



XI Encontro Nacional de Pesquisa em Ciência da Informação
Inovação e inclusão social: questões contemporâneas da informação
Rio de Janeiro, 25 a 28 de outubro de 2010

GT8 - Informação e Tecnologia
Modalidade de apresentação: Pôster

CATEGORIZAÇÃO E CLASSIFICAÇÃO DE DOCUMENTOS UTILIZANDO REDES NEURAIS ARTIFICIAIS

Magali Rezende Gouvêa Meireles
Pontifícia Universidade Católica de Minas Gerais
Beatriz Valadares Cendón
Universidade Federal de Minas Gerais

RESUMO

A categorização é um mecanismo que simplifica as interações do indivíduo com o meio que o cerca, facilitando o processo de recuperação de informação. O objetivo geral deste trabalho é verificar a eficiência e a viabilidade do uso de Redes Neurais Artificiais para categorizar e classificar documentos utilizando-se as referências bibliográficas citadas. Como, naturalmente, ao buscarmos novas informações já nos utilizamos de experiências anteriores, as categorias encontradas serão referência para a recuperação de artigos que apresentem características similares. Como as citações são uma importante fonte de informação sobre o relacionamento existente entre os documentos, os artigos serão categorizados e recuperados utilizando-as como atributo. São apresentadas, inicialmente, as abordagens teóricas relacionadas ao processo de categorização, às técnicas bibliométricas e ao uso de Redes Neurais Artificiais. A seguir, é descrita a técnica utilizada no trabalho, que se utiliza de uma Rede Neural Artificial para reconhecer similaridades entre entidades e, posteriormente, agregá-las em categorias. Finalmente, são apresentados os resultados preliminares da pesquisa, agrupando, em categorias, os artigos selecionados para esta demonstração.



1 INTRODUÇÃO

Com o crescimento da *Web*, desde o seu início em meados dos anos 90, várias atividades profissionais, dentre elas atividades comerciais, culturais e educacionais, têm se beneficiado de acessos constantes aos conteúdos *online*. Mesmo com tanta informação disponível e com a possibilidade de se utilizar inúmeras ferramentas, nem sempre o usuário atinge o objetivo de sua busca. Segundo Nachmias e Gilad (2002), 71% dos usuários demonstram frustração ao buscarem informações específicas na *Web*. Uma das discussões presente no trabalho destes autores ressalta o fato de que o processo de busca na *Web* exige habilidades para utilizar as ferramentas disponíveis, conhecimento de técnicas de pesquisa e capacidade cognitiva para organizar a pesquisa.

Normalmente, o usuário utiliza interfaces por meio das quais traduz sua necessidade de informação em forma de questões ou palavras-chave ou examina os documentos na busca de informações relevantes (SOUZA, 2006). Como o percentual de usuários insatisfeitos com o resultado da busca é significativamente alto, constata-se que o processo de recuperação de informação continua sendo um problema identificado na literatura. Este trabalho propõe uma alternativa ao processo de recuperação da informação, avaliando a possibilidade da utilização das referências listadas nos documentos para a recuperação destes documentos. Como as citações são uma importante fonte de informação sobre o relacionamento, inclusive semântico, existente entre os documentos, os artigos serão categorizados e recuperados utilizando-as como atributo.

O objetivo geral do trabalho é verificar a eficiência e a viabilidade do uso de RNAs para categorizar e classificar documentos, no processo de recuperação de informação, utilizando-se as referências bibliográficas citadas.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Categorização e classificação

Historicamente, a noção de categoria tem sido abordada de modos diversos. Segundo Xavier (2008), o conceito de categorias, como discutido nos dias de hoje,



nasceu com Aristóteles. “*Categorias*” é o primeiro dos cinco tratados que compõem “*Organon*”, a obra que expõe a lógica aristotélica.

Aristóteles determinou dez gêneros supremos que constituem as *Categorias*: substância, quantidade, qualidade, relação, lugar, tempo, posição, posse, ação e paixão (MORA *apud* XAVIER, 2008). Há quem afirme que Aristóteles não se limitou a esta tabela de categorias. Alguns dos seus discípulos defenderam a idéia de que o número de categorias proposto pelo filósofo era indeterminado. Estas categorias, segundo Xavier (2008), possibilitavam a ordenação do pensamento, fornecendo elementos para a perfeita caracterização do objeto a ser estudado.

Segundo Jacob (2004), categorização é o processo de dividir o mundo em grupos de entidades cujos membros têm similaridades entre eles dentro de um determinado contexto. Agregar as entidades em categorias leva o indivíduo a perceber ordem no mundo que o circunda. Segundo Barite (2000), as categorias são usadas como ferramentas para se descobrir certas regularidades do mundo material, concluindo-se que todos os objetos, pelo menos os que pertencem ao mundo material, possuem certas propriedades. Barite, afirma, ainda, que as categorias são abstrações simplificadas usadas pelos classificacionistas para investigar regularidades de objetos do mundo físico e ideal e para representar noções.

Utilizando-se o pensamento de Piaget, citado por Rojas (2005), o conhecimento é construído pelo sujeito com base na assimilação, na integração e na reorganização de estruturas que lhe permitem interpretar o mundo e interagir com ele. Portanto, para que o conhecimento possa ser disseminado, são realizados processos cognitivos conhecidos como classificação, ordenação e organização da informação.

A classificação é, portanto, uma capacidade conceitual humana, que utiliza as categorias como ferramentas. O processo de classificação envolve a associação ordenada e sistemática de cada entidade a uma única classe dentro de um conjunto de classes mutuamente exclusivas e que não se sobrepõem (JACOB, 2004).

2.2 Técnicas bibliométricas

A bibliometria oferece um conjunto de métodos e medidas para estudar a estrutura e o processo da comunicação científica. A análise de citações é a abordagem bibliométrica mais conhecida (BORGMAN; FURNER, 2002) e identifica relacionamento de



um documento com outro independentemente da presença de termos iguais nos documentos avaliados.

As citações bibliográficas denotam relações particulares entre o artigo citante e o citado e, nesse sentido, a citação pode ser entendida como um processo social no momento em que considera toda a experiência anterior do autor, sua rede de conhecimento e suas próprias considerações (LEAL, 2005).

Cronin, citado por Leal (2005), afirma que o hábito de citar mostra conformidade e consistência no ato da produção intelectual, muitas vezes, governada por conhecimento tácito e normas internalizadas. Os fatos citados no texto ganham credibilidade quando se apresenta a literatura utilizada, interligando o leitor a outras fontes de informação sobre o assunto. De acordo com Alvarenga (2009), cria-se um fluxo de informação e de comunicação entre aquele que cita e aquele que lê, proporcionando uma dinâmica circulante, que mantém viva a criação de um autor, concedendo-lhe significativo reconhecimento. Todos os aspectos relacionados confirmam a importância das interpretações dos resultados obtidos pela aplicação da análise de citações como método de avaliação de produtividade e de tendências da literatura científica no tempo e no espaço físico.

2.3 Redes Neurais Artificiais

Um neurônio artificial pode ser representado por um modelo matemático simplificado de processamento existente em um neurônio biológico. Uma rede neural artificial, RNA, pode ser definida como uma topologia interconectada de neurônios artificiais, na qual, tipicamente, se pode identificar neurônios de entrada, neurônios internos e neurônios de saída. A maneira como os neurônios estão organizados e conectados depende da arquitetura da rede. As RNAs implementam algoritmos que procuram alcançar um desempenho próximo dos sistemas neurológicos naturais, por meio de técnicas como aprendizado pela experiência e generalização a partir de situações similares.

As RNAs são utilizadas, basicamente, em problemas de aproximação, predição, classificação, categorização e otimização. A grande maioria das aplicações reportadas na literatura se concentra na área industrial (MEIRELES; ALMEIDA; SIMÕES, 2003). Em um Sistema de Recuperação de Informação, SRI, as RNAs podem ser usadas para realizar o



casamento de padrões entre as consultas e os documentos do acervo do sistema (SOUZA, 2006).

3 METODOLOGIA

A proposta geral do trabalho é realizar a categorização de artigos utilizando-se os dados contidos em suas referências. Para isso, será utilizada uma RNA, que receberá estes dados como entradas e apresentará, como saídas, as categorias de artigos agrupados por similaridades.

Para iniciar esta primeira fase da metodologia, serão cadastradas, em um banco de dados, todas as referências dos artigos previamente selecionados. O armazenamento das referências na base de dados minimiza a ocorrência de referências duplicadas.

Para validar a proposta sugerida anteriormente, foi proposto o treinamento de uma RNA, utilizando-se dez artigos fictícios, com algumas referências em comum. Os dez artigos apresentaram, juntos, um total de 64 referências.

Depois de identificados os atributos dos artigos que foram agrupados na mesma categoria, será proposto e implementado um SRI, para recuperar artigos que se relacionem às categorias propostas. Assim, o usuário poderá optar por recuperar, em outro banco de dados, apenas os artigos que estejam classificados em categoria já definida pela RNA e selecionada por ele, atendendo a sua necessidade.

4 RESULTADOS PRELIMINARES DA PESQUISA

Como resultado do estudo preliminar realizado, observou-se que, após o treinamento da RNA, os artigos foram agrupados em quatro categorias mutuamente exclusivas. A primeira categoria selecionou os artigos 1, 2 e 3, que apresentaram cinco referências comuns aos três artigos e quatro referências comuns aos dois primeiros artigos. Estes artigos foram publicados nos anos de 2002, 2003 e 2003 respectivamente.

A segunda categoria agrupou os artigos 4, 5 e 6, que apresentaram sete referências comuns aos três artigos e uma referência comum aos artigos 4 e 6. Os artigos foram publicados em 2005, 2006 e 2006.



A terceira categoria foi composta pelos artigos 7, 8 e 9. Seis referências eram comuns aos três artigos e duas referências, comuns aos artigos 8 e 9. Considerou-se 2007, 2008 e 2006, como o ano de publicação destes artigos.

O artigo 10 foi indicado para a quarta categoria. Este artigo tinha referências comuns aos outros oito, dos nove artigos restantes, e seu ano de publicação era 2005.

Para realizar esta categorização, a rede utilizou o ano de publicação do artigo e as referências de cada um deles. Após alguns ajustes nos parâmetros da RNA, o resultado apresentado sugere que a categorização tenha sido realizada com sucesso, agrupando artigos que apresentavam o maior número de referências em comum e anos de publicação próximos.

Deste modo, o processo de categorização poderá ser utilizado para selecionar, dentre um conjunto de artigos de outra base de dados, aqueles que forem identificados como pertencentes às categorias definidas anteriormente pela RNA.

ABSTRACT

Categorization is a mechanism that simplifies the interactions of the individual with the environment that surrounds him, easing the process of information retrieval. The aim of this study is to verify the efficiency and viability of using Artificial Neural Networks to categorize and classify documents using the references cited. As, naturally, by seeking new information we use past experiences, the found categories will be reference to the recovery of items that exhibit similar characteristics. As citations are an important source of information concerning the relationship between the documents, articles will be categorized and retrieved using them as an attribute. Firstly, the theoretical approaches regarding the process of categorization, the bibliometric techniques and the use of Artificial Neural Networks are presented. Next, the technique used at work is described, which uses an Artificial Neural Network to recognize similarities between entities and then aggregate them into categories. Finally, we present the preliminary results of research, grouping, into categories, the selected articles for this demonstration.

REFERÊNCIAS

ALVARENGA, Lídia. **Citações como método de trabalho na bibliometria**: notas de aula da disciplina Análises no processo classificatório para Sistemas de Recuperação de Informação, PPGCI – UFMG, 24f, 2009.



BARITE, Mario Guido. **The Notion of “Category”: Its Implications in Subject Analysis and in the Construction and Evaluation of Indexing Languages.** Knowledge organization, v. 27, n.1/n.2, p. 4-10, 2000.

BORGMAN, Christine L; FURNER, Jonathan. **Scholarly Communication and Bibliometrics.** Annual Review of Information Science and Technology, v. 36, n. 1, p. 2-72, 2002. Disponível em: <http://link.periodicos.capes.gov.br/sfxlcl3?url_ver=Z39.88-2004&url_ctx_fmt=infofi/fmt:kev:mtx:ctx&ctx_enc=info:ofi/enc:UTF-8&ctx_ver=Z39.88-2004&rft_id=info:sid/sfxit.com:azlist&sfx.ignore_date_threshold=1&rft.object_id=954925458043&svc.fulltext=yes> Acesso em: 12 maio 2010.

JACOB, Elin. **Classification and categorization: a difference that makes a difference.** Library trends, v. 52, n. 3, p. 515-540, 2004.

LEAL, Igor Campos. **Análise de citações da produção científica de uma comunidade: a construção de uma ferramenta e sua aplicação em um acervo de teses e dissertações do PPGCI - UFMG.** 2005. 94f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal de Minas Gerais. Belo Horizonte. Disponível em: <http://www.bibliotecadigital.ufmg.br/dspace/bitstream/1843/VALA-6KFNN9/1/mestrado_igor_campos_leal.pdf>. Acesso em: 19 abr 2010.

MEIRELES, Magali Rezende Gouvêa; ALMEIDA, Paulo Eduardo Maciel; SIMÕES, Marcelo Godoy. **A Comprehensive Review for Industrial Applicability of Artificial Neural Networks.** IEEE Transactions on Industrial Electronics. EUA, v. 50, n. 3, p. 1-18, 2003.

NACHMIAS, Rafi; GILAD, Amir. **Needle in a Hyperstack: Searching for Information on the World Wide Web.** Journal of Research on Technology in Education, v. 34, n. 4, p. 475-486, 2002.

ROJAS, Miguel Ángel Rendón. **Relación entre los conceptos: información, conocimiento y valor. Semejanzas y diferencias.** Disponível em: <<http://revista.ibict.br/index.php/ciinf/article/view/628/560>>. Acesso em: 03 nov. 2009.

SANTOS, Boaventura de Souza. **Um discurso sobre as ciências.** São Paulo: Cortez, 2003.

SOUZA, Renato Rocha. **Sistemas de Recuperação de Informações e Mecanismos de Busca na web: panorama atual e tendências.** Perspectivas em ciência da informação, Belo Horizonte, v. 11, n. 2, p. 161-173, mai./ago. 2006.



XI Encontro Nacional de Pesquisa em Ciência da Informação
Inovação e inclusão social: questões contemporâneas da informação
Rio de Janeiro, 25 a 28 de outubro de 2010

XAVIER, Beatriz Rêgo. **As categorias de Aristóteles e o conhecimento científico.**
Pensar, Fortaleza, v. 13, n. 1, p. 57-64, jan./jun. 2008.