Extração semiautomática de taxonomia para domínios especializados usando técnicas de mineração de textos

Fabiane dos Reis Braga
Chefe do Centro de Informações Nucleares da Comissão Nacional de Energia Nuclear (CNEN), Doutorado em Sistemas Computacionais de Alto Desempenho pela COPPE/UFRJ, fabiane@cnen.gov.br, http://buscatextual.cnpq.br/buscatextual/visualizacv.do?id=K4709086J8.

Submetido em: 22/08/2017. Aprovado em: 24/10/2017. Publicado em: 22/02/2018.

RESUMO

Apresenta metodologia para a extração semiautomática de uma taxonomia de conceitos, utilizando técnicas de mineração de textos, a partir de um corpus textual. A classificação de textos é uma prática natural do ser humano e uma tarefa crucial para se trabalhar com grandes repositórios. A técnica de agrupamento (*clustering*) de documentos fornece uma estrutura lógica e compreensível que facilita a organização, a navegação e a busca. A maioria dos algoritmos de agrupamento utiliza o modelo de saco de palavras (*bag of words*) para representar um documento. Este modelo gera alta dimensionalidade dos dados, ignora o fato de que diferentes palavras podem ter o mesmo significado e não considera o relacionamento entre elas, presumindo que as palavras são independentes umas das outras. A metodologia proposta apresenta a combinação de um modelo de representação de documentos por conceitos com um método de agrupamento hierárquico de documentos baseado na frequência da coocorrência dos conceitos e uma técnica de rotulação mais representativa, com o objetivo de produzir uma taxonomia de conceitos que possa refletir uma estrutura do domínio do conhecimento. A metodologia foi avaliada em um corpus textual composto de documentos científicos relacionados à área nuclear extraídos da produção científica da Comissão Nacional de Energia Nuclear (CNEN). Os resultados confirmaram que a mineração de textos serve como poderosa técnica para gerenciar conhecimento encapsulado em grandes coleções de documentos e, assim, apoiar a gestão das atividades de pesquisa da área.

Palavras-chave: Gestão do conhecimento. Agrupamento de documentos. Agrupamento por conjunto de itens frequentes.

Semiautomatic extraction of taxonomy for specialized domains using text mining techniques

ABSTRACT

Presents a methodology for the semiautomatic extraction of a taxonomy of concepts, using techniques of text mining, from a textual corpus. The classification of texts is a natural practice of the human being and a crucial task to work with large repositories. The document clustering technique provides a logical and understandable structure that facilitates organization, navigation and search. Most clustering algorithms use the word of mouth (bag of words) model to represent a document. This model generates a high dimensionality of the data, ignores the fact that different words can have the same meaning and does not consider the relationship between them, assuming that the words are independent of each other. The proposed methodology presents the combination of a document representation model with a hierarchical document grouping method based on the frequency of co - occurrence of the concepts and a more representative labeling technique with the objective of producing a taxonomy of concepts that may reflect structure of the knowledge domain. The methodology was evaluated in a textual corpus composed of scientific documents related to the nuclear area extracted from the scientific production of the National Commission of Nuclear Energy (CNEN). The results confirmed that text mining serves as a powerful technique for managing encapsulated knowledge in large collections of documents and thus supporting the management of research activities in the area.

Keywords*: Knowledge management. Grouping of documents. Grouping by set of frequent items.*

Extracción semiautomática de taxonomía para dominios especializados usando técnicas de minería de textos

RESUMEN

Se presenta una metodología para la extracción semiautomática de una taxonomía de conceptos, utilizando técnicas de minería de textos, a partir de un corpus textual. La clasificación de textos es una práctica natural del ser humano y una tarea crucial para trabajar con grandes repositorios. La técnica de agrupamiento (clustering) de documentos proporciona una estructura lógica y comprensible que facilita la organización, la navegación y la búsqueda. La mayoría de los algoritmos de agrupación utilizan el modelo de bolsa de palabras para representar un documento. Este modelo genera una alta dimensionalidad de los datos, ignora el hecho de que diferentes palabras pueden tener el mismo significado y no considera la relación entre ellas, presumiendo que las palabras son independientes unas de otras. La metodología propuesta presenta la combinación de un modelo de representación de documentos por conceptos con un método de agrupación jerárquica de documentos basado en la frecuencia de la coocorrencia de los conceptos y una técnica de rotulación más representativa con el objetivo de producir una taxonomía de conceptos que pueda reflejar una estructura del dominio del conocimiento. La metodología fue evaluada en un corpus textual compuesto de documentos científicos relacionados al área nuclear extraídos de la producción científica de la Comisión Nacional de Energía Nuclear (CNEN). Los resultados confirmaron que la minería de textos sirve como una poderosa técnica para administrar conocimiento encapsulado en grandes colecciones de documentos y, de esa forma, apoyar la gestión de las actividades de investigación del área.

Palabras clave*: Gestión del conocimiento. Agrupación de documentos. Agrupamiento por conjunto de elementos frecuentes.*

INTRODUÇÃO

O conceito de taxonomia torna-se cada vez mais importante à medida que o volume de informações aumenta exponencialmente, e os usuários adquiriram papel-chave tanto na produção como no uso e categorização da informação. As taxonomias são aplicadas para portais institucionais, repositórios institucionais, Web Semântica, ontologias, gestão de informação (IM) e gestão de conhecimento (KM) como um novo motor de consulta ao lado das ferramentas de pesquisa tradicionais. De acordo com Hodge (2000), as taxonomias, como as ontologias e os tesauros, são estruturas de classificação que formam os principais tipos de estruturas de organização e representação do conhecimento. O processo de construção e manutenção de uma taxonomia, quando envolve grandes coleções de textos, demanda tempo e é custoso, tornando-se extremamente complexo. Por esta razão, e devido ao grande volume e riqueza de documentos textuais digitais, surge a necessidade de buscar técnicas automatizadas que auxiliem na identificação de padrões para grandes volumes de dados textuais, que apoiem a condução desse processo visando aperfeiçoá-lo (KASHYAP et al.,2005; ECHARTE et al., 2007).

A metodologia de mineração de textos, conhecida como *Text Mining* (FELDMAN e DAGAN, 1995) surgiu em razão da demanda para o tratamento de dados textuais, escritos em linguagem natural não estruturada, possibilitando encontrar padrões e tendências em conjuntos de documentos, classificar documentos ou ainda comparar documentos. Desta forma, a mineração de textos vem possibilitando às instituições transformar grandes volumes de textos em conhecimentos úteis às suas estratégias.

Gerar uma taxonomia a partir de documentos envolve desafios tais como encontrar relações conceituais ou mapas conceituais que exigem que diversas questões especiais sejam resolvidas. No entanto, técnicas que utilizam algoritmos de agrupamento têm produzido bons resultados. O agrupamento de documentos é uma das mais importantes técnicas da mineração de textos que aborda a classificação não supervisionada de documentos em diferentes agrupamentos, onde os documentos de cada agrupamento compartilham algumas propriedades em comum de acordo com alguma medida de similaridade. Documentos no mesmo agrupamento apresentam alta similaridade, mas são dissimilares aos documentos que estão em outros agrupamentos (HAN e KIMBER, 2001). Algoritmos rápidos e de alta qualidade de agrupamento de documentos desempenham um papel importante para uma navegação e organização eficaz de informações. Devido as suas características, esta técnica tem sido bastante utilizada para apoiar a geração automática ou semiautomática de taxonomias.

Este trabalho explorou o processo de descoberta de conhecimento e utilizou as técnicas de mineração de textos em uma base de dados textuais de artigos científicos no desenvolvimento de uma metodologia para a geração semiautomática de uma taxonomia de conceitos apresentando uma nova forma de organização desses conhecimentos.

TRABALHOS RELACIONADOS

Nos últimos anos, foram realizadas pesquisas para geração de taxonomia usando uma combinação de várias técnicas. Algumas das abordagens utilizadas para gerar taxonomia automática e semiautomática incluem o seguinte (KASHYAP et al., 2005):

- aplicação de técnicas de PLN (processamento de linguagem natural) para gerar uma taxonomia de conceitos e suas relações;
- uso de abordagens de aprendizagem supervisionada que exigem uma coleção de exemplos de treinamento;
- agrupamentos e métodos de mineração de dados para facilitar a pesquisa, categorização e visualização de dados;
- uso do dicionário WordNet (banco de dados lexical), Web e um tesauro.

O uso de técnicas de descoberta de conhecimento em banco de dados (*knowledge discovery in database* - KDD) na geração de taxonomias, especialmente em relação à tarefa de agrupamento hierárquico, é uma questão que já foi explorada por vários autores. Woon e Madnick (2009) apresentaram um novo método para a construção automática de taxonomias para domínios específicos de pesquisa. A metodologia proposta utiliza frequências de coocorrência de termos como um indicador da proximidade semântica entre os termos. Para apoiar a criação automatizada de taxonomias eles apresentaram uma simples modificação da medida básica de distância e descreveram um conjunto de procedimentos pelos quais essas medidas podem ser convertidas em estimativas da taxonomia desejada. Punera et al. (2005) propuseram um método de geração de hierarquia usando agrupamento de cima para baixo. Os autores geram uma taxonomia com cada nó associado a uma lista de categorias. Cada nó da folha possui apenas uma categoria. Este algoritmo, basicamente, usa dois centroides de categorias que estão mais distantes como as sementes iniciais e, em seguida, aplica o algoritmo *spherical k-means*. Cada categoria é atribuída a um grupo (*cluster*) se a maioria dos seus documentos pertencer ao grupo (sua relação excede um parâmetro predefinido). Caso contrário, esta categoria está associada a ambos os grupos. Esse método gera uma taxonomia com uma categoria possivelmente ocorrendo em múltiplos nós.

Kashyap et al. (2005) apresentaram uma estrutura de experimentação para a construção de taxonomia automatizada a partir de um grande corpus de documentos que envolve: (a) a geração de uma hierarquia de grupos de documentos usando o algoritmo *bisecting K-means* (um algoritmo hierárquico divisivo) com a métrica de distância de cosseno; (B) extração de taxonomia desta hierarquia; E (c) a atribuição de rótulos a nós na taxonomia. Eles recorreram a um conjunto de técnicas de agrupamento e PLN e parâmetros identificados para formar a base de uma estrutura de experimentação.

O algoritmo F*requent Itemset based Hierarchical Clustering* (FIHC) para agrupamento foi desenvolvido por (FUNG et al., 2003) e se baseia na ideia de conjuntos de itens frequentes (*Frequent Item Sets* FIs) proposta por (AGRAWAL e SRIKANT, 1994). Esta técnica emprega a noção de FIs para a construção e organização dos agrupamentos em uma hierarquia de tópicos. O FIs é um conjunto de termos que coocorrem ou ocorrem conjuntamente em uma fração mínima de documentos. Espera-se que documentos no mesmo agrupamento compartilhem mais FIs em comuns do que com os documentos que estão nos outros agrupamentos. Algumas características importantes dessa abordagem proposta são: redução da dimensionalidade do vetor de documentos, criação de agrupamentos com maior precisão, número de agrupamentos como parâmetro de entrada opcional e facilidade para navegação pelos agrupamentos que apresentam descrições significativas. No entanto, ele ignora o importante relacionamento semântico entre os termos.

Neste trabalho apresenta-se uma metodologia que propõe um método de agrupamento de documentos inspirado no algoritmo FIHC, mas que se baseia na noção de conjuntos de conceitos frequentes para a geração dos agrupamentos e organização dos mesmos numa árvore de tópicos, já que este algoritmo não considera a relação semântica entre os termos. A escolha do método FIHC se deve ao fato de que em comparação com outros métodos de agrupamento como UPGMA aglomerativo (JAIN e DUBES,1998; KAUFMAN e ROUSSEEUW, 1990), o *Bisecting k-means* (JAIN e DUBES,1998; KAUFMAN e ROUSSEEUW, 1990) e o HFTC (BEIL; ESTER; XIAOWEI, 2002), o FIHC, teve melhor desempenho em termos de acurácia, eficiência e sensibilidade a parâmetros, pois permite que o usuário defina a quantidade de agrupamentos, e escalabilidade. Além disso, a estrutura hierárquica da árvore de tópicos gerada por este método permite uma navegação mais eficiente, pois apresenta uma rotulação representativa dos agrupamentos.

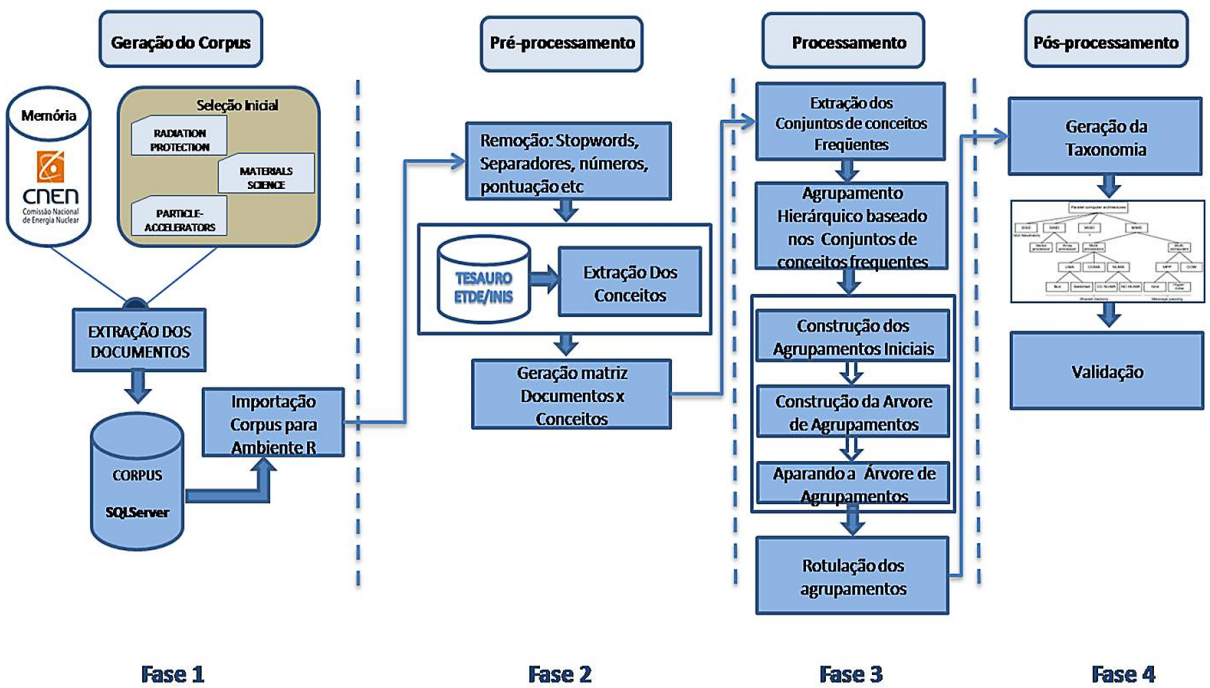
METODOLOGIA PARA EXTRAÇÃO DA TAXONOMIA

A metodologia para extração da taxonomia visa cobrir todas as etapas do processo de mineração de texto, a partir da seleção de documentos que farão parte do corpus até a geração semiautomática de uma estrutura taxonômica. Uma nova abordagem do método hierárquico de agrupamento de documentos, com base na noção de conjuntos de conceitos frequentes, foi aplicada e o resultado final é uma árvore de grupos onde os nós podem ser vistos como tópicos e subtópicos. A estrutura utilizada para gerar a taxonomia a partir de documentos textuais é ilustrada na figura 1.

GERAÇÃO DO CORPUS

O corpus foi composto por documentos da biblioteca digital da Comissão Nacional de Energia Nuclear (CNEN). São 1841 trabalhos científicos que abordaram várias áreas no domínio nuclear. Estes documentos estavam inicialmente em formato PDF e foram convertidos para o formato de texto. O ambiente de desenvolvimento foi o *software open source* R (R *Development Core Team*, 2012), poderosa ferramenta para análise de dados e representação gráfica que oferece excelente estrutura para fins de mineração de texto. Essa estrutura permite aos usuários trabalhar de forma eficiente com textos e metadados correspondentes e transformar os textos em representações estruturadas, onde os métodos existentes podem ser aplicados, por exemplo, para agrupamento ou classificação.

Figura 1 – Estrutura para extração da taxonomia



Fonte: Braga (2013)

PRÉ-PROCESSAMENTO DE DOCUMENTOS

Antes do processamento, a coleção de documentos passa por várias etapas de pré-processamento. O pré-processamento é um passo muito importante porque pode afetar o resultado de um algoritmo de agrupamento. O procedimento de pré-processamento consiste nas seguintes subetapas:

- *tokenização*: é utilizada para decompor o documento em cada termo que o compõe. Nossos textos foram divididos em palavras;
- *stopword*: remoção de dígitos e pontuação;
- eliminação de espaço em branco e conversão dos textos em minúsculas;
- etiquetagem de *Part-of-Speech* (POS): foi utilizado o conjunto de etiquetas PENN *Treebank Tag* (MITCHELL et al., 1993).

EXTRAÇÃO DOS TERMOS E CONCEITOS

Um tesauro desempenha papel essencial nos sistemas de recuperação de informações. Em particular, o tesauro de um domínio específico de conhecimento melhora consideravelmente a eficácia da recuperação de informações. Consiste de termos, cada um representando um conceito específico de domínio. O Thesaurus ETDE/INIS contém a terminologia controlada para indexação de todas as informações dentro dos âmbitos do *International Nuclear Information System* (INIS) e da *Energy Technology Data Exchange* (ETDE). A terminologia destina-se a ser utilizada nas descrições de assuntos para a entrada ou recuperação de informações nesses sistemas.

Os conceitos são termos agrupados por significado. Para cada termo, o dicionário de sinônimos ETDE/INIS identificou aqueles com o mesmo significado através da relação preferencial USE ou SEE, UF (*Used For*) e a construção de nossos conjuntos de sinônimos, ou seja, cada conjunto de sinônimos é equivalente a um conceito, como mostrado na figura 2.

Figura 2 − Conjuntos de sinônimos

Conceito 4999 “climates”
Conceito 5000 “**climatic change**” “global climate change”
Conceito 5001 “**nuclear energy**” “atomic energy”

Fonte: Braga (2013)

Como os termos correspondem à representação linguística de conceitos em textos (SAGER et al., 1980), o próximo passo é identificar e extrair de cada texto os termos e os termos múltiplos (n-gramas) de interesse para o nosso estudo. Primeiro, identifica-se como tais termos são estruturados sintaticamente. Avaliando dicionários de vocabulários técnicos, descobrimos que a maioria dos termos técnicos consiste principalmente em frases nominais contendo adjetivos, substantivos e algumas preposições e raramente contém verbos, advérbios e conjunções (KATZ e JUSTESON, 1995). A estrutura dos termos técnicos pode ser ilustrada através da avaliação de fontes de diferentes domínios, mas, para este estudo, apenas o tesauro ETDE/INIS foi utilizado.

REPRESENTAÇÃO DE DOCUMENTO BASEADA EM CONCEITO

No modelo de espaço vetorial, um documento é representado como um vetor de atributos d = (tf_{t1},..... tf_{ti}), onde tf_t retorna a frequência absoluta do termo t € T no documento d € D, onde D é o conjunto de documentos e T = {t1, t2,..... ti} é o conjunto de todos os termos diferentes encontrados em D. No método proposto, a medida Tf-Idf (*Term Frequency - Inverse Document Frequency*) é usado no vetor de representação do documento. Esta estatística avalia a importância de um termo para um documento em uma coleção ou corpus, e aumenta a precisão do agrupamento. A importância aumenta proporcionalmente ao número de vezes que uma palavra aparece no documento, mas é compensada pela frequência da palavra no corpus. Em outras palavras, se um termo/palavra aparecer várias vezes em um documento, mas também aparece várias vezes no corpus como um todo, ele receberá uma pontuação menor.

Em última análise, cada documento d é representado por um vetor de peso de conceito. No método proposto, os conceitos são identificados como um conjunto de termos que possuem significados comuns ou que possuem uma relação de sinonímia. A figura 3 mostra o processo de geração da matriz documento-conceito.

A partir de conjuntos de sinônimos, os termos que apresentam essa relação são substituídos pelo conceito principal nos documentos aos quais estão associados. O conceito de tamanho inferior a quatro caracteres e uma frequência inferior a cinco foi eliminado. O peso de cada conceito C no documento d é calculado como:

$$Pc = Cfc \times id_{fc}$$

Onde Cfc é a soma de cada frequência do termo dos termos associados ao conceito e id_{fc} é a frequência inversa do documento do conceito C, calculando o número de documentos nos quais o conceito C aparece. No final de cada documento, d é representado por um vetor de pesos dos conceitos.

$$d = (P_{c1}, P_{c2,} P_{c3,}.....,P_{ci})$$

PROCESSAMENTO

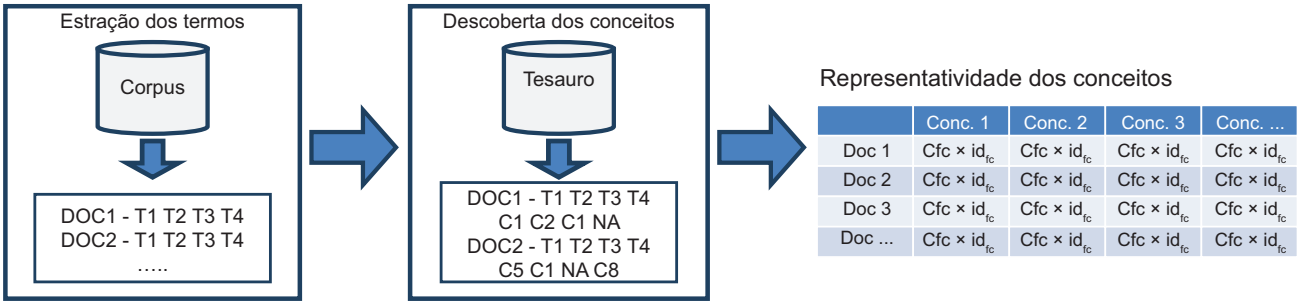
Ao contrário dos métodos aglomerativos e divisivos que são "*document-centered*", onde a similaridade entre os documentos é ponto chave na construção dos agrupamentos no método FIHC, no qual se baseia este trabalho, a medida de coesão de um agrupamento é feita diretamente usando os conjuntos de itens frequentes, no nosso caso, os conjuntos de conceitos frequentes, é "*cluster-centered*".

Os documentos que aparecem num mesmo agrupamento compartilham mais conjuntos de conceitos do que aqueles que estão em outros agrupamentos.

O algoritmo para agrupamento de documentos pode ser resumido em três fases: construção dos grupos iniciais, criação de uma árvore de grupos (tópico) e poda da árvore de grupos, caso haja muitos grupos ou, se o usuário quiser refinar a estrutura da taxonomia.

Construindo os grupos: o algoritmo *Apriori* (AGRAWAL e SRIKANT, 1994) foi utilizado no método proposto para gerar os conjuntos de conceitos frequente. Para cada conjunto de conceitos frequente, um grupo inicial é construído para incluir todos os documentos que contêm o conjunto de conceitos específico. Os grupos iniciais se sobrepõem porque um documento pode conter vários conjuntos de conceitos frequentes. Esse conjunto de conceitos frequente é usado como rótulo do grupo para sua identificação. Para cada documento, o "melhor" grupo inicial é identificado e o documento é atribuído apenas ao grupo inicial mais correspondente. A adequação de um grupo C_i para um documento doc_j é medida por uma função de pontuação usando os conceitos dos grupos frequentes dos grupos iniciais. Após essa etapa, cada documento pertence exatamente a determinado grupo. O conjunto de grupos pode ser visto como um conjunto de tópicos no conjunto de documentos.

Figura 3 – Processo para geração da matriz documentos *versus* conceitos



Fonte: Braga (2013)

Construindo a árvore de grupos: na árvore de grupos, cada grupo (exceto o nó raiz) tem exatamente um pai. O tópico de um grupo pai é mais geral que o tópico de um grupo filho, e eles são "semelhantes" até certo ponto. Cada grupo usa um conjunto de k-conceito frequentes como seu rótulo de grupo. Um grupo com um rótulo de grupo de conjunto de conceito k aparece no nível k na árvore. A árvore do grupo é criada de baixo para cima, escolhendo o "melhor" pai no nível k-1 para cada grupo no nível k. O rótulo do grupo do pai deve ser um subconjunto do rótulo do grupo do filho. Tratando todos os documentos no grupo filho como um documento único, o critério para selecionar o melhor pai é semelhante ao da escolha do melhor grupo para um documento.

Podando a árvore de grupos: o objetivo da poda de árvores é remover eficientemente os grupos específicos sobrepostos com base na noção de similaridade intergrupo. A ideia é que, se dois grupos irmãos são muito parecidos, eles devem ser incorporados em um único grupo. Se um grupo filho é muito semelhante ao pai (alta similaridade intergrupo), substitua o grupo filho pelo seu grupo pai. O grupo pai também incluirá todos os documentos do grupo filho.

ROTULANDO OS GRUPOS

As hierarquias dos documentos fornecem uma coleção de pontos de vista em diferentes níveis de granularidade, facilitando a visualização e a análise de grandes coleções de documentos. Os tópicos utilizados como descritores, para cada nível da hierarquia, desempenham um papel importante na assistência à navegação na árvore e na descrição abrangente do grupo. Um dos problemas das metodologias de geração semiautomática e automática de taxonomias é o processo de identificação do tópico ou lista de tópicos que é fundamental para discriminar cada grupo.

Muitas das abordagens existentes para rotulagem de agrupamento hierárquico são baseadas na avaliação da frequência dos termos dentro dos documentos do mesmo grupo; podemos mencionar Popescul e Ungar (2000), que propuseram dois métodos em suas pesquisas. O primeiro método baseia-se no significado do teste Qui-quadrado para detectar diferentes usos de palavras em diferentes grupos em uma hierarquia de documentos. O segundo método seleciona palavras que ocorrem com frequência em um grupo e efetivamente descreve o agrupamento de outros grupos interessados. Glover et al. (2002) mostraram como uma abordagem simples para listar os termos mais relevantes para cada grupo, ou seja, ordenar os termos com o uso de cálculos estatísticos, pode fornecer uma boa descrição do grupo, diferenciando o grupo de irmãos e pais na hierarquia.

O modelo de rotulagem proposto no trabalho combina as características estatísticas do grupo e seus descendentes em uma pontuação, gerando uma lista de tópicos e subtópicos, que posteriormente são enriquecidos com termos equivalentes extraídos do tesauro ETDE / INIS.

O processo de rotulagem proposto procede da seguinte forma:

- para cada grupo gerado, o algoritmo extrai um conjunto formado pelos conceitos mais frequentes. Um conceito é considerado frequente se estiver contido em uma quantidade mínima de documentos desse grupo. O suporte é fornecido para definir a quantidade mínima de documentos;
- depois de selecionar os conceitos mais frequentes, o algoritmo nos permite expandir os conceitos originais usando o dicionário de sinônimos ETDE/ INIS, somando-lhes os termos equivalentes ao conceito, isto é, aqueles com a notação UF no dicionário de sinônimos;
- de acordo com o esquema de enriquecimento proposto associado a cada um desses conceitos, seus termos equivalentes no tesauro são extraídos. Por exemplo, se o conceito principal é "Nuclear Energy", seu termo equivalente seria "Atomic Energy"; Se o conceito principal for "Radiation protection", os termos equivalentes seriam "health physics", "nuclear safety", "protection (radiation)," "radiation hygiene", "radiation safety", "radiological protection" e "safety" (Nuclear). Assim, se a consulta do usuário contiver o termo "Atomic Energy", os documentos que contenham esse termo serão recuperados assim como aqueles que contenham "Nuclear Energy".

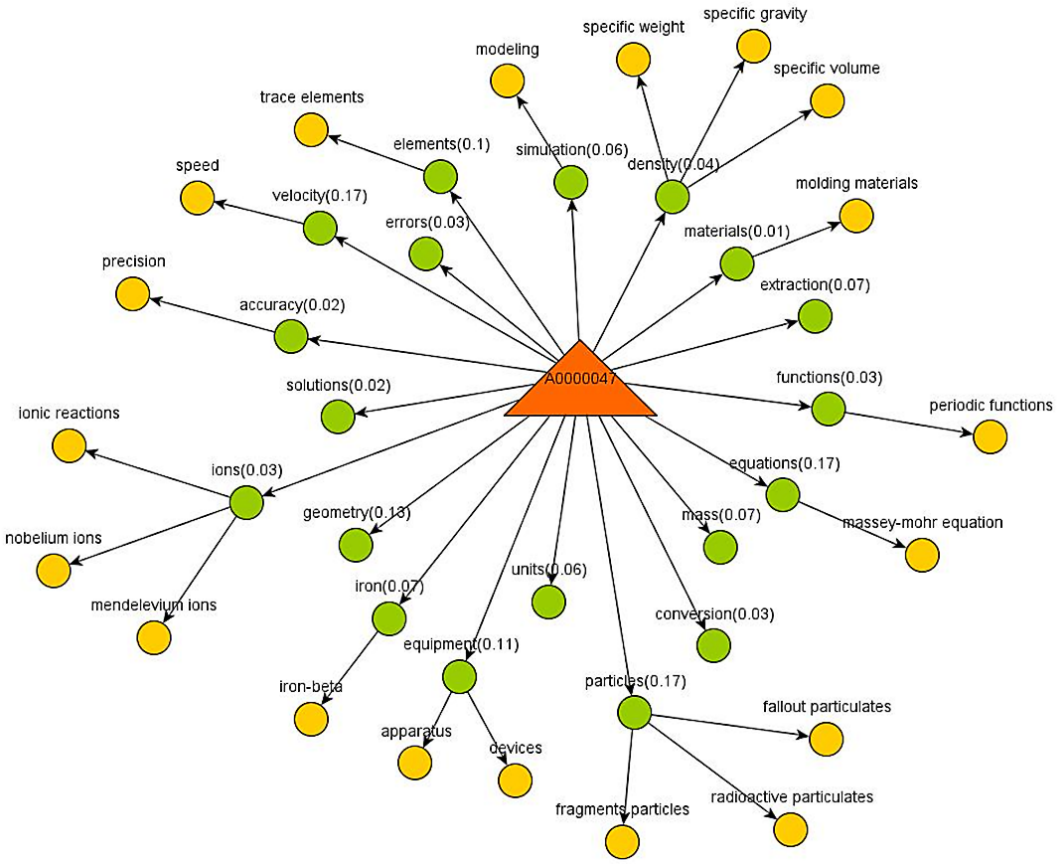
ANÁLISE E DISCUSSÃO DOS RESULTADOS

Neste trabalho propõe-se um modelo para representação dos documentos baseado em conceitos que considera as relações semânticas entre as palavras. Os termos que compõem os documentos são substituídos por conceitos do Tesauro ETDE/INIS. Na figura 4 pode-se ver a representação do documento A0000047 cujo título é "*Particle trajectory tracing and 2D electromagnetic fields simulations by finite element method*". O triângulo representa o documento. Os círculos mais internos representam os conceitos identificados no documento, a partir do tesauro, o número entre parênteses ao lado de cada conceito é o peso calculado daquele conceito usando a medida Tf x Idf. Os círculos mais externos associados aos conceitos são os termos pelos quais aquele conceito pode ser referenciado. Por exemplo, os termos "*specific weight*", "*specific gravity*" e "*specific volume*" podem ser usados para representar o conceito "*density*".

Este documento está associado à área de "Física das partículas elementares e de campos" segundo os especialistas. No modelo de representação proposto observa-se que os conceitos que representam o documento conseguem traduzir em linhas gerais a área de pesquisa ao qual o documento está associado. A vantagem da utilização deste modelo é que ele reduz as dimensões dos documentos e aprimora a acurácia no agrupamento de documentos sem que haja uma perda significativa na representatividade do documento.

A metodologia proposta apresentou nova estrutura para descoberta de conhecimento em documentos através da construção de uma taxonomia (hierarquia de conceitos) e a categorização dos textos por conceitos a partir da hierarquia. A estrutura taxonômica gerada capturou a hierarquia dos conceitos no corpus, com grau razoável de precisão.

Figura 4 – Representação de um documento



Fonte: Braga (2103)

Tabela 1 − Extrato dos GFIs gerados

CONJUNTO DE GFIs			
1 - GFIs (nível 1)			
radiation_protection	velocity	storage	simulation
2 - GFIs (nível 2)			
radiation_protection, safety	calibration, energy	energy, thickness	emission, energy
3 - GFIs (nível 3)			
Energy, materials nuclear_ energy	Energy, materials, safety	Energy, irradiation, materials	Energy, levels, materials

Fonte: Braga (2103)

A quantidade de GFIs gerados reflete o número inicial de agrupamentos que irá compor a árvore de agrupamentos. Para a geração dos GFIs aplicou-se o algoritmo Apriori na matriz de documentos versus conceitos sendo gerados 133 GFIs. Definir quantidade ideal de agrupamentos é tarefa complexa, pois não há critérios precisos para se basear esta decisão, assim, no primeiro momento, de forma a não se produzir uma árvore muito ampla e/ou profunda foi utilizado um critério subjetivo. Na tabela 1 é mostrado um extrato dos GIFs gerados.

AVALIAÇÃO DA TAXONOMIA

Partindo-se da disponibilidade dos dados resultantes do retorno dos formulários preenchidos pelos especialistas, fez-se a análise estatística para se verificar, no que se refere à classificação dos documentos, a concordância interespecialistas e entre os especialistas e a resposta do algoritmo. A análise dos gráficos indicou que os níveis de concordância entre as classificações feitas pelos especialistas e a classificação do algoritmo, bem como os níveis de concordância interespecialista não foram altos, mas são promissores.

A partir desses resultados buscou-se, então, entender as possíveis causas desse desempenho. Uma das razões encontradas foi o fato de que os especialistas apresentam níveis de conhecimento diferenciados sobre o domínio, o que acabou impactando nas avaliações individuais.

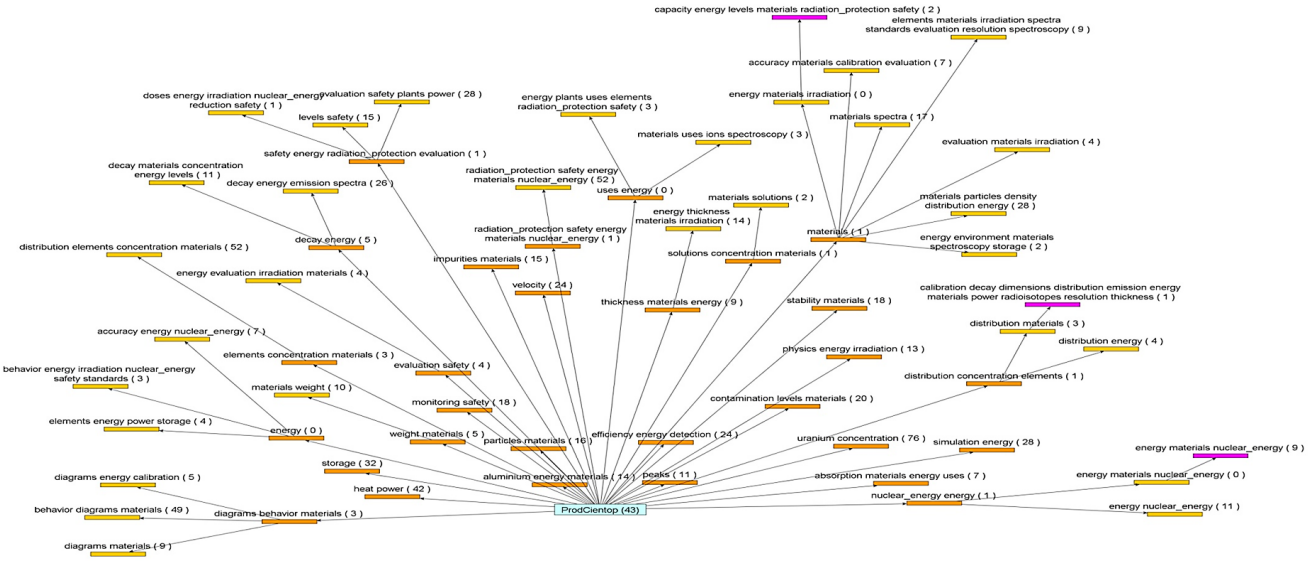
Outro motivo que pode ter contribuído para esse resultado se deve ao fato de que os resumos de alguns documentos não apresentavam a maioria dos conceitos mais frequentes do documento, e assim, os conceitos selecionados pelo algoritmo eram diferentes daqueles selecionados pelos especialistas. Quando solicitado, o texto completo foi fornecido ao especialista. Também foi observado que o especialista, pela sua experiência, utilizava conceitos que não necessariamente apareciam em determinado resumo, mas que estavam presentes na lista de conceitos, ou então, indicavam outros conceitos que não estavam presentes na lista. A ausência de conceitos mais específicos foi um ponto levantado por todos os especialistas.

Com base nos resultados e nas observações feitas, verifica-se que, para melhor desempenho, as etapas da metodologia referentes à identificação e extração dos conceitos, a partir dos textos, a geração dos GFIs, no que diz respeito à quantidade e à construção do rótulo do agrupamento merecem, no futuro, estudo mais aprofundado a fim de atender as solicitações formuladas.

AVALIAÇÃO DA ESTRUTURA TAXONÔMICA

A taxonomia gerada (figura 5) foi submetida aos especialistas para que fizessem suas críticas e sugestões quanto a sua estrutura e a sua usabilidade como ferramenta de apoio ao mapeamento conceitual da produção científica da CNEN.

Figura 5 – Parte da estrutura taxonômica gerada



Fonte: Braga (2013)

Uma das observações feitas pelos especialistas dizia respeito à generalidade dos conceitos extraídos. Os especialistas, acostumados ao uso do tesauro, esperavam encontrar maior quantidade de conceitos mais específicos, similares às relações NT (*narrow terms*), que os ajudassem a caracterizar melhor os nós. Também foi levantada a questão de que alguns agrupamentos pais e filhos apresentavam os mesmos rótulos, o que gerou dúvidas, pois o ideal é que os agrupamentos sejam diferenciados pelos seus rótulos para sua melhor caracterização. Esse fato pode ter sido ocasionado em função do valor do suporte utilizado para o cálculo dos conceitos mais frequentes, sinalizando um estudo mais aprofundado em relação a esse valor.

Outro ponto levantado pelos especialistas foi em relação à construção dos rótulos dos agrupamentos. A utilização de um conjunto de conceitos na caracterização de um agrupamento, em vez de apenas um conceito, foi vista como benéfica, pois trouxe melhorias na representatividade dos agrupamentos.

CONSIDERAÇÕES FINAIS

A organização automática de textos em linguagem natural por tópicos é tarefa desafiadora, pois envolve não apenas a identificação de tópicos, mas também a organização adequada. Ambas as tarefas exigem conhecimento de que as pessoas geralmente adquirem ao longo do tempo através da qualificação profissional.

O conhecimento gerado pela taxonomia pode ser usado para facilitar processos de organização e recuperação de informações, bem como sua própria compreensão da coleção textual organizada, ou mesmo servir como suporte para sistemas de suporte à decisão. Mas é importante que as técnicas sejam desenvolvidas para auxiliar especialistas no domínio, a fim de facilitar a compreensão e o uso dos conhecimentos adquiridos.

Este artigo apresentou uma nova abordagem para a geração semiautomática de uma taxonomia a partir da coocorrência de conceitos. Além de ser um passo no processo de criação de ontologias, esta técnica pode ser útil para melhor compreensão do domínio associado ao corpus em estudo.

Além disso, os resultados indicam que ainda há problemas técnicos que devem ser superados antes que este método possa ser totalmente utilizado.

O resultado final, após a aplicação da metodologia, é que a taxonomia gerada poderia permitir um mapeamento conceitual da produção científica da CNEN e que esse conhecimento poderia apoiar a gestão das atividades de pesquisa da instituição. Os resultados mostram que essa abordagem pode ser viável, embora melhorias e aprimoramentos sejam necessários para aumentar sua eficiência.

REFERÊNCIAS

AGRAWAL, R.; SRIKANT, R. Fast Algorithms for Mining Association Rules in Large Databases. In: INTERNATIONAL CONFERENCE ON VERY LARGE DATA BASES, 20., 1994, Santiago de Chile. *Proceedings*... Santiago de Chile: Morgan Kaufmann, 1994. p. 487-499.

AIEA. INIS Multilingual Thesaurus.Vienna: AIEA, 2004. Disponível em:<https://nkp.iaea.org/INISMLThesaurus/>. Acesso em: 10 set. 2011.

BEIL F.; ESTER, M.; XIAOWEI, Xu. Frequent term-based text clustering. In: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 8., 2002. *Proceedings*… New York: ACM, 2002. p. 436-442.

BRAGA, F. *Metodologia para Extração Semiautomática de Uma Taxonomia de Conceitos a partir da Produção Cientifica da Área Nuclear utilizando Técnicas de Mineração de Textos*. 2013. Tese (Doutorado em Engenharia Civil - Sistemas computacionais de alto desempenho) - Universidade Federal do Rio de Janeiro, COPPE/UFRJ, Rio de Janeiro, 2013.

ECHARTE, F. et al. Ontology of Folksonomy: A New Modeling Method. In: SEMANTIC AUTHORING, ANNOTATION AND KNOWLEDGE MARKUP WORKSHOP (SAAKM), 2007, Whistler. *Proceedings*… Whistler, British Columbia, 2007.

FELDMAN, R.; DAGAN, I. Knowledge discovery in textual databases (KDT). In: INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY (KDD-95), 1., 1995, Cambridge. *Proceedings...* Cambridge: AAAI/MIT Press, 1995. p. 112-117.

FUNG, B. C.M. et al. Hierarchical document clustering using frequent itemsets. In: SIAM INTERNATIONAL CONFERENCE ON DATA MINING , SDM, 3., 2003, San Francisco. *Proceedings* … San Francisco, 2003. p. 59-70.

GLOVER, E. et al. Inferring hierarchical descriptions. In: INTERNATIONAL CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT, 11., 2002, New York. *Proceedings* … New York: ACM, 2002, p. 507-514.

HAN, J.; KIMBER, M. *Data Mining*: concepts and Techniques. New York: Morgan Kaufmann, 2001.

HODGE, G. *Systems of knowledge organization for digital libraries*: beyond traditional authority files. Washington: The Digital Library Federation, 2000.

JAIN, A. K.; DUBES, R. C. *Algorithms for Clustering Data*. Cliffs, NJ: Prentice-Hall, 1988.

KASHYAP, V. et al.TaxaMiner: an experimentation framework for automated taxonomy bootstrapping. *International Journal of Web and Grid Services*, v. 1, n. 2, p. 240-266, 2005.

KATZ S.M.; JUSTESON T.S. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, v. 1, n. 1, p. 9-27, 1995.

KAUFMAN, L.; ROUSSEEUW, P.J. *Finding groups in data*: an introduction to cluster Analysis. New York: Jonh Wiley & Sons, 1990.

MITCHELL; M.; SANTORINI, B.; MARCINKIEWICZ, M. A. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, v. 19, n. 2, p. 313-330. Disponível em: <ftp://ftp.cis.upenn.edu/pub/treebank/doc/cl93.ps.gz >. Acesso em:15 ago. 2011.

POPESCUL, A. E.; UNGAR, L. *Automatic labeling of document clusters*. 2000. Disponível em: <ftp://ftp.cis.upenn.edu/pub/datamining/public_html/Publications/labels.pdf>. Acesso em: 11 set. 2012.

PUNERA, K.; RAJAN, S.; GHOSH, J. Automatically learning document taxonomies for hierarchical classification. In: INTERNATIONAL CONFERENCE ON WORLD WIDE WEB (WWW), 14, 2005, New York. *Proceedings*… New York: ACM, 2005, p. 1010-1011.

R DEVELOPMENT CORE TEAM. R: a1 language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2012. Disponível em: <http://www.R-project.org/>. Acesso em: 11 set. 2012.

SAGER, J. C.; DUNGWORTH, D.; MCDONALD, P. F. *English Special Language*: principles and practice in science and technology: Wiesbaden: Oscar Brandstetter Verlag KG, 1980.

WOON, W. L.; MADNICK, S. Asymmetric information Distances for Automated Taxonomy Construction. *Knowledge Information Systems*, v. 21, n. 1, p. 91-111, 2009.