma ferramenta computacional de apoio à revisão sistemática. In *Proc.: Congresso Brasileiro de Software (CBSOFT 10), Salvador, Brazil*.
- Zipf, G. K. (1949). *Human Behaviour and the Principle of Least Effort*. Addison-Wesley.