Finding Related Articles by a Bibliometric Approach

Anderson Poltronieri
Elias Oliveira
Departamento de Ciências da Informação
Universidade Federal do Espírito Santo
Campus de Goiabeiras, Av. Fernando Ferrari, s/n,
Cx Postal 5011, 29060-970 – Vitória, ES.
blue http://www.inf..ufes.br/~elias
elias@inf.ufes.br,anderson@tradusa.com.br

Abstract

The increasing amount of electronic documents and the diversity of scientific areas are turning the traditional manual work in libraries of clustering documents an impratical job. Hence, an automatic solution is needed in order to keep the same pace as the new documents are generated. The increase of the digital library services, in particular in Brazil, is an example of this growth. Nevertheless, when looking at the provided metadata struture offered by some of the management content software used in this area, we notice that important features are negleted. In this paper we are going to present a factorial analysis on the bibliographical references of a set of articles. By this analysis we are going to show a novel way of creating a web of related subject papers. Therefore, this paper presents a methotology for scientific document clustering based on their citations. This methodology is based on the decomposition of the citation *vs.* document matrix by the Singular Values Decomposition. By this decomposition we calculate the degree of similarity between any two documents grouping them by their similarities with respect to their citations. We carried out an experiment in order to validate our assumptions. The experiments was performed on a set of 70 articles from an *on-line* Brazilian journal: Datagramazero. The results showed that this methodology can well be used as an extra tool for a digital library for locating, searching and classifying documents, as we could create a graph of the most related papers of this journal only by looking at their citations frequencies.

Palavras-chave: Classificação automática de documentos; Recuperação da informação; Extração de Semântica Latente.

1 Introdução

Todos nós ao lermos um artigo, tese ou qualquer outro documento científico algumas vezes nos perguntamos: *Seria este documento uma cópia?* , ou talvez, *Quem escreveu algo parecido?* , *Quem são os autores mais importantes neste assunto?* , ou mesmo, *Em que classe de documentos eu coloco este novo documento que chegou às minhas mãos?* .

Com o grande aumento de publicações, sejam elas no mundo digital ou mesmo nos acervos de bibliotecas, estas dúvidas têm surgido com maior freqüência, visto a dificuldade, nas ferramentas de recuperação de documentos, de se obter algo relevante aos nossos anseios de busca. No entanto, assim como crescem as publicações digitais, também cresce o número de estudiosos que têm pesquisado novos algoritmos, metodologias e técnicas que permitam, particularmente no mundo digital, um melhor tratamento da verificação de similaridades de documentos, categorização e recuperação dos mesmos com mais rapidez e precisão.

No contexto dos documentos digitais, vários modelos para indexação e classificação foram propostos na literatura. Para citar apenas alguns: Redes Bayesianas [1]; Modelo Vetorial de representação da informação, com sua extensão para a *Latent Semantic Indexing* (LSI) [2] e Redes Neurais Artificiais [3], dentre outras técnicas de tratamento da informação em meio digital.

Diferentemente de outros procedimentos, como os acima citados, que envolvem, em princípio, a análise do conteúdo dos documentos, a nossa proposta metodológica é independente do conteúdo do documento. A análise por nós proposta é feita apenas com a utilização de dois tipos de informação: (1) as referências bibliográficas existentes ao final de cada documento científico e (2) a freqüência de citações destes itens de referência manifestas ao longo do documento.

Este trabalho está organizado da seguinte forma: na Seção 2 apresentamos uma breve discussão sobre as motivações cognitivas pelas quais um autor vem a citar outro em seu trabalho. Na Seção 3, apresentamos a estrutura de dados, por nós utilizada, para posterior processamento da informação intrínseca existente na freqüência de citações nos documentos citantes. O modelo estatístico aqui utilizado para extrair a *semântica* existente na relação frequência de citações e documentos citantes é apresentada na Seção 4. A seguir, na Seção 5, descrevemos como automaticamente classificamos os documentos baseados apenas em suas freqüências de citações. Como forma de avaliação de nosso procedimento, o resultado obtido pela nossa metodologia é comparada com outra [4] onde a técnica utilizada envolveu a análise dos conteúdos dos documentos envolvidos. Por fim, na Seção 6 apresentamos nossas conclusões e apontamos alguns possíveis caminhos de continuações deste trabalho.

2 Influência Intrínseca de Citações

Visto que o foco de nosso trabalho está na análise das referências bibliográficas citadas em um documento, então como podemos utilizar tal informação para avaliar similaridades entre documentos?

Para identificarmos a relação entre dois documentos utilizando referências e frequência de citações, nos basearemos na *hipótese* da *Influência Intrínseca de Citações* [5]. Portanto, partimos do princípio de que muito provavelmente, se em alguma parte de um documento encontramos uma citação a um outro, o documento que cita aborda um assunto relacionado com o tratado no documento citado, ainda que as opiniões sobre o assunto sejam divergentes. Logo, quanto maior o número de vezes que um documento fizer menção a um outro, podemos dizer que maior será a coincidência dos assuntos tratados em ambos. Outra forma de ver isso é pensarmos também que maior será a similaridade de áreas do conhecimento às quais pertençam os documentos, ainda que um e outro possam conter suas variações temáticas. Através das referências bibliográficas, segundo Nazer [6], podemos também identificar laços intelectuais entre autores, linhas de pesquisa e até mesmo laços sociais. Tentando extrair estas características embutidas nas referências e citações, a técnica utilizada consiste na contagem das citações e co-citações existentes nos documentos analisados [5]. Através desta contagem nos será possível identificar o grau de influência que uma referência bibliográfica, e conseqüentemente de seu autor, exerce em um documento no qual seu trabalho é citado. Desta forma, o primeiro passo a ser tomado para a identificação das similaridades é a construção de uma matriz onde cada coluna desta matriz representará um documento e cada linha desta matriz conterá o número de vezes que um trabalho científico, ou livro, é citado em um documento. Este número será ponderado por um fator relativo a freqüência com que trabalho científico também apareça nos outros documentos analisados. Este fator será maior se o trabalho científico aparecer em outros poucos documentos e será menor se ocorrer com freqüência na maioria dos outros documentos.

Esta forma de representar nossos documentos, tal como acima descrito, é semelhante àquela utilizada por pesquisadores da área de recuperação da informação quando utilizando o conhecido modelo vetorial [7]. No presente trabalho estamos trocando as freqüências dos termos pela freqüência das (*co*)-citações. Na seção seguinte fazemos uma explanação mais concreta desta forma de representação aqui discutida.

3 Relacionando Citações *vs.* Documentos

A construção da matriz documentos *vs.* referências citadas é o primeiro passo da técnica proposta neste trabalho. Como já dito anteriormente, este processo consiste na contagem das citações e co-citações existentes no interior de cada documento analisado [5]. No nosso caso, fizemos esta contagem de forma manual de quantas vezes uma referência bibliográfica foi citada em um

documento, por exemplo d_1, d_2, ..., d_n. Na coluna representativa de um documento, digamos d_4, a primeira linha representa a freqüência com que a referência R1 aparece no texto. Já a segunda linha deste mesmo documento nos mostra que a referência R2 não aparece neste documentos, muito embora apareça 5 vezes no documento d_6. Lembrando que este valor de freqüência já sofreu uma atenuação em decorrência também da freqüência com que a mesma referência aparece nos outros documentos sendo, portanto, um valor nominal de freqüência. Desta forma, a Tabela 1 nos mostra um exemplo desta montagem da matriz para seis documentos exemplos.

Trabalho	Documentos					
Citado	d_1	d_2	d_3	d_4	d_5	d_6
R1	3	0	4	0	1	4
R2	0	4	0	3	0	5
R3	3	3	3	0	5	3
R4	2	1	0	3	2	2
R5	1	3	3	4	0	2
R6	0	0	1	1	0	2
R7	6	0	9	3	0	1

Table 1: Representação vetorial dos documentos. Matriz **Documento** *vs.* **Referências**.

Note que, com esta estratégia de representação, nós transformamos nosso problema de manipulação de documentos textuais em um outro problema matemático. O trabalho agora é identificar similaridades entre estes vetores: documentos. Para isso nós lançaremos mão de um ferramental já amplamente utilizado na Estatística, chamado de Análise de Componentes Principais [8]. Na próxima seção descreveremos como utilizamos esta técnica para extrair uma certa *semântica* entre os documentos por nós analisados.

4 Modelo para Extração de Semântica

A Análise de Componentes Principais tem por objetivo descrever dados contidos em quadros *indivíduos–caracteres numéricos* de forma mais concisa [8].

Este tipo de análise é muito útil para um conjunto de dados, com mais de três caracteres, onde estes caracteres apresentam uma interligação intrínseca não claramente explícita. Isto porque, com até três caracteres, nós conseguimos facilmente construir uma representação gráfica para um exame visual. Mas quando este número de caracteres cresce, o exame visual se torna impossível.

Como já mencionado na Seção 3, os documentos representados na Tabela 1 existem em um espaço geométrico 7-dimensional de referências bibliográficas. No caso mais geral diríamos *n*-dimensional, onde *n* é o número de referências bibliográficas consideradas na análise. Neste trabalho adotaremos o *coseno* do ângulo entre os vetores (documentos) como forma de avaliarmos a similaridade entre os mesmos. Entretanto, quando projetamos os vetores deste espaço em um outro espaço possivelmente mais reduzido devido a correlação existente entre seus caracteres, o que obtemos é uma distorção da imagem dos vetores originais dado a projeção destes neste novo espaço. Ou seja, obteremos uma imagem transformada assim como *o efeito da sombra* de um objeto projetada sobre a parede. Este efeito faz com que vetores que antes não apresentavam diretamente nenhuma relação com alguma característica, no nosso caso, alguma referência, passam a apresentar tal relação. Portanto, a construção de novos indivíduos *sintéticos* é feita pela combinação dos caracteres iniciais por meio dos *fatores*. Estes fatores aparecem, portanto, por um método de combinações lineares entre as referências descritivas de um documento [8]. O cálculo matricial [9] nos provê o ferramental de que necessitamos para a projeção de vetores tais como, por exemplo, os apresentados na Tabela 1 em outro espaço, onde uma *semântica implícita*, a *sombra dos vetores originais*, possa ser tornada explícita através da projeção do espaço de representação dos vetores originais em um outro plano. Vale salientar que a escolha deste plano de projeção não é das mais triviais tarefas em problemas como o nosso [2]. A exata proporção de redução do espaço inicial para que apresente as características desejadas de extração é um problema ainda em aberto na literatura e que é tipicamente resolvido de forma empírica [4, 2]. Por isso, adotaremos a mais simples das alternativas, a qual consiste em simplesmente eliminar os indivíduos linearmente dependentes na transformação de um espaço para o outro. Para maiores detalhes o leitor interessado por consultar [9].

Por exemplo, utilizando-nos de uma técnica de decomposição de matrizes, a *Singular Value Decomposition* [9], podemos a partir da matriz apresentada na Tabela 1, gerarmos uma nova matriz apresentada na Tabela 2.

Trabalho Citado	Documentos					
	d_1	d_2	d_3	d_4	d_5	d_6
R1	3,14	1,03	3,91	0,91	2,00	2,19
R2	-0,04	3,91	-0,11	3,68	0,64	4,42
R3	3,23	2,24	2,76	-0,16	4,63	3,86
R4	0,98	1,92	1,06	1,55	1,15	2,53
R5	1,43	2,31	2,66	3,75	-0,51	2,88
R6	0,50	0,87	0,79	1,12	0,10	1,11
R7	5,82	-0,19	9,11	2,90	-0,07	1,27

Table 2: Documentos projetados de seu plano original para um novo plano semântico.

Neste novo plano semântico, documentos que antes tinham valor nulo em uma característica, como por exemplo o documento d_1 com respeito a referência R2, passam a apontar uma certa relação com esta característica. Observe que ainda pouco expressiva. Já outros documentos aumentaram muito sua expressão de co-relação característica, como é o caso do documento d_5 com respeito a referência R1. Enquanto que o próprio documento d_5 teve uma atenuação com respeito a característica R3.

5 Categorizando Documentos Automaticamente

Para validarmos nossa proposta, utilizamos uma amostra de 70 artigos da revista DataGramaZero (`http://www.dgz.org.br`) a fim de ser submetida ao processo automático de categorização de documentos pela análise apenas de suas citações e co-citações.

Uma vez tendo as citações e co-citações de cada artigo devidamente contadas, veja Seção 3, submetemos nossa matriz de citações *vs.* documentos aos processos brevemente descritos na Seção 4. Os artigos utilizados da revista foram os apresentados na Tabela 3. Onde *Ord*, corresponde a ordem de apresentação dos artigos da revista neste trabalho. *Mês-Ano/SeqRevista* diz respeito ao mês e ano em que o artigo foi publicado na revista e *SeqRevista* a seqüência interna de publicação do artigo na revista. Desta forma, o 46 artigo utilizado para experimento, o `abr04/03`, traduz-se pelo terceito artigo da revista de Abril de 2004.

Ord.	Mês-Ano/ SeqRevista	Ord.	Mês-Ano/ SeqRevista	Ord.	Mês-Ano/ SeqRevista	Ord.	Mês-Ano/ SeqRevista
1	ago01/01	18	abr00/02	35	out00/01	53	ago04/01
2	ago01/03	19	abr00/03	36	jun01/05	54	abr04/04
3	dez00/02	20	abr01/01	37	jun01/04	55	fev04/04
4	dez99/05	21	abr01/02	38	abr02/02	56	fev04/03
5	dez99/01	22	abr01/03	39	abr02/03	57	fev04/02
6	dez99/03	23	jun00/01	40	abr03/01	58	fev04/05
7	dez01/05	24	abr01/04	41	abr03/02	59	ago04/04
8	dez01/04	25	ago00/01	42	abr03/03	60	ago04/05
9	dez00/03	26	dez00/04	43	abr03/04	61	dez04/04
10	fev01/01	27	jun00/02	44	abr04/01	62	dez02/05
11	dez01/03	28	jun00/04	45	abr04/02	63	dez04/02
12	ago01/04	29	fev00/03	46	abr04/03	64	dez03/06

13	out01/04	30	fev01/02	47	ago02/01	65	jun03/04
14	out01/03	31	fev00/01	48	ago02/03	66	jun04/02
15	out01/02	32	jun00/03	49	ago02/04	67	out02/06
16	out01/01	33	jun01/01	50	ago03/01	68	out02/02
17	abr00/01	34	out00/02	51	ago03/04	69	jun04/05
				52	ago03/05	70	out03/01

Table 3: Artigos da DatagramaZero utilizados para os experimentos.

A categorização dos 70 artigos aqui analisados resultou nos agrupamentos apresentados na Tabela 4. Em nossos experimentos não permitimos a sobre-posição entre grupos, ou seja, não permitimos que um artigo possa pertencer a mais de um grupo.

Categorias	Artigos
A	1 *2 3* 4 **5 8** 18 21 *22* 24 *26* **28 29 31** 32 33 **35** 37 39 40 46 48 56 57 58 64 67
B	6 **19 20 36** 54 65 69
C	9 10 **11** 14 **16 17 34** 38 42 44 49 50 52 60 63 66 68 70
D	15 27 41
E	30 53
D	43 51 55 61

Tabela 4: Categorias geradas pelo nosso procedimento.

Nosso processo de classificação agrupou os documentos em 6 categorias. Por serem documentos da área de Ciência da informação, e por conseguinte, terem vários documentos de referência de importância global nesta área, podemos perceber que documentos de várias edições são agrupadas numa mesma categoria. Como trabalhamos com referência bibliográfica, referências como "*CASTELLS, M. A Sociedade em Rede, São Paulo: Paz e Terra, 1999."* e "*LEVY, P., cibercultura, São Paulo, Editora 34, 1999" s*ão citadas em vários artigos de edições diferentes. Como estes artigos outros também aparecem fazendo que possamos criar os vínculos de ligação entre os documentos.

Comparamos nossos resultados de nossa metodologia com uma outra baseada na análise de termos do documento [4]. Usamos este trabalho como referência, pois, dos 70 artigos classificados, 36 encontram-se no trabalho no referenciado. Desta comparação podemos observar semelhanças com alguns grupos criados por Ramiro[4]. Em [4] encontramos um grupo *E* com os seguintes documentos (02, 03, 22,26). Este grupo está completamente agrupado em nosso resultado. No entanto ela se encontra em nosso trabalho no grupo A, itens em destaque e sublinhado na tabela 4, juntamente com outros documentos. Também podemos observar que o grupo A do trabalho de Ramiro[4], itens em

negrito na tabela 4, foi dividido em em 3 subgrupos. Interpretamos tal divisão do grupo pelo fato de que, embora da mesma área, conforme definido por Ramiro[4], tais documentos possuem referências bibliográficas distintas o que provoca esta divisão em nosso trabalho, visto que nos baseamos apenas pelas citações e referências.

Da mesma forma, outros documentos que se diziam de grupos distintos no trabalho[4], em nosso trabalho aparecem juntos, como é o caso dos documentos 9 e 10. Observando os termos descritores das categorias criados por [4], podemos perceber que estes documentos tratam de assuntos muito próximos e por este motivo possuem grande probabilidade citarem documentos em comum. De fato ambos os artigos citam a referência *"CASTELLS, M. - La era de la informacion: Economia, Sociedade, y cultura... 1997, Vol1."* , sendo citado 2 vezes pelo documento 9 e 8 vezes o documento 10. Tais citações, alidadas à *Influência Intrínseca* de outras citações, obtida como mostrado no itens 2 e 3 deste artigo, nos permite dizer que estes documentos estão numa mesma área, e de fato estão muito próximos se levarmos em conta os descritores das categorias definidos por Ramiro[4].

Também pudemos observar que alguns documentos não foram classificados em nosso trabalho. Tal fato se deve a distinção das citações destes documentos dos demais documentos analisados, como é o caso dos documento 7 e 12.

De posse dos resultados obtidos, podemos dizer que a metodologia adotada nos fornece grupos de documentos que podemos afirmar ser de áreas afins, ainda que com pequenas variações de subtemas. No entanto, usada sozinha a técnica pode isolar documentos que possuam referências muito distintas das utilizadas pela maioria dos autores.

A técnica também pode criar grupos dentro de uma mesma área de atuação de acordo com as escolas mais influêntes para os autores dos documentos de uma mesma categoria, como é o caso dos documentos da categoria A de [4] que se dividiu em 3 grupos no nosso trabalho. Nestes casos apenas uma ferramenta de análise semântica de todo o conteúdo poderia dizer que os documentos ditos de grupos distintos pela metodologia proposta pertencem ao mesmo grupo. Então de posse dos dois resultados, o semântico e o refencial, poderíamos criar uma categoria com ambos os grupos, se for o caso, usando os dados de referências para criar subcategorias.

6 Conclusões

Neste trabalho apresentamos uma metodologia de agrupamento de documentos científicos através somente da análise da freqüência de citações e co-citações existentes no texto. A utilização apenas das citações e co-citações

nos diferencia das técnicas convencionais, as quais envolvem a análise de todo o conteúdo do texto documental.

Nós utilizamos como forma de validação de nossa proposição um conjunto de artigos disponíveis *on-line* através da revista Datagramazero. Os agrupamos de acordo com nossa metodologia e comparamos nossos resultados com aqueles produzidos em [4]. Neste último trabalho, os agrupamentos foram feitos através da extração de todos os termos existentes nos documentos. A comparação dos resultados no mostrou que nossa metodologia pode servir como uma boa ferramenta a mais de apoio ao difícil problema de recuperação de documentos similares em uma biblioteca digital.

Referências

[1] Neapolitan RichardE. Learning Bayesian Networks. New Jersey, USA: Pearson & Prentice-Hall; 2004.

[2] Deerwester ScottC, Dumais SusanT, Landauer ThomasK, Furnas GeorgeW, Harshman RichardA. Indexing by Latent Semantic Analysis. Journal of the American Society of Information Science 1990;41(6):391–407.

[3] Haykin S. Neural Networks – A Comprehensive Foundation. Pearson Education; 1998.

[4] Ramiro ThiagoBortolo, Monteiro Valéria, Azevedo LíviaLopes, Teixeira Sergio, Oliveira Elias. Atribuindo Títulos de Assuntos na Categorização Automática de Documento. In: XXI Congresso Brasileiro de Biblioteconomia, Documentação e Ciência da Informação. Curitiba; 2005. .

[5] Egghe L, Rousseau R. Introduction to Informetrics: Quantitative Methods in Library, Documentation and Information Science. Amsterdam: Elsevier Science; 1990.

[6] White H, Wellman B, Nazer Nancy. Does Citation Reflect Social Structure? Journal of the American Society for Information Science and Technology 2004;55(2):111–126.

[7] Baeza-Yates R, Ribiero-Neto B. Modern Information Retrieval. 1st ed. New York: Addison-Wesley; 1998.

[8] Johnson RichardA, Wichern DeanW. Applied Multivariate Statistical Anaysis. New Jersey: Prentice Hall; 1992.

[9] Golub GeneH, Loan CharlesFVan. Matrix Computations. 3rd ed. Baltimore, MD: Johns Hopkins University Press; 1996.