Comparação do processo de categorização de documentos utilizando palavras-chave e citações em um domínio de conhecimento restrito

Comparison of document categorization process using key words and citations in a restricted knowledge domain

Magali Rezende Gouvêa MEIRELES[1]
Beatriz Valadares CENDÓN[2]
Paulo Eduardo Maciel de ALMEIDA[3]

Resumo

O processo de categorização requer a extração de elementos representativos de um documento de modo que sua essência possa ser utilizada na identificação de similaridades e na geração de categorias. O objetivo deste trabalho é analisar as dificuldades e os resultados encontrados em dois processos diferentes de categorização de documentos de um domínio de conhecimento restrito, o primeiro, baseado no uso de palavras-chave e o segundo, na utilização de citações para representação dos documentos. Para exemplificar a utilização de diferentes atributos na representação dos documentos foram realizados dois experimentos. O primeiro utilizou um algoritmo de categorização baseado em palavras-chave. O segundo experimento gerou as categorias, utilizando Redes Neurais Artificiais, a partir das citações dos artigos. Em um domínio de conhecimento restrito, como o utilizado neste trabalho, foi possível evidenciar a dificuldade na formação de grupos que utilizam palavras-chave como atributo do processo de categorização devido à grande similaridade existente entre aquelas utilizadas pelos autores dos artigos. As citações podem ser, como mostrado no segundo experimento, um atributo alternativo e mais eficiente para esse processo. A formação de um grupo de artigos com um expressivo acoplamento bibliográfico e uma forte relação semântica comprovou a validade do método proposto. A presente pesquisa detalha a metodologia utilizada nos experimentos, mostrando a importância de uma criteriosa fase de pré-processamento para a confiabilidade das bases de dados. Este estudo pode contribuir com as pesquisas relacionadas à representação de documentos em processos de categorização e de recuperação de informação.

Palavras-chave: Análise de citação. Categorias. Inteligência artificial. Recuperação da informação. Representação da informação.

Abstract

The categorization process requires the extraction of representative elements from a document so that its essence can be used to identify similarities among documents and generate categories. The objective of this study was to analyze the difficulties and results from two

[1] Pontifícia Universidade Católica de Minas Gerais, Instituto de Ciências Exatas e Informática, Programa de Pós-Graduação em Informática. R. Walter Ianni, 255, São Gabriel, 31980-110, Belo Horizonte, MG, Brasil. Correspondência para/*Correspondence to*: M.R.G. MEIRELES. *E-mail*: <magali@pucminas.br>.
[2] Universidade Federal de Minas Gerais, Escola de Ciência da Informação, Programa de Pós-Graduação em Ciência da Informação. Belo Horizonte, MG, Brasil.
[3] Centro Federal de Educação Tecnológica de Minas Gerais, Departamento de Computação, Laboratório de Sistemas Inteligentes. Belo Horizonte, MG, Brasil.
Recebido em 26/3/2014, reapresentado em 25/2/2015 e aceito para publicação em 25/3/2015.

different processes of document categorization in a restricted knowledge domain. The first one was based on the use of keywords and the second was based on the use of citations for document representation. To illustrate the use of different attributes in document representation, two experiments were conducted. The first one used a categorization algorithm based on keywords. The second experiment generated categories, using Artificial Neural Networks, from the citations of the articles. In the restricted knowledge domain, as used in this study, it was difficult to form groups that use keywords as attributes of the categorization process due to the great similarity of keywords used by the authors. The citations can be, as shown in the second experiment, an alternative and more efficient attribute for the categorization process of these documents. The formation of a set of articles with significant bibliographic coupling and a strong semantic relationship validated the method proposed. The article details the methodology used in the experiments, showing the importance of careful pre-processing phase for the reliability of the databases. This study may contribute to the research related to the representation of documents in categorization processes and information retrieval.

Keywords: *Citation analysis. Categories. Artificial inteligence. Information retrieval. Information representation.*

Introdução

O processo de categorização é um processo natural do ser humano que busca, na criação de grupos, um modo de organizar as informações e simplificar suas interações com o mundo. Segundo Nicolelis (2011), a busca incessante do indivíduo por conhecimento é caracterizada por sua própria história evolutiva, por seu estado dinâmico global a cada momento no tempo e pelas representações internas que ele mantém do corpo e do mundo. Fazendo associações de tudo que lhe parece novo aos grupos de imagens e conceitos já estabelecidos, o indivíduo é capaz de reorganizar as estruturas já concebidas e entender com mais simplicidade o ambiente que o cerca.

Com a participação crescente da *Internet* na vida dos usuários dos sistemas computacionais, o interesse pelo desenvolvimento de ferramentas que pudessem organizar e recuperar uma coleção de documentos disponível em meio digital tornou-se cada vez maior. A categorização é a tarefa de agrupar entidades relacionadas; já o processo de classificação associa uma entidade a um dos grupos previamente definidos na categorização. As Redes Neurais Artificiais (RNA), que também podem ser utilizadas nesses processos, são modelos matemáticos inspirados nas estruturas neurais biológicas, as quais adquirem conhecimento por meio de um processo de aprendizado e utilizam as forças de conexão existentes entre os neurônios, conhecidas como pesos sinápticos, para armazená-lo (Haykin, 1994). A capacidade do cérebro de guardar conhecimento, reorganizando e alterando as forças de conexão entre os neurônios, tem inspirado pesquisadores a desenvolver e aprimorar o funcionamento das RNA.

Os processos de categorização podem utilizar diferentes atributos para identificar similaridades entre os documentos e a cada escolha de atributo estão associadas diferentes considerações. Em alguns domínios de conhecimento restrito a utilização de palavras-chave pode se tornar um obstáculo ao processo de geração de grupos, visto que muitos dos artigos apresentam palavras-chave semelhantes ou relacionadas, o que dificulta a identificação de características que possam diferenciar e criar os grupos. Este trabalho avalia, em um domínio de conhecimento restrito, as dificuldades encontradas e os resultados gerados por processos de categorização que utilizam palavras-chave e citações. A fase de pré-processamento dos dados utilizados é detalhada, ressaltando sua importância para a confiabilidade das análises posteriores.

A fundamentação teórica do trabalho, destaca as características e as dificuldades encontradas nos processos de categorização que usam palavras-chave e citações na representação dos documentos. O tópico procedimentos metodológicos, apresenta a metodologia utilizada, definindo o *corpus* empírico e ressaltando a fase de pré-processamento de dados e aquelas relacionadas aos experimentos práticos realizados. A discussão dos resultados e as conclusões são colocadas nas seções finais.

Toda informação precisa ser representada para, posteriormente, ser recuperada. O primeiro passo inclui a extração de alguns elementos do documento ou a atribuição de características ao mesmo de modo que sua essência possa ser capturada e apresentada ao usuário. Segundo Chu (2003), este é um processo que envolve desafio, complexidade e sutileza. A categorização é o processo utilizado para descrever a informação por meio de categorias.

Representação de documentos por meio de palavras

Segundo Croft *et al.* (2010), as configurações das ferramentas de busca refletem a aplicação para a qual foram designadas. Além do processo de busca, o processo de indexação desempenha um papel importante na construção da ferramenta. Este define a estrutura que será utilizada na busca, basicamente composta pela aquisição e processamento do texto e pela criação dos índices. Para composição desses índices, além de palavras, podem ser usadas frases ou expressões, nomes de pessoas, datas e *links* em uma página da *web* relacionados ao documento. Algumas palavras especiais são consideradas por alguns autores, em casos específicos, suficientes para distinguir certo documento de outro (Lensu & Koikkalainem, 1999). Na tentativa de capturar o contexto em que as palavras estão inseridas, a representação de documentos também pode ser realizada extraindo-se as principais ideias de tais documentos (Phuc & Hung, 2008) ou ainda as frases que possam substituir um conjunto de palavras (Bakus *et al*., 2002).

Kim *et al.* (2008) classificam os itens dos documentos a serem analisados como estruturados e não estruturados. Itens estruturados estão associados a uma determinada representação e apresentam o mesmo formato em um mesmo tipo de documento. Tomando-se o exemplo de uma patente, que é um documento rico em informações e muito explorado em processos de categorização e de classificação, os itens estruturados são campos como autores e afiliações, número da patente, data de aplicação, data de publicação e palavras-chave. Os itens não estruturados têm o formato de texto livre e são diferentes em tamanho e conteúdo. Em patentes, estes itens são campos de resumo, reivindicações e descrição da invenção. Em alguns estudos relacionados à classificação de patentes apenas uma das classes de itens é utilizada. Richter e MacFarlane (2005) descrevem a utilização de metadados para a classificação desse tipo de documento e utilizam, como atributos do processo, códigos de classificação, inventores e nomes dos autores. Kim *et al.* (2008) apresentam um mapa de patentes que considera ambos os aspectos para a classificação, avaliando as palavras-chave, que vêm de um item estruturado, e a frequência com que são apresentadas no texto, um campo não estruturado. O objetivo da análise do mapa de patentes é identificar as tecnologias emergentes e prever as possíveis áreas de desenvolvimento. Lee *et al.* (2009) utilizaram palavras-chave para criar os mapas de patentes e identificar as áreas com baixa densidade de invenções e, assim, definir potenciais campos para criação de novas tecnologias.

Os processos de categorização e de classificação são utilizados como ferramentas de apoio à recuperação de informação. Algumas pesquisas relatam a frustração dos usuários com a qualidade dos documentos recuperados e associam esse resultado a uma formulação ineficiente da pergunta de busca. Kajanan *et al*. (2014) propõem um método para fornecer consultas baseadas em textos fornecidos pelo usuário. Esse é um método que deve ser utilizado quando já se identificou um conjunto relevante de textos que podem ser usados na implementação do método. Segundo esses autores, existem evidências conclusivas as quais comprovam que, geralmente, falta efetividade na formulação da busca, resultando em repetidas reformulações da mesma. Algumas razões citadas estão relacionadas ao fato de que as primeiras pesquisas são, frequentemente, realizadas em domínios pouco conhecidos e, portanto, não se sabe quais termos melhor capturam a essência dos tópicos de interesse. Além disto, os pesquisadores tendem a utilizar termos de busca que são ou muito genéricos ou muito específicos.

A utilização das palavras como representação temática dos documentos pode apresentar problemas relacionados às propriedades linguísticas das mesmas (Smeaton, 1991; Kuramoto, 2002). Algumas das dificuldades encontradas são enumeradas a seguir:

- Polissemia: palavras que apresentam diferentes significados dependendo do domínio de conhecimento no qual estão inseridas;

- Sinonímia: duas palavras designando o mesmo significado;

- Combinação de palavras em ordem diferente: a ordem pode alterar o sentido da expressão.

Kuramoto (2002) ressaltou ainda que as palavras extraídas automaticamente podem substituir inadequadamente os descritores, porque, ao serem isoladas de um texto, deixam de ter o valor atribuído pelo autor ao redigir o documento. O descritor torna-se uma unidade da língua, um vocabulário de designação genérica, e

deixa de ser, como desejado, uma unidade do discurso relacionada à realidade do autor.

Representacão de documentos por meio de citações

Segundo Borgman e Furner (2002), a análise de citações é a abordagem bibliométrica mais conhecida entre os métodos matemáticos e estatísticos que estudam e analisam eventos relacionados a variáveis quantificáveis na produção escrita. Ela identifica a relação de um documento com outro independentemente da presença de termos iguais nos mesmos. Os resultados obtidos pela aplicação da análise de citações têm confirmado a importância do método para avaliar a produtividade e as tendências da literatura científica no tempo e no espaço físico, mapeando, em alguns casos (Andrade & Jung, 2013), a frente de pesquisa de específicas áreas de conhecimento.

Acoplamento bibliográfico e co-citação são exemplos de estudos na avaliação de similaridades de documentos como mostrado na Figura 1. O grau do acoplamento bibliográfico existente entre A e B é resultado da frequência com que os documentos são citados simultaneamente por A e B, como no caso de C. O foco da análise do estudo de co-citação está nos documentos citados e é resultado da frequência com que E e F são citados, por exemplo, por D (Lai & Wu, 2005). Uma diferença relevante entre os dois estudos é a temporalidade relativa. Os documentos cujo grau de acoplamento bibliográfico é avaliado são mais recentes que suas citações. Por outro lado, o foco do estudo de co-citação está nos citados e, portanto, anteriores àqueles que os citam.

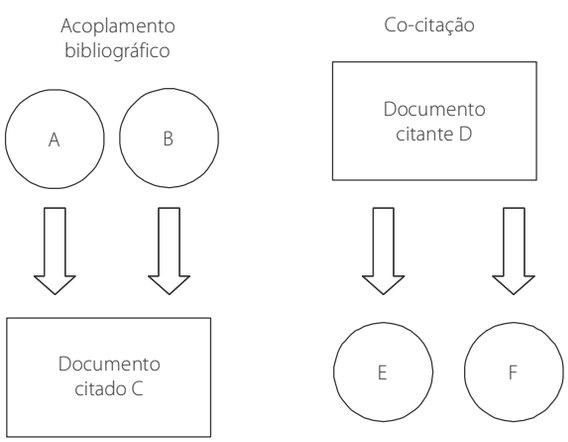

Figura 1. Exemplos de acoplamento bibliográfico e co-citação.
Fonte: Lai e Wu (2005).

De acordo com Hjorland (2008), o acoplamento bibliográfico foi introduzido por Kessler, em 1963, e a análise de co-citação foi sugerida, independentemente, por Marshakova e Small ambos em 1973. Kessler (1963), em seus experimentos, encontrou um alto grau de correlação semântica entre os documentos agrupados usando o critério de acoplamento bibliográfico. Marshakova (1973) entendeu que a análise por co-citação utiliza lógica oposta àquela usada pelo acoplamento bibliográfico. Small (1973) concluiu que a análise por co-citação gera uma ferramenta para monitorar o desenvolvimento de campos científicos e para avaliar o grau de relacionamento entre documentos específicos. Mesmo que muitos autores reconheçam as citações como fonte de informação para o estudo da inovação e da mudança tecnológica, é importante frisar que a análise das mesmas, utilizada na identificação de pesquisas emergentes, precisa considerar as diferenças de tempo que, muitas vezes, os documentos levam para ser citados.

Muitos aspectos devem ser levados em conta ao se optar pela classificação de documentos pelas citações. Macias-Chapula (1998) destaca que, para os sociólogos, as mesmas só podem ser compreendidas se as condições sociais que predispõem os cientistas a realizá-las forem analisadas. Alvarenga (1998) ressalta que as críticas à utilização de técnicas bibliométricas não se restringem à sua abordagem quantitativa, mas estendem-se, também, às suas vinculações com possíveis tendências de pesquisas consideradas legitimadoras de ideologias dominantes, em que se privilegiam os meios em detrimento dos fins. MacRoberts M.H. e MacRoberts B.R. (1989) relatam dificuldades, como o elevado número de autocitações em um documento e o fato de que citações afirmativas e negativas são tratadas sem distinção.

Considerando-se, novamente, o exemplo de patentes, é possível afirmar que pelo menos um tipo de relacionamento pode ser determinado pelo emprego de análise de citações. Se uma patente é muito citada, então, é, possivelmente, o fundamento para as citantes (Fujii *et al*., 2007). Muitos autores têm apresentado propostas de classificação de patentes utilizando co-citação (Liu & Shih, 2011). Outros têm reconhecido as citações como uma

fonte de informação para associar invenções e cientistas, por meio do acoplamento bibliográfico existente entre os documentos, para estudar o desenvolvimento tecnológico de uma área específica do conhecimento (Hall *et al*., 2002) e para estimar o valor econômico deste tipo de documento (Sapsalis *et al*., 2006).

Mesmo que as motivações dos autores para citar outros trabalhos possam ser diferentes de pessoa para pessoa, as citações estão relacionadas ao que os autores, nesse contexto, escolheram para representar as suas respectivas publicações. Elas podem, ainda, ser decompostas em títulos e autores, criando uma fonte rica para a representação do documento (Chu, 2003).

Procedimentos metodológicos

Para comparar a eficiência dos processos de categorização utilizando palavras-chave e citações, foram realizados dois experimentos com uma base de testes. Esta foi formada por 200 artigos de um domínio de conhecimento restrito publicados entre 2001 e 2010. Utilizou-se a busca avançada da Biblioteca Digital *IEEE Xplore* e os artigos selecionados continham a expressão "*neural network"* nos títulos. Foram escolhidos, a partir da definição do ano, os 20 primeiros artigos publicados no periódico *IEEE Transactions on Neural Networks* com acesso ao texto completo e cujas citações apresentassem todos os dados exigidos para o cadastramento da base relacionada. As referências dos artigos selecionados somaram um total de 6 015 e os artigos geraram um grupo de 661 palavras-chave selecionadas pelos próprios autores.

Duas bases de dados foram geradas: a primeira, contendo as informações relativas aos artigos e, a segunda, aquelas relativas às citações dos mesmos. Ao se utilizar, na representação dos documentos, atributos associados às palavras-chave ou às citações, é preciso lidar com as inconsistências encontradas no próprio campo das palavras-chave e nos títulos desses documentos, bem como nos títulos e autores das citações. A etapa de pré-processamento deste trabalho envolveu desde a identificação de dados incorretos, utilizando um programa na linguagem Java, até o ajuste manual dos dados, tomando-se como referência as informações contidas na Biblioteca digital *IEEE Xplore*.

O primeiro experimento utilizou o *software* Tanagra e as palavras-chave dos artigos da base de teste como atributo do processo de categorização. O arquivo utilizado na entrada do Tanagra possuía a informação da presença ou da ausência das 661 palavras-chave presentes nos 200 artigos escolhidos. Foram selecionados os processos que geravam de três a 20 categorias.

Para avaliar a frequência com que estas palavras predominantes nos grupos gerados pelo experimento eram encontradas nos campos correspondentes dos artigos da base de dados, foi utilizado um algoritmo baseado na análise de n-gramas. Uma sequência específica de *n* letras ou palavras em um documento é denominada n-grama (Croft *et al*., 2010). A frequência com que essas sequências são encontradas nos documentos tem sido objeto de análise em diferentes problemas e domínios, tais como sumarização de texto, a detecção de erros de ortografia e correção e a categorização automática de textos (Tomovic *et al*., 2006). A partir da frequência com que uma dada sequência de palavras é encontrada em um conjunto de documentos é possível, por exemplo, extrair expressões para representá-los e identificar, posteriormente, similaridade semântica entre eles.

O segundo experimento utilizou um método de categorização baseado em RNA e as citações destes 200 artigos como atributo do processo. A proposta era categorizá-los, agrupando-os pelo grau de acoplamento bibliográfico existente entre eles. Foi utilizado o *toolbox Neural Network*, do *software Matlab* (*Matrix Laboratory*), que emprega as redes *Self Organizing Maps* (SOM). Estas, utilizadas em processos de categorização (Braga *et al*., 2000), são estruturas não supervisionadas que possuem capacidade de auto-organização. São mais semelhantes às estruturas neurobiológicas do que aquelas que utilizam informações fornecidas por um supervisor, denominadas supervisionadas. A partir do instante em que a rede identifica regularidades entre os dados de entrada, ela desenvolve uma habilidade para gerar representações internas para codificar as características de entrada e criar novos grupos automaticamente.

Assim, a presença ou a ausência de cada uma das 6 015 citações dos 200 artigos selecionados para a base de testes foi o atributo binário utilizado pela RNA no processo de categorização (Meireles *et al*., 2014). Para gerar o arquivo de entrada da RNA, cada uma das citações

recebeu um código numérico durante o cadastramento da base de testes. Esse código, presente na representação do artigo, definia a presença da citação no mesmo. Para garantir que cada uma delas estivesse cadastrada com um único código numérico, mesmo que referenciada por artigos distintos, foi elaborado um programa na linguagem *Java* para identificar as duplicidades e, de modo apropriado, conciliar as referências em uma base de dados final para ser utilizada pela RNA (Meireles, 2012).

A categorização utilizando palavras-chave

Na maioria das topologias utilizadas, os artigos se concentraram em um único grupo, o que se justifica por se tratar de uma base de dados de um domínio específico de conhecimento. Foi selecionada para análise a topologia que gerava 13 grupos de artigos. Nela, a maioria dos 200 artigos foi distribuída em quatro grupos, denominados Grupo A, Grupo B, Grupo C e Grupo D. A Tabela 1 apresenta as características dos mesmos e as palavras-chave que predominavam nos artigos agrupados em cada um deles.

Como um dos critérios utilizados para a composição da base de dados era a presença da expressão "*neural network*" nos títulos dos artigos, grande parte deles apresentava a mesma como palavra-chave, o que justifica o maior número de artigos reunido no Grupo A. Porém, nem todos eles apresentavam essa expressão. Existem palavras comuns entre os artigos que os relacionam e, de algum modo, justificam a criação do grupo.

Para validar os resultados apresentados neste experimento, a palavra-chave foi utilizada como unidade de n-grama pelo algoritmo que apresentou a frequência com que as mesmas foram utilizadas pelos autores. A frequência com que as palavras-chave relacionadas às encontradas pelo *software* Tanagra são utilizadas pelos autores dos artigos da base de dados é apresentada na Tabela 2.

As palavras-chave predominantes nos grupos do experimento realizado com o *software* Tanagra são muito utilizadas pelos autores dos artigos, conforme comprova a última coluna da tabela. A análise de 2-gramas confirmou que, mesmo não estando presente em todos os textos, a expressão "*neural networks*", no plural, foi utilizada 118 vezes no campo de palavras-chave, e a "*neural network*", 88. Em alguns casos, essas duas palavras estavam associadas a expressões do tipo "*delayed neural networks*", "*recurrent neural network*", "*discontinuous neural networks*", "*graph neural networks*", "*recursive neural networks*", dentre outras. A grande frequência de apresentação da expressão no campo de palavras-chave justifica o fato de que os outros grupos gerados no experimento, significativamente menores, também eram caracterizados por palavras-chave derivadas dessa expressão principal.

A categorização realizada por RNA utilizando as citações dos artigos

Para gerar as categorias, a RNA utilizou as informações contidas na base de dados pré-processada. Depois de ajustar alguns parâmetros da RNA e variar sua topologia, que está associada ao número de neurônios da camada de saída, optou-se por definir saídas que gerassem 4, 9, 10, 12, 16, 25 e 36 categorias de artigos. O objetivo dessa variação era verificar se, alterando o número de grupos criados na saída da RNA, os mesmos documentos continuariam agrupados. Foram escolhidas as quatro topologias que apresentavam artigos concentrados no mesmo número de grupos. Aquelas que geravam 10, 12, 16 e 25 categorias concentravam a maioria dos artigos em 7 grupos. Essa opção foi realizada na expectativa de que, concentrando os documentos no

Tabela 1. Características dos quatro principais grupos gerados na topologia de 13 categorias.

Grupos	Artigos (n)	Palavras-chave predominantes	Comentários (%)
Grupo A	141	*Neural network*	35 dos artigos apresentava a palavra-chave predominante
Grupo B	8	*Global asymptotic stability*	100 dos artigos apresentava a palavra-chave predominante
Grupo C	35	*Recurrent neural network*	97 dos artigos apresentava a palavra-chave predominante
Grupo D	7	*Radial basis function neural network*	100 dos artigos apresentava a palavra-chave predominante

Fonte: Meireles (2012).

 http://dx.doi.org/10.1590/2318-08892016002800007

Tabela 2. Frequência com que as palavras-chave foram encontradas nos artigos.

Palavras-chave predominantes nos quatro principais grupos gerados pelo *software* de categorização	Palavras-chave dos artigos da base de testes	Frequência com que são utilizadas pelos artigos (n)	Frequência das palavras-chave relacionadas
Neural network	*Artificial neural network (ANN)*	1	60
	Artificial neural network	1	
	Artificial neural networks	3	
	Neural network	19	
	Neural networks	27	
	Neural networks (NN)	2	
	Neural networks (NNs)	7	
Global asymptotic stability	*Asymptotic stability*	4	27
	Asymptotical stability	2	
	Global asymptotic stability	6	
	Global asymptotical stability	2	
	Stability	10	
	Stability analysis	3	
Recurrent neural network	*Recurrent networks*	2	39
	Recurrent neural network	9	
	Recurrent neural network (RNN)	9	
	Recurrent neural networks	13	
	Recurrent neural networks (RNNs)	6	
Radial basis function neural network	*Radial basis function*	1	9
	Radial basis function (RBF) network	1	
	Radial basis function networks	1	
	Radial basis function neural network	1	
	Radial basis function neural network (RBFNN)	3	
	Radial basis function (RBF) neural networks	1	
	Radial basis function neural networks (RBFNNs)	1	

Fonte: Elaborado pelos autores (2014).

mesmo número de grupos, a RNA estivesse mantendo, nesses grupos, os mesmos documentos. A Figura 2 sumariza o método. As topologias associadas ao número de categorias geradas são designadas pelas letras TP na Figura 2. O termo genérico TPxx representa as topologias TP12, TP16 e TP25, também, utilizadas no experimento.

A análise dos resultados gerados pelo método encontrou, em um dos grupos criados pela RNA no conjunto final de categorias, seis artigos com um grande número de citações em comum. Cada dois dos seis artigos possuíam de 7 a 21 citações em comum, apresentando um alto nível de acoplamento bibliográfico.

Resultados e Discussão

Como a escolha da base de dados deste trabalho restringiu o domínio de pesquisa - limitando a seleção dos artigos de um periódico especializado àqueles que apresentassem a expressão "*neural network*" em seus títulos -, o grande grupo encontrado pelo *software* Tanagra apresentava essa expressão como palavra-chave predominante. A análise de 2-gramas confirmou que as expressões "*neural networks*" e "*neural network*" somadas foram utilizadas mais de 200 vezes no campo de palavras-chave. Nem sempre a expressão foi utilizada pelos autores, mas surgiu, em muitos casos, associada a outras. Isso justifica a dificuldade do *software* em identificar diferenças no processo de categorização, criando, então, um grande grupo formado por 141 artigos. Por outro lado, a análise da frequência de ocorrência das palavras-chave predominantes nos grupos do primeiro experimento comprova que essas palavras são muito utilizadas pelos autores dos artigos da base de testes.

O segundo experimento, que utilizava a presença das citações dos artigos como atributo de categorização,

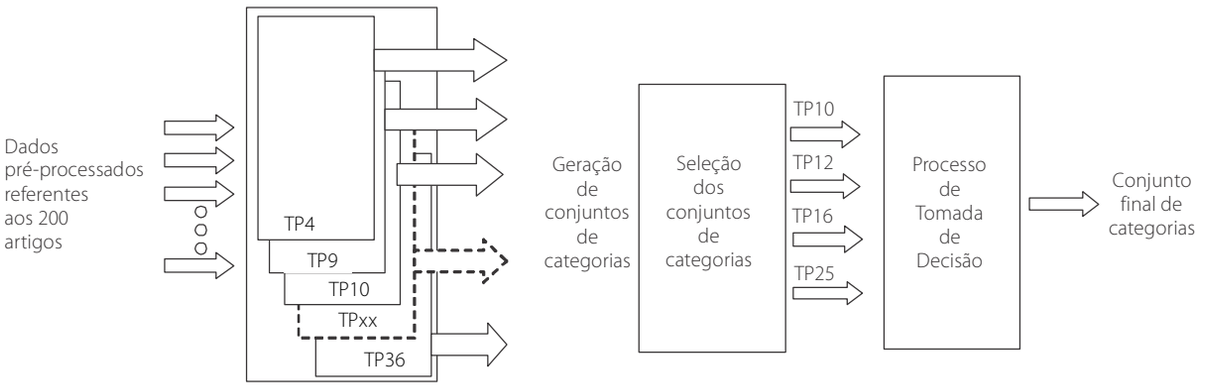

Figura 2. Representação do método de categorização utilizando citações e Redes Neurais Artificiais.

Fonte: Elaborado pelos autores (2014).

Nota: TP: Topologias; TPxx: Topologia genérica.

conseguiu identificar, dentre os grupos gerados pela RNA, um conjunto de seis artigos que, além de apresentar um significativo grau de acoplamento bibliográfico, também apresentava uma forte relação semântica, comprovada pela especificidade dos temas abordados e do expressivo número de palavras-chave em comum. O método é, entretanto, como discutido no referencial teórico (MacRoberts M.H. & MacRoberts B.R., 1989; Alvarenga, 1998; Macias-Chapula, 1998), limitado por várias restrições práticas. Muitas vezes, os autores não estão cientes da relevância de associar o seu documento a outro pelo grau de acoplamento bibliográfico e podem até omitir deliberadamente algumas referências.

Os artigos encontrados no segundo experimento não foram identificados como pertencentes à mesma categoria do primeiro. Como já destacado na introdução deste texto, em domínios específicos de conhecimento, grande parte dos artigos apresenta palavras-chave semelhantes ou relacionadas, dificultando a identificação de características que possam diferenciar as categorias quando o atributo de categorização é a palavra-chave. Os estudos que ilustram a utilização de análise de citações usam, como neste trabalho, um *corpus* de um domínio de conhecimento restrito. A efetividade do método reside na oportunidade de se identificar o relacionamento de documentos, independentemente da presença de termos iguais, abrindo-se a possibilidade de se obter informação, como discutido no exemplo de patentes no referencial teórico (Hall *et al*., 2002), para o estudo do desenvolvimento tecnológico de uma área específica de pesquisa. O fato de que uma patente cita outra deve ser um indicativo de que o conhecimento flui entre os dois documentos. Neste trabalho, a comparação realizada e a constatação de diferenças nas categorias resultantes, sugere a possibilidade de utilização da citações como um complemento ao processo de categorização que utiliza palavras-chave.

Conclusão

Na presente pesquisa, a edição e a limpeza dos dados consumiram grande parte do tempo dedicado à fase de pré-processamento. A base de dados foi construída a partir de informações extraídas de textos disponíveis na *Internet* e os códigos em *html*, utilizados para a apresentação de acentos ou expressões alfanuméricas, foram substituídos para que, tanto o programa que preparava os dados para a entrada da RNA quanto o *software* que utilizava as palavras-chave como atributo, pudessem processar corretamente os dados. O trabalho de edição exigiu que fossem uniformizadas as palavras constantes nos títulos dos trabalhos e campos de palavras-chave, que se apresentavam, em alguns casos, no singular, em outros, no plural ou com o acréscimo de aspas ou de hífens. Foram identificadas ausências e simplificações de palavras e nomes de autores. Os erros e inconsistências presentes em uma base de citações podem gerar um cálculo incorreto de índices associados à citação, como

http://dx.doi.org/10.1590/2318-08892016002800007

os descritos neste trabalho, e nos processos de categorização e recuperação de informação que utilizam essas bases de dados.

Os resultados experimentais encontrados confirmam que a escolha do atributo utilizado no processo de categorização deve considerar a natureza da base de dados e os recursos disponíveis ao pesquisador. No caso deste trabalho, a base de dados construída a partir de um domínio de conhecimento restrito dificulta a formação de grupos que utilizam palavras-chave como variável do processo de categorização, devido à grande similaridade existente entre as palavras selecionadas pelos autores dos artigos. Como existe uma grande repetição, confirmada pela análise da frequência com que foram utilizadas pelos autores, o processo de categorização por palavras-chave mostrou-se pouco eficiente para gerar grupos de artigos especificamente relacionados em um domínio de conteúdo tão restrito. As questões associadas à escalabilidade podem ser exploradas em um estudo futuro, discutindo-se a possibilidade de categorizar um grupo maior de documentos e implementando os métodos discutidos em uma base de dados com características distintas.

Em um domínio de conhecimento específico, a categorização de artigos utilizando citações pode ser considerada uma importante alternativa. As pesquisas de acoplamento bibliográfico vêm sendo explorados com o objetivo de avaliar a similaridade entre os documentos e têm sido reconhecidos como fonte de informação para o estudo da inovação e do desenvolvimento de uma área específica do conhecimento. As razões que levam um autor a selecionar um trabalho e relacioná-lo em sua lista de citações serão sempre difíceis de explicitar. Mas, mesmo considerando a limitação prática relacionada à possibilidade de que as citações possam ser tendenciosas ou preconcebidas, elas podem se tornar uma importante fonte de informação nos processos de representação de documentos e, consequentemente, nos de categorização e de recuperação de informação.

Agradecimentos

Pelo apoio financeiro da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Processo BEX 18208-12-6).

Colaboradores

Todos os autores contribuíram na concepção e desenho do estudo, análise de dados e redação final.

Referências

Alvarenga, L. Bibliometria e arqueologia do saber de Michel Foucault: traços de identidade teórico-metodológica. *Ciência da Informação*, v.27, n.3, 1998. Disponível em: <http://www.scielo.br/pdf/ci/v27n3/27n3a02.pdf>. Acesso em: 4 jan. 2014.

Andrade, F.S.; Jung, C. F. Análise de referências utilizadas por pesquisadores na revista Gestão & Produção. *Transinformação*, v.25, n.1, 2013. Disponível em: <http://periodicos.puc-campinas.edu.br/seer/index.php/transinfo/article/view/1784/1700>. Acesso em: 4 jan. 2014.

Bakus, J. *et al.* A SOM-Based document clustering using phrases. In: International Conference on Neural Information Processing, 9., 2002, Singapore. *Proceedings*... Singapore: Nanyang Technological University, 2002. p.2212-2216.

Borgman, C.L.; Furner, J. Scholarly communication and bibliometrics. *Annual Review of Information Science and Technology*, v.36, n.1, p.2-72, 2002.

Braga, A.P. *et al. Redes neurais artificiais*: teoria e aplicações. Rio de Janeiro: LTC, 2000.

Chu, H. *Information representation and retrieval in the digital age*. New Jersey: Information Today, 2003.

Croft, W.B. *et al. Search engines*: Information retrieval in practice. New Jersey: Pearson Addison Wesley, 2010.

Fujii, A. *et al.* Introduction to the special issue on patent processing. *Information Processing Management*, v.43, n.5, p.1149-1153, 2007.

Hall, B.H. *et al.* The NBER patent citations data file: Lessons, insights and methodological tools. In: Jaffe, A.B.; Trajtenberg, M. (Ed.). *Patents, citations & innovations*. London: MIT Presss. 2002. p.403-459.

Haykin, S. *Neural networks*: A comprehensive foundation. New Jersey: Prentice Hall, 1994.

Hjorland, B. What is Knowledge Organization (KO)? *Knowledge Organization*, v.35, n.2/3, p.86-101, 2008.

Kajanan, S. *et al.* Efficient automatic search query formulation using phrase-Level analysis. *Journal of the Association for Information Science and Technology*, v.65, n.5, p.1058-1075, 2014.

Kessler, M.M. Bibliographic coupling between scientific papers. *American Documentation*, v.14, n.1, p.10-25, 1963.

Kim, Y.G. *et al*. Visualization of patent analysis for emerging technology. *Expert Systems with Applications*: *An International Journal*, v.34, n.3, p.1804-1812, 2008.

Kuramoto, H. Sintagmas nominais: uma nova proposta para a recuperação da informação. *DataGramaZero*, v.3, n.1, 2002. Disponível em: <http://www.dgz.org.br/fev02/F_I_art.htm>. Acesso em: 4 dez. 2013.

Lai, K.K; Wu, S.J. Using the patent co-citation approach to establish a new patent classification system. *Information Processing and Management*: *An International Journal*, v.41, n.2, p.313-330, 2005.

Lee, S. *et al*. An approach to discovering new technology opportunities: Keyword-based patent map approach. *Technovation*, v.29, p.481-497, 2009.

Lensu, A.; Koikkalainen, P. Similar document detection using self-organizing maps. In: International Conference on Knowledge-Based Intelligent Information & Engineering Systems, 3., 1999, Adelaide. *Proceedings*... Adelaide: Institute of Electrical and Electronics Engineers Inc., 1999. p.174-177.

Liu, D.R.; Shih, M.J. Hybrid-patent classification based on patent-network analysis. *Journal of the American Society for Information Science and Technology*, v.62, n.2, p.246-256, 2011.

Macias-Chapula, C.A. O papel da informetria e da cienciometria e sua perspectiva nacional e internacional. *Ciência da Informação*, v.27, n.2, 1998. Disponível em: <http://revista.ibict.br/cienciadainformacao/index.php/ciinf/article/view/342/1689>. Acesso em: 10 dez. 2013.

MacRoberts, M.H.; MacRoberts, B.R. Problems of citation analysis: A critical review. *Journal of the American Society for Information Science*, v.40, n.5, p.342-349, 1989.

Marshakova, I.V. A system of document connection based on references. *Scientific and Technical Information Serial of VINITI*, v.6, n.2, p.3-8, 1973.

Meireles, M.R.G. Categorização de documentos a partir de suas citações: um método baseado em redes neurais artificiais. 2012. Tese (Doutorado em Ciência da Informação) - Universidade Federal de Minas Gerais, Belo Horizonte, 2012.

Meireles, M.R.G. *et al*. Bibliometric knowledge organization: A domain analytic method using artificial neural networks. *Knowledge Organization*, v.41, n.2, p.145-159, 2014.

Nicolelis, M. *Muito além do nosso eu*: a nova neurociência que une cérebros e máquinas e como ela pode mudar nossas vidas. São Paulo: Companhia das Letras, 2011.

Phuc, D.; Hung, M. X. Using SOM based graph clustering for extracting main ideas from documents. In: IEEE International Conference on Research, Innovation and Vision for the Future. *Computing & Communication Technologies*. Ho Chi Minh, 6., 2008. Ho Chi Minh City: Vietnam National University, 2008. p.209-214.

Richter, G.; Macfarlane, A. The impact of metadata on the accuracy of automated patent classification. *World Patent Information*, v.27, n.1, p.13-26, 2005.

Sapsalis, E. *et al*. Academic versus industry patenting: An in-depth analysis of what determines patent value. *Research Policy*, v.35, n.10, p.1631-1645, 2006.

Small, H. Co-citation in the scientific literature: A new measurement of the relationship between two documents. *Journal of the American Society for Information Science*, v.24, n.4, p.265-269, 1973.

Smeaton, A.F. Prospects for intelligent, language-based information retrieval. *Online Review*, v.15, n.6, p.373-382, 1991. Available from: <http://www.emeraldinsight.com/toc/or/15/6>. Cited: 4 Dec. 2013.

Tomovic, A. *et al*. n-Gram-based classication and unsupervised hierarchical clustering of genome sequences. *Computer Methods and Programs in Biomedice*, v.81, n.2, p.137-153, 2006.