GOOD: UM MECANISMO PARA ORDENAÇÃO COERENTE DE ARTIGOS CIENTÍFICOS

Andrés Velasteguí C., Wallace A. Pinheiro, Marcos V. Peixoto*
Instituto Militar de Engenharia (IME), Departamento de Ciências da Computação –
Praça General Tibúrcio, 80, 22290-270, Praia Vermelha, Rio de Janeiro, RJ, Brasil.
**afvelastegui@gmail.com*

RESUMO

A pesquisa científica é uma das atividades mais importantes das comunidades acadêmicas, pois através dela se motiva a discussão científica e a disseminação do conhecimento. A primeira fase dessa atividade requer uma busca exaustiva de informações relevantes que possam contribuir na redação de artigos e teses. Através dos sistemas de recomendações um pesquisador tem acesso a diversos documentos científicos, não obstante, a ordenação coerente das informações obtidas pode constituir um grande problema. Alguns trabalhos propõem soluções para resolver problemas similares, porém os algoritmos utilizados apresentam grande complexidade computacional, o que dificulta o seu uso prático. Este trabalho propõe uma alternativa para organizar documentos através de algoritmos com complexidade menores que outros trabalhos relacionados. Para isso, foi estabelecido um framework, chamado GOOD (**G**enetic **O**rder **O**f **D**ocuments), que inclui algumas técnicas de mineração de dados. Os resultados iniciais mostram que o framework proposto fornece uma significativa diminuição na complexidade temporal e espacial para organizar uma sequência de artigos, conservando a coerência da sequência final obtida.

Palavras-chave: sistemas de recomendações, clusterização de documentos, ordenação de documentos.

ABSTRACT

Scientific research is one of the most important activities in the academic community, because it promotes the scientific discussion and the dissemination of knowledge. The first phase of this activity requires an exhaustive search of relevant information that may assist in the edition of articles and theses. Through academic recommendation systems, a researcher has acess to numerous scientific papers; however, the coherent ordering of the information obtained can be a major problem. Some works propose solutions to solve similar problems, but the algorithms used have high computational complexity, which hinders its practical use. This paper proposes an alternative to organizing throught algorithms with lower complexity than other related research. For that, we implement a framework called GOOD (**G**enetic **O**rder **O**f **D**ocuments), which includes some techniques of data mining. Initial results show that the proposed framework provides a significant decrease in the tem-

poral and spatial complexity to organize a sequence of articles preserving the coherence of the result obtained.

Keywords: recommendation systems, document clustering, document organization.

INTRODUÇÃO

Na última década, a comunidade científica tem incrementado seu interesse pelos sistemas de recomendações, com atenção especial para os sistemas acadêmicos. Desde 2007, essas aplicações têm experimentado um rápido crescimento (PARK, 2012), chegando a ser um instrumento fundamental para promover a discussão científica e a disseminação do conhecimento. Assim, estudos recentes mostram que os maiores sistemas de recomendações acadêmicos (Google Scholar e Microsoft Academic Search) disponibilizam aproximadamente 114 milhões de documentos escritos no idioma inglês (KHABSA, 2014).

O bem-sucedido desenvolvimento desses sistemas gira em torno da recuperação de informação relevante, no entanto, as necessidades dos pesquisadores, hoje em dia, exigem pesquisas mais complexas e refinadas, normalmente envolvendo uma forte carga semântica. Por exemplo, é fácil obter um grande volume de dados relevantes de diversas fontes, mas tarefas como: a revisão exaustiva, a organização coerente e a forma de apresentar a informação coletada ainda são um desafio para os novos pesquisadores.

Entre os problemas que um usuário tem no momento de utilizar estes sistemas podem-se citar: (i) uma limitada revisão da literatura (PARK, 2012; KHABSA, 2014), (ii) a existência de informação duplicada (ORTEGA, 2014), (iii) a desvalorização de instituições, assuntos e/ou autores (ORTEGA, 2014) e (iv) a falta de um mecanismo de encadeamento coerente da informação.

Trabalhos prévios abordaram esses problemas sob diferentes perspectivas, tais como: decidir em que sequência apresentar um conjunto de informações pré-selecionado (KARAMANIS, 2004), representar e medir a coerência entre partes de um texto (BARZILAY, 2008), encontrar artigos relevantes a partir de uma representação de interesse dada (BASU, 2011), diminuir a complexidade e melhorar a escalabilidade dos sistemas de recomendações (KARAMANIS, 2009; SARWAR, 2000, 2001), estudar as propriedades de ordenar informação no domínio das notícias (BARZILAY, 2011), implementar algoritmos para a estruturação de textos (BARZILAY, 2004, 2011; KARAMANIS, 2009; LAPATA, 2003), encadear notícias utilizando a sabedoria das multidões (RODRIGUES, 2010), encadear notícias relacionadas a partir de duas notícias pré-selecionadas (SHAHAF, 2010), encadear notícias utilizando reconhecimento de implicação textual (CAVALCANTE, 2013), entre outros.

O presente trabalho propõe uma alternativa complementar aos sistemas existentes, pois o foco deste estudo é a organização coerente de artigos (permutação) para facilitar a leitura abrangente de um determinado tema.

A permutação é um conceito que expressa a ideia de arranjos de objetos

não repetidos e que conservam certa ordem para satisfazer critérios específicos (MACMAHON, 2001). Dado que nesta pesquisa, pretende-se ordenar artigos não repetidos e que conservem a coerência da leitura, podemos compará-lo com o problema do caixeiro viajante (TRICOIRE, 2010). Assim, os artigos podem ser vistos como as cidades a ser visitadas e as relações de coerência entre os artigos, como os caminhos entre as cidades. Consequentemente, a complexidade de permutar documentos é da ordem exponencial ou fatorial. Deste modo, pode ser fácil ordenar um conjunto de 5 artigos. Mas o que aconteceria se um pesquisador dispusesse de uma grande coleção de artigos relacionados ao mesmo tema? Qual critério deveria ser usado para selecionar artigos relevantes? E quais seriam os critérios para ordená-los?. Por exemplo, considerando um repositório com 100 artigos e sequências de tamanho 7, teríamos: $100! / 7! \times (100 - 7)! = 1.6 \times 10^{10}$ possíveis sequências a serem geradas. Uma alternativa de permutação seria ordenar os artigos em função da data, mas isto nem sempre pode gerar uma leitura abrangente. Além disso, existe a possibilidade de que artigos mais novos expliquem de uma melhor maneira um assunto mais antigo e vice-versa. Então, o sequenciamento coerente da informação não deveria ficar limitado a um simples critério temporal.

O propósito deste trabalho é analisar um conjunto de técnicas que permitam diminuir a complexidade de ordenar artigos, considerando a hipótese de que não é necessário calcular a relação de coerência entre todos os artigos da coleção para encadear documentos e gerar uma leitura coerente (sequência de artigos).

Como resultado desta pesquisa foi implementado um framework, chamado GOOD (***G**enetic **O**rder **O**f **D**ocuments*), que permite explorar um grande espaço de busca, procurando a diversidade de conteúdo e a ordenação coerente de artigos para facilitar uma leitura abrangente sobre um assunto para os usuários.

O trabalho está organizado da seguinte forma: na Seção 2 se define o conceito de coerência; a Seção 3 apresenta o framework, a Seção 4 mostra um experimento para avaliar a hipótese, a Seção 5 mostra os resultados e discussão e a Seção 6 expõe as conclusões do trabalho.

DEFINIÇÃO DE COERÊNCIA

Segundo (STORRER, 2002), a coerência é uma característica que reflete a progressão lógica e complementar das ideias para tornar um texto claro e compreensível. Baseados nesse conceito, são apresentadas as definições de coerência que serão utilizadas ao longo desta pesquisa.

Definimos a **coerência de uma leitura ou sequência** *S* de comprimento *m*, como uma medida escalar que representa a soma das coerências parciais entre os pares de artigos adjacentes (A_i, A_{i+1}):

$$\text{coerência}(S) = \sum_{i=1}^{m-1} \text{coerência}(A_i, A_{i+1}).$$

A coerência entre um par de artigos (A_i, A_j) é definida como a soma dos valores de similaridade (*Sim*), implicação textual (*Imp*) e da análise temporal (*Temp*)

calculados entre eles mediante a função:

$$f\text{coerência}(\mathrm{A}_i,\ \mathrm{A}_j) = \mathrm{Sim}(A_i, A_j) + \mathrm{Imp}(A_i, A_j) + \mathrm{Temp}(A_i, A_j)$$

A similaridade (*Sim)* - é uma métrica utilizada pela área de Recuperação da Informação para determinar o valor de semelhança entre uma consulta e um documento. Adotamos o *score* de similaridade implementado por (MCCANDLESS, 2010) para quantificar essa relação entre um par de artigos por meio da função:

$$f_{Sim}(\mathrm{A}_i,\ \mathrm{A}_j) = \mathrm{coord}(\mathrm{A}_i,\ \mathrm{A}_j) \cdot \mathrm{queryNorm}(\mathrm{A}_i) \cdot (\textstyle\sum_{i=1}^{n} \mathrm{TF}^{1/2}\ (\mathrm{t}_i \in \mathrm{A}_j) \cdot \mathrm{IDF}^2(\mathrm{t}_i \in \mathrm{A}_j) \cdot \mathrm{boost}(\mathrm{t}_i \in \mathrm{A}_j) \cdot \mathrm{norm})$$

Onde: A_i é o artigo de interesse, A_j é o artigo comparado, *coord* é o número de termos de A_i que constam em A_j; *queryNorm* é um fator de normalização para os termos de A_i; a frequência de termo ou *TF* é o número de vezes que um termo t_i aparece em A_j; a frequência inversa de documento ou *IDF* é o número de documentos que contêm um termo *t* específico; *boost* é um fator de importância para um termo de A_j; e *norm* é um fator de importância para um campo de indexação.

Implicação (Imp) - é a relação unidirecional que permite identificar se a hipótese *h* pode ser inferida a partir de um texto *t*. Esta medida é composta de dois valores: (i) a similaridade léxica *simLex* e (ii) *ref*, que é a verificação do título de A_i nas referências de A_{i+1}. O valor de *Imp* é calculado por meio da função:

$$f_{Imp}(\mathrm{A}_i,\mathrm{A}_j) = \frac{\sum_{i=1,j=1}^{n} simLex(h_j,t_i)}{\sum_{i=1}^{n} TF(h_i)} + \mathrm{ref}(\mathrm{A}_i,\ \mathrm{A}_j),$$

Para encontrar a similaridade léxica, cada hipótese do vetor A_{i+1}= [h_1, h_2, h_3, ..., h_m] é procurada no vetor de textos de A_i= [t_1, t_2, t_3, ..., t_n] aplicando a seguinte função:

$$SimLex(t_i, h_j) = \begin{cases} max(TF(h_j), TF(t_i)) & ,\text{se } h_j \subset \text{sinônimo}(t_i); \\ -TF(h_j) & ,\text{caso contrário.} \end{cases}$$

Logo, o título de A_j, é procurado mediante expressões regulares entre as referências de A_j. Então, *ref* toma um valor conforme a regra:

$$ref = \begin{cases} 0.5 & ,\text{se título}(A_i) \subset \text{referências}(A_j); \\ 0.0 & ,\text{caso contrário.} \end{cases}$$

Análise temporal (Temp) - é um fator baseado na diferença de anos entre um par de artigos. A menor diferença corresponde um maior valor de relação temporal. Esse valor é normalizado pela variável α no intervalo de [0,1] e calculado mediante a função:

$$f_{Temp}(A_i, A_{i+1}) = \frac{1}{(ano(A_i) - ano(A_{i+1}))^2 + \alpha}$$

Assim, considerando um valor de α=1.00, um artigo de 2011 terá uma maior possibilidade de ser encadeado com um de 2010 ou de 2012, em vez de um artigo de 2008 ou de 2014. Isto pode ser observado na Figura 1.

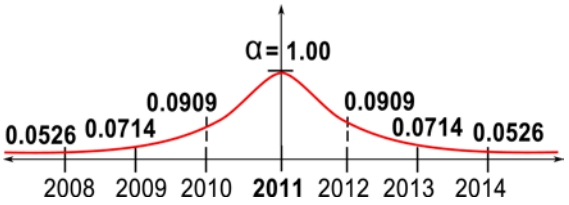

Figura 1. *Exemplo da Análise Temporal*

O FRAMEWORK GOOD

O framework consta de dois conjuntos de componentes: (i) o tratamento das informações mostrado na Figura 2 e (ii) os cálculos de coerência e o mecanismo de encadeamento apresentado na Figura 3. A seguir, é explicado o funcionamento dos componentes que realizam o tratamento das informações.

- Mediador – realiza o *download* de artigos no formato PDF e a geração de metadados correspondente (arquivos em formato Bibtex).
- Conversor – converte os artigos de formato PDF para texto e junta os metadados para gerar artigos completos.

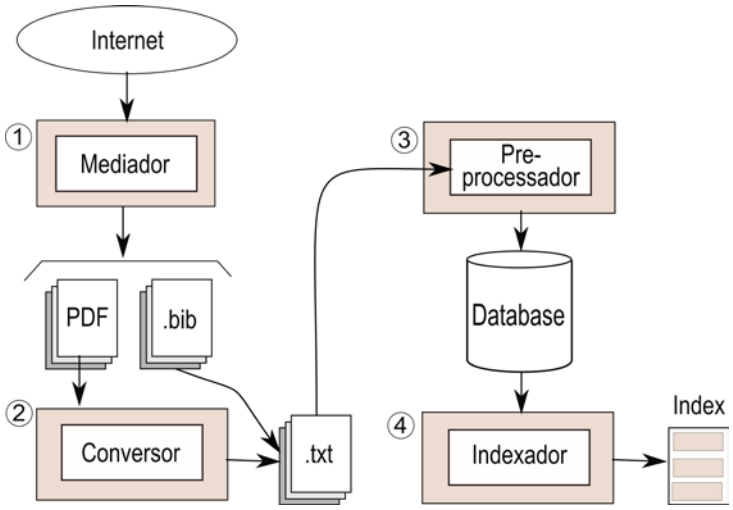

Figura 2. *Tratamento das Informações*

- Pré-processamento – divide o artigo completo em seções (título, autor, data de publicação, jornal, abstract, conteúdo e referências) para facilitar a sua análise e indexação.
- Indexador – executa processos de tokenização, remoção de palavras irrelevantes, contagem de termos, aplicação de métricas de frequência, e cria os índices no sistema de arquivos.

A Figura 3 mostra as técnicas utilizadas para diminuir a complexidade de permutar artigos a partir de uma coleção de documentos.

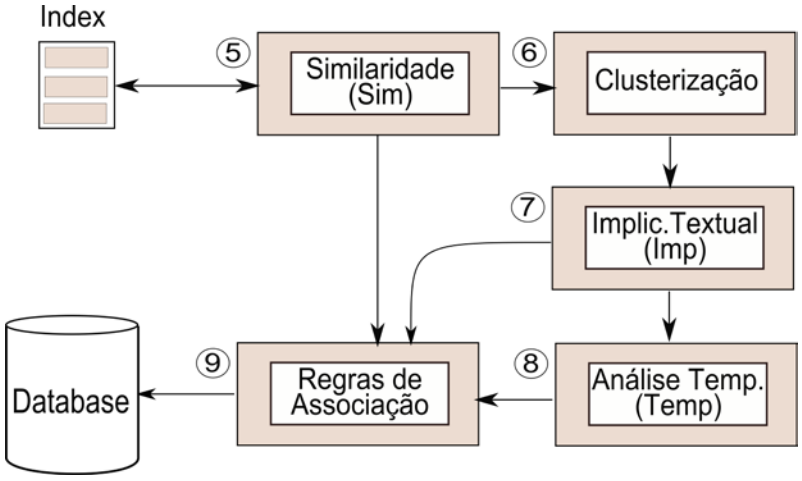

Figura 3. *Cálculos de Coerência e Mecanismo de Encadeamento*

- Similaridade (Sim) – a partir de um documento de interesse selecionado são recuperados todos os documentos relevantes da coleção.
- Clusterização – tem dois objetivos: (i) encontrar o *cluster* mais próximo[1] e (ii) diminuir o número de documentos para os processos de implicação textual e análise temporal. Dentre os três algoritmos de clusterização analisados (*Bisecting k-means*, Clusterização por Árvores de Sufixos (STC) e Lingo) foi selecionado o *Bisecting k-means,* pois ele garante a geração de clusters não superpostos. Uma comparação dos algoritmos de clusterização é mostrada na Tabela 1.

Tabela 1. *Comparação de Algoritmos de Clusterização (JANRUANG, 2011; NAVIGLI, 2010; OSINSKI, 2005, 2007)*

Métrica		**Lingo**	**STC**	**Bisecting K-means**
Densidade	Eps, MinPts	Alta	Baixa	Baixa
Extração de conceito	Descrição	Boa	Boa	Limitada
	N-gramas	1-10	1-5	1
Superposição	Doc. Repetidos Clusters	Sim	Sim	Não
Escalabilidade	Processamento >1000 doc.	Baixa	Alta	Baixa

- Implicação Textual (Imp) – quantifica a relação de implicação entre o artigo de interesse e cada artigo do seu cluster mais próximo.
- Análise Temporal (Temp) – calcula a relação de análise temporal entre o artigo de interesse e os artigos do seu cluster mais próximo.

Após calcular a similaridade, a implicação textual e a análise temporal, cada par de artigos é submetido a um conjunto de regras de associação para garantir um mínimo de qualidade na relação de coerência. Essas regras são detalhadas a seguir:

1 Grupo de documentos que guardam um contexto comum com o artigo de interesse

$$Regras = \begin{cases} salvar, & \text{se (Sim} \geq \text{mínSim) e (Imp} \geq \text{mínImp);} \\ salvar, & \text{se (Sim} \geq \text{mínSim) e (Temp} \geq \text{mínTemp);} \\ salvar, & \text{se (Imp} \geq \text{mínImp) e (Temp} \geq \text{mínTemp);} \\ excluir, & \text{em outros casos.} \end{cases}$$

Onde: mínSim, mínImp e mínTemp são os valores mínimos que deve ter uma relação de coerência para ser considerada como aceitável e ser salva no banco de dados. Esses valores deveriam ser configurados por um grupo de especialistas.

Finalmente, as relações armazenadas no banco de dados são a base para o mecanismo de encadeamento apresentado no Algoritmo 1, onde: n é o comprimento da sequência desejado, m é o índice das leituras geradas, o símbolo ∅ representa um conjunto sem elementos e i é o índice que varia até o tamanho da sequência.

A função de incremento do comprimento das sequências é mostrada no Algoritmo 2.

Algoritmo 1. *Mecanismo de Encadeamento*

Parâmetros: n ← 7, conjuntoLeituras ← ∅, m ← 1, i ← 2, selecionados ← ∅.
função ENCADEARARTIGOS (A_0)
selecionados ← adicionar (A_0)
cluster ← buscarClusterMaisPróximo (A_0)
para todo artigo ∅ cluster **faça**
    $leitura_m$ ← adicionar (A_0)
    $leitura_m$ ← adicionar (artigo)
    selecionados ← adicionar (artigo)
    conjuntoLeituras ← adicionar ($leitura_m$)
fim para
enquanto i < n **faça**
    i ← i+1
    incrementarComprimento (conjuntoLeituras)
fim enquanto
devolve conjuntoLeituras
fim função

Algoritmo 2. *Função de Incremento do Comprimento das Sequências*

função INCREMENTARCOMPRIMENTO(conjuntoLeituras)
para toda leitura ∅ conjuntoLeituras **faça**
    cluster ← buscarClusterMaisPróximo (último artigo da leitura)
    para todo artigo ∅ cluster **faça**
    se artigo ∅ selecionados **então**
    $leitura_m$ ← adicionar (artigo)
    fim se
    selecionados ← adicionar (artigo)
    conjuntoLeituras ← adicionar ($leitura_m$)
        fim para
    fim para
    remover leituras com comprimento < i
    devolve conjuntoLeituras
fim função

EXPERIMENTO: AVALIAÇÃO DO ENCADEAMENTO

O objetivo deste experimento é verificar se o mecanismo de encadeamento é capaz de gerar uma leitura abrangente a partir das relações de coerência armazenadas no banco de dados. Para isso, calculamos a correlação de uma sequência ou leitura gerada pelo mecanismo de encadeamento e as sequências geradas por um grupo heterogéneo de 12 avaliadores, entre eles, professores e alunos de mestrado do Instituto Militar de Engenharia. A métrica de avaliação usada foi o coeficiente de correlação de postos de Spearman (ρ_s). Esta medida de associação não paramétrica descreve a direção e o grau em que uma variável ordinal está linearmente relacionada com outra e permite rejeitar a hipótese nula (afirmar que as duas variáveis não estão relacionadas). Considerando uma coleção com 2.000 documentos referentes a *"data mining"* e um tamanho de leitura arbitrário de 7 artigos (observações), o mecanismo gerou 23 sequências. Dentre elas, foi selecionada a melhor sequência para ser avaliada. Segundo (MYERS, 2010), um valor de ρ_s □ 0.786 para o número de observações selecionado, permitiria rejeitar a hipótese nula. As sequências do mecanismo e as sequências geradas pelo grupo de avaliação são mostradas na Tabela 2.

Tabela 2. *Sequências geradas pelo mecanismo e pelo grupo avaliador*

Encadeamento	Ordem						
Mecanismo	1	2	3	4	5	6	7
Pesquisador 1	1	5	3	6	2	4	7
Pesquisador 2	2	3	1	4	6	5	7
Pesquisador 3	4	1	3	2	6	5	7
Pesquisador 4	1	2	4	3	5	6	7
Pesquisador 5	4	2	1	6	3	5	7
Pesquisador 6	2	1	6	3	5	4	7
Pesquisador 7	1	3	6	5	2	4	7
Pesquisador 8	4	1	5	6	3	2	7
Pesquisador 9	4	2	3	1	6	5	7
Pesquisador 10	2	5	1	6	3	4	7
Pesquisador 11	2	3	1	6	5	4	7
Pesquisador 12	1	3	2	4	5	6	7

RESULTADOS

A Figura 4 mostra a relação linear entre a sequência gerada pelo mecanismo e cada sequência ordenada pelo grupo avaliador. Os pontos correspondem às observações do mecanismo comparadas com as observações humanas; a linha "F" representa o ajuste da correlação; a área entre as linhas "C" é o intervalo de confiança (95%) e a área entre as linhas "P" é o intervalo da predição de valores. Desses gráficos, pode-se ver que 58.3% das sequências contêm observações dentro do intervalo de confiança, enquanto o resto das sequências tem apenas um valor fora desse intervalo.

Analisando a Figura 5, que representa a função de densidade de probabilidade de Spearman, a maior parte de valores se encontra no intervalo [0.53-0.75],

obtendo um valor de Spearman médio de 0.682. Embora o número de amostras não seja alto (7 observações), ele é suficiente para afirmar que o valor obtido fica muito próximo do valor que permite rejeitar a hipótese nula e aceitar que existe uma forte correlação entre as sequências geradas pelo mecanismo e as sequências montadas pelos avaliadores.

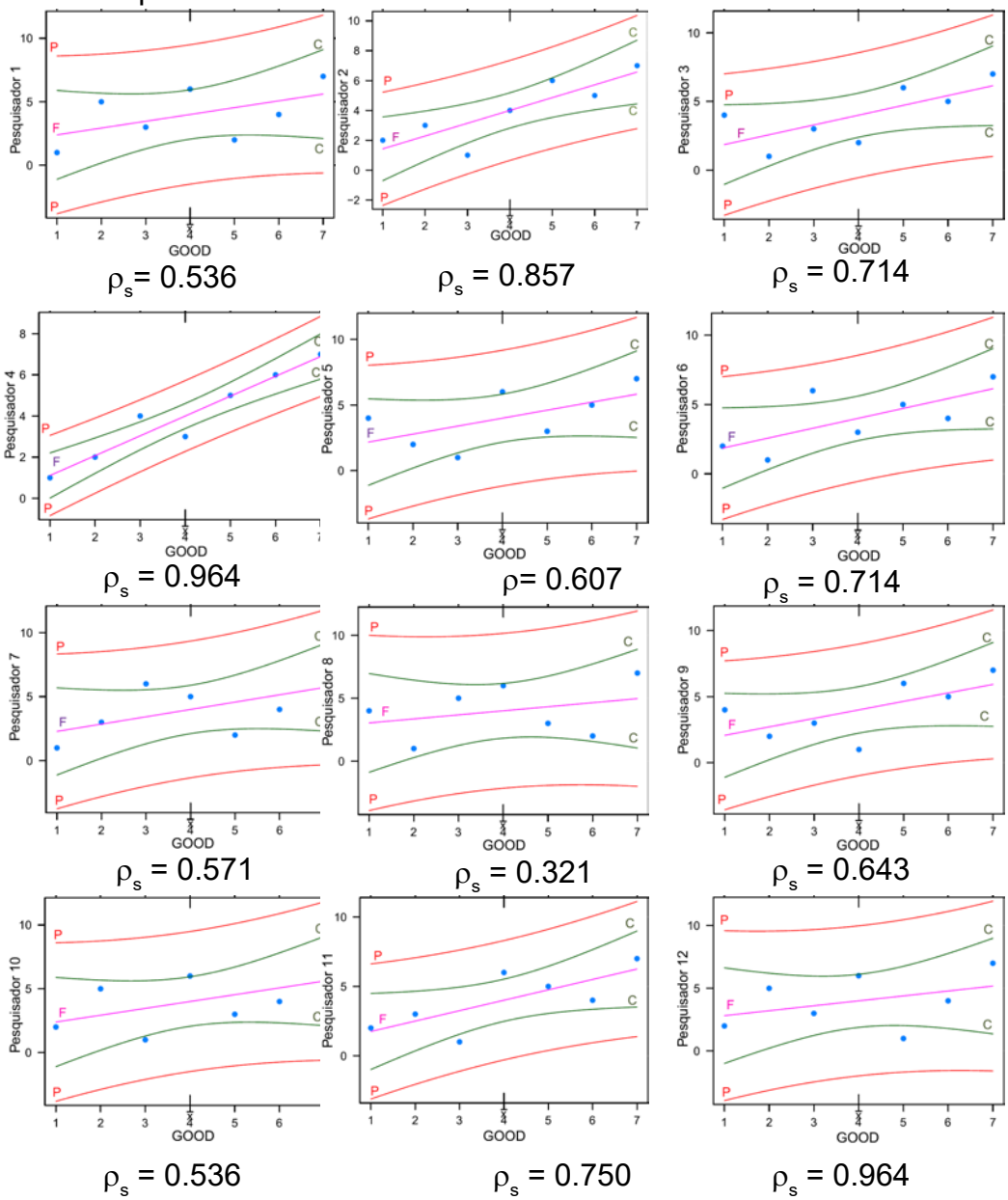

Figura 4. Sequências geradas pelo mecanismo e pelo grupo avaliador

Na Tabela 3, é analisada a complexidade e o coeficiente de Spearman médio alcançado pelo mecanismo de encadeamento proposto em comparação aos mecanismos desenvolvidos em outras pesquisas.

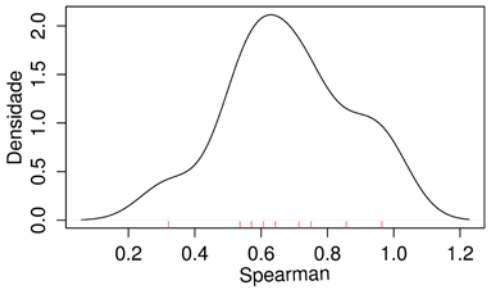

Figura 5. *Densidade de probabilidade dos coeficientes de Spearman*

Um fato a ser destacado na Tabela 3 é que na complexidade do mecanismo proposto o valor de *M* corresponde aproximadamente a ln(N), sendo inferior aos obtidos nos trabalhos relacionados.

Tabela 3. *Comparação da Complexidade vs. Spearman médio de vários trabalhos*

Autor	Complexidade	Spearman Médio
(RODRIGUES, 2010)	$O(N^3)$	0.400
(SHAHAF, 2010)	$O(N^2)$	Sem dados
(CAVALCANTE, 2013)	$O(N^2)$	0.784
O mecanismo proposto	$O(N \times M) \approx O(N \ln(N))$	0.682

CONCLUSÕES

A utilização de várias técnicas de mineração de dados diminuíram significativamente a complexidade do problema de ordenação de documentos científicos. Entre as técnicas utilizadas merecem especial destaque os algoritmos de clusterização, pois eles permitiram encontrar um contexto comum para a análise de implicação textual e análise temporal. Embora tenha sido utilizado o *Bisecting k-means*, que não gera clusters superpostos, ainda é necessário verificar os resultados ao se empregar outros algoritmos de clusterização. As regras de associação garantiram um mínimo de qualidade na relação entre pares de documentos, enquanto que o mecanismo de encadeamento garantiu a diversidade de artigos na leitura, ampliando o espaço de busca.

De acordo com os resultados alcançados, é possível encadear artigos de maneira coerente sem calcular a relação de coerência entre todos os documentos de uma coleção. Contudo, trabalhos futuros devem ser conduzidos de modo a: (i) propor uma alternativa para buscar padrões de coerência que permitam selecionar a melhor sequência dentre o espaço de solução apresentado pelo framework, (ii) verificar os resultados do encadeamento usando outros algoritmos de clusterização, (iii) incluir na definição de coerência a ponderação de pesos e/ou outros critérios, e (iv) estabelecer um corpus de teste para este tipo de aplicações.

REFERÊNCIAS BIBLIOGRÁFICAS

- *BARZILAY, R.; Lapata, M. Modeling local coherence: An entity-based approach; Computational Linguistics; v.34, n. 1, p.1-34,* **2008**.
- *BARZILAY, R.; Lee, L.; Catching the drift: Probabilistic content models, with applications to generation and summarization. arXiv preprint cs/0405039,* **2004**.
- *BARZILAY, R.; Elhadad, N. Inferring Strategies for Sentence Ordering in Multidocument News Summarization. arXiv preprint arXiv:1106.1820, v. 17, p. 35–55,* **2011**.
- *BASU, C.; Cohen, W. Technical paper recommendation: A study in combining multiple information sources. arXiv preprint arXiv:1106.0248, v. 1, p. 231–252,* **2011**.
- *CAVALCANTE, P.; Pinheiro, W. Mecanismo de Encadeamento de Notícias por Reconhecimento de Implicação Textual; sbbd2013.cin.ufpe.br.* **2013**, *p. 1–6,* ***2013***.
- *JANRUANG, J.; Guha, S. Semantic suffix tree clustering. Em: First IRAST International Conference on Data Engineering and Internet Technology; DEIT* **2011**.
- *KARAMANIS, N. Evaluating centering for information ordering using corpora. Computational Linguistics; v.35, n.1, p.29-46,* **2009**.
- *KARAMANIS, N. Evaluating centering-based metrics of coherence for text structuring using a reliably annotated corpus; Em: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, p.391,* **2004**.
- *KHABSA, M.; Giles, C. L. The number of scholarly documents on the public web. PloS one; v. 9, n. 5, p. e93949,* **2014**.
- *LAPATA, M.; Lascarides, A. A Probabilistic Account of Logical Metonymy. Computational Linguistics, v. 29, n. 2, p. 261–315,* **2003**.
- *MCCANDLES, M. Hatcher, E. and Gospodnetic, O.; Lucene in action: Covers Apache Lucene 3.0. Manning Publications Co.,* **2004**.
- *MACMAHON, P.A. Em Combinatory Analysis, American Mathematical Soc. 2001, v.137.*
- *Myers, J.; Well, A.; Lorch, R; Research design and statistical analysis. Routledge,* ***2010***.
- *NAVIGLI, R.; Crisafulli, G. Inducing word senses to improve web search result clustering. Em: Proceedings of the 2010 conference on empirical methods in natural language processing. Association for Computational Linguistics. p. 116-126,* **2010**.
- *ORTEGA, J.L; Aguillo, I. Microsoft academic search and Google scholar citations: Comparative analysis of author profiles. Journal of the Association for Information Science and Technology, v.65, n.6, p.1149-1156 ,* **2014**.
- *OSINSKI, S.; Weiss, D. A concept-driven algorithm for clustering search results.* **I***ntelligent Systems, IEEE, v.20, n.3, p.48-54,* **2005**.
- *OSINSKI, S.; Weiss, D. Clustering Search Results with Carrot2.* **2007**, .
- *PARK, D. et al. A literature review and classification of recommender systems research. Expert Systems with Applications; v. 39, n. 11, p. 10059–10072,* **2012**.
- *RODRIGUES, T.; Pinheiro, W. Relacionando Notícias Web: Uma Abordagem Causal e Temporal. Em: 9th International Information and Telecomunication Technologies Symposium,* **2010**.
- *SARWAR, B. et al. Analysis of recommendation algorithms for e-commerce. Em: Proceedings of the 2nd ACM conference on Electronica commerce. ACM, v. 5, n. 1/2, p. 158–167,* **2000**.
- *SARWAR, B. et al. Item-based collaborative filtering recommendation algorithms. Em: Proceedings of the 10th international conference on World Wide Web. ACM, p.285-295,* **2001**.

- *SHAHAF, D.; Guestrin, C. Connecting the dots between news articles. Em: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, p.623-632,* **2010**.
- *STORRER, A. Coherence in text and hypertext. Document Design,* **2002,** *v. 3, n. 2, p. 156–168.*
Tricoire, F. et al. Heuristics for the multi-period orienteering problem with multiple time windows. Computers & Operations Research. v.37, n.2, p.351-367, **2010**.