GeCa - Um ambiente para apoio à criação Semi-automática de Categorias

Kelly Assis de Souza[1], José Marques Pessoa[2,3], Crediné Silva de Menezes[1,2]

[1]PPGI - Departamento de Informática, Centro Tecnológico
Universidade Federal do Espírito Santo, Av. Fernando Ferrari, s/n, Vitória,Es,Brasil, Tel:(27)3335-2654

[2]PPGEE, Departamento de Engenharia Elétrica, Centro Tecnológico,
Universidade Federal do Espírito Santo, Av. Fernando Ferrari, s/n, Vitória-Es, Tel: (27) 3335-2654

[3]ICLMA– Universidade Federal de Mato Grosso (UFMT)
Rod. MT 100, Km 4, Pontal do Araguaia – MT– Brasil. (27) 3340-9266

kasouza@click21.com.br, jmpessoa@npd.ufes.br, credine@inf.ufes.br

Resumo. *O surgimento da Internet trouxe consigo a facilidade de divulgação e acesso à informação, dando origem entretanto a um novo problema: com tantos documentos disponíveis, como alcançar a informação desejada? Uma solução para este problema, é organizar documentos em categorias. Neste artigo, é apresentado um ambiente para criar categorias onde documentos possam ser armazenados de forma organizada e que nas quais, novos documentos possam ser classificados automaticamente. Para que estes objetivos fossem alcançados, foram utilizadas técnicas de agrupamento automático de documentos e algoritmos de treinamento.*

Palavras-chave: agrupamento automático de documentos, algoritmos de treinamento, categorização de documentos, ambientes de educação a distância.

Abstract. *The sprouting from the Internet brought with himself the facility of disclosure and access to the information, creating however a new problem: with so many available documents, how to find information needed? A solution for this problem, is to organize documents in categories. In this article, is presented an environment to create categories where documents can be stored in an organized way and that new documents can be automatically classified. For these were utilized automatic techniques of document grouping and algorithms of training.*

Key words: automatic grouping documents, training algorithms, document categorization, environments of distance education.

1. Introdução

O surgimento da Internet possibilitou uma maior democratização no acesso e distribuição da informação e transformou a comunicação contemporânea, possibilitando a comunicação massiva e a interatividade. Hoje, é possível que pessoas separadas geograficamente, mas que possuam traços e vontades semelhantes, se conectem, dando origem a um novo conceito: as comunidades virtuais (Rheingold 2001). Dentre os diversos tipos de comunidade virtual, merecem destaque os ambientes para educação à distância, como por exemplo, O AmCorA (Menezes 2000).

As comunidades virtuais oferecem diversas formas de troca de informações entre os seus participantes. Há os fóruns, também conhecidos como *Usenet Newsgroup*, que armazenam todas as discussões decorridas sobre um determinado tema durante a sua existência, dificultando a localização eficiente de mensagens sobre um determinado tópico dentro do tema discutido. Uma variação dos fóruns são os encontros eletrônicos, onde os seus participantes geram uma grande quantidade de documentos em um determinado intervalo de tempo (Roussinov 1999).

Um outro serviço encontrado é o de esclarecimentos, por meio de perguntas e respostas, que se torna mais eficiente quando as questões previamente respondidas são acessadas rapidamente, evitando, assim,

que perguntas (e respostas) sejam repetidas indefinidamente, aumentando, ainda mais o volume de informação disponível. Um exemplo deste serviço é o Qsabe (Pessoa 1997), um ambiente inteligente para roteamento de perguntas, utilizado pelo Amcora.

Há, também, recursos voltados para a comunidade científica, onde *sites* tais como o Citeseer (Bollacker 1998) (http://www.citeseer.com), a biblioteca digital da ACM (http://www.acm.org/dl), a biblioteca da CAPES (http://www.periodicos.capes.gov.br) e outros, disponibilizam artigos sobre diversas áreas de estudo, armazenando, portanto, um grande volume de informação, que se não estiver devidamente organizada, torna impraticável a utilização desses serviços.

Os motivos que levam a criação de uma comunidade virtual são inúmeros e as formas de socialização entre seus membros são diversas, mas todas giram em torno de um único objeto: a informação. Sendo assim, tais comunidades sofrem com a desorganização das informações disponíveis, que é proporcional, de forma não linear, ao número de participantes ativos.

Uma solução para o problema descrito anteriormente é organizar a informação em diretórios ou categorias. Neste artigo, é apresentado um ambiente de apoio à criação e manutenção semi-automática de categorias, que utiliza como ponto de partida coleções de documentos sobre os temas a serem organizados. Para alcançar este objetivo, foram utilizados *self-organizng maps* (SOM), que possibilitam o agrupamento automático de documentos e o algoritmo de Rocchio, utilizado para ajustar a definição das categorias encontradas.

Na seção 2, será apresentada uma revisão bibliográfica sobre a geração automática de categorias. Na seção 3, discutiremos os algoritmos utilizados. Na seção 4, apresentamos a proposta de um ambiente. A seção 5 apresenta experimentos com o protótipo construído e na seção 6, são apresentadas as considerações finais.

2. Trabalhos Correlatos

A organização de documentos em diretórios é bastante discutida na literatura, havendo diversas finalidades e formas de criação das categorias.

No Yahoo! (Labrou 1999), um portal de pesquisa da Internet, que apresenta documentos organizados em diretórios, a classificação de novos documentos e a criação de categorias é feita manualmente. O fato de novos documentos serem classificados por seres-humanos faz com que a definição de um diretório apenas pelo seu nome seja suficiente. Este tipo de abordagem inviabiliza a automatização da classificação de documentos e dificulta o trabalho dos responsáveis pela manutenção dos diretórios.

O CORA (McCallum 2000), um portal para pesquisa de artigos na área da ciência da computação, tem a organização de seus diretórios também feita de forma manual, porém como a classificação de documentos é feita de forma automática, seus diretórios recebem uma definição, que é criada por meio de um conjunto de treinamento, algumas palavras-chave, atribuídas manualmente às categorias, e por um algoritmo de treinamento (o algoritmo Bootstrapping), que refina a definição das categorias.

Há, também, casos onde ocorre a automatização não somente da definição das categorias, como também da sua identificação. É o caso do algoritmo apresentado em (Han 2000). Nele, não há qualquer indicação sobre os diretórios a serem criados. As categorias surgem à medida que os documentos são agrupados e são definidas por meio do *centroid* dos documentos que as compõem, que é modificado a cada vez que um novo documento é associado ao diretório durante a fase de treinamento.

Em (Goren-Bar 2000), os autores relatam experimentos realizados com duas formas de definição automática de categorias e classificação automática de documentos. Em primeiro lugar, uma coleção de documentos foi gerada, sendo a sua classificação feita manualmente. Depois, estes exemplos foram apresentados a dois tipos de rede neural e os resultados comparados com a abordagem manual. Em ambos os casos, as categorias não foram criadas inicialmente, mas resultaram dos agrupamentos gerados. No primeiro tipo, o SOM, os exemplos foram apresentados, o treinamento foi realizado e, então, as classificações realizadas. O segundo tipo, uma rede neural do tipo LVQ (*Learning Vector Quantization*), também sofreu os mesmos processos, mas por ser uma rede de aprendizado supervisionado, na fase de treinamento, foram informados os resultados desejados para cada entrada. Após o treinamento, cada rede possuía um conjunto de categorias (agrupamentos), definidas por meio da rede neural, através da qual a classificação de novos documentos era realizada.

Um outro exemplo, é encontrado em (Aggarwal 1999). Nele, é apresentado um método supervisionado para criar categorias para classificação de documentos. O algoritmo utilizado é conhecido como *projected clustering*. Os agrupamentos são definidos inicialmente baseados em uma taxonomia preexistente e, a cada iteração, a dimensionalidade projetada dos agrupamentos é reduzida (alguns elementos do vetor de representação são ajustados em zero) e agrupamentos com similaridades maiores que um limiar são reunidos. No final do processo, os agrupamentos resultantes se tornam categorias, cuja definição é dada pelos *centroid* dos documentos nelas agrupados.

3. Referencial Teórico

Nossa proposta, a ser apresentada em seções subseqüentes, se baseia em 2 técnicas. A primeira delas, denominada SOM, é utilizada para agrupar os termos semelhantes, permitindo assim a criação de categorias. A segunda, um algoritmo de aprendizagem, é utilizada para fazer sintonia fina no sistema, a partir de *feedback* do usuário. A seguir, fazemos uma apresentação das técnicas, visando preparar o leitor para um entendimento apropriado da solução proposta.

3.1. O Self-organizing Map (SOM)

O *Self-Organizing Map* (SOM), desenvolvido por Telvo Kohonen, na década de 80 (Kohonen 1982, 1995), é um modelo de rede neural do tipo feedforward e de aprendizado não supervisionado (Lin 1991). Também conhecido como Rede de Kohonen, o SOM é capaz de arranjar um espaço de alta dimensão, por exemplo, os termos de uma coleção de documentos, em um espaço de dimensão menor, por exemplo, os assuntos tratados nessa coleção. A figura 1 (Roussinov 1998), apresenta o SOM, com duas camadas.

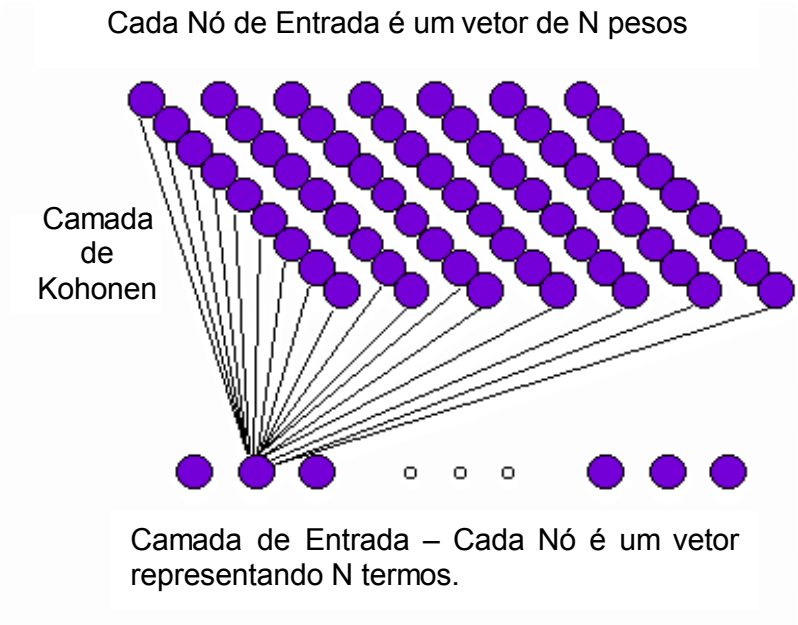

Figura 1 – Rede de Kohonen para organização de termos

O SOM funciona da seguinte forma: os pesos da rede são inicializados aleatoriamente, o sinal de entrada (que pode ser, por exemplo, uma representação para um documento) é fornecido sem que se indique a saída desejada. De acordo com este sinal, um neurônio de saída responderá melhor; e será o neurônio vencedor. Assim, o neurônio vencedor e seus vizinhos têm seus pesos ajustados de forma a responder melhor do que antes à entrada apresentada. O ajuste dos pesos faz com que sinais semelhantes obtenham a melhor resposta de neurônios vizinhos. Sendo assim, sinais semelhantes apresentados ao SOM, gerarão um agrupamento em determinada região da rede.

3.2. Algoritmo de Treinamento

Um tipo especial de algoritmo, são os algoritmos de treinamento, capazes de alterar a definição de categorias (vetores de características) ao longo de sua utilização, a fim de possibilitar que os métodos de classificação obtenham resultados mais precisos, utilizando, para isso, dados de treinamento. Tais algoritmos podem ser divididos em dois tipos: algoritmos on-line e de lote.

Os algoritmos on-line têm um exemplo de treinamento apresentado a cada vez. Eles atualizam os vetores de pesos baseados no exemplo apresentado, e o descartam, permanecendo apenas com o vetor atualizado. Os algoritmos de lote, por outro lado, otimizam seus vetores de pesos baseados em um conjunto completo de dados de treinamento de uma só vez (Uden 2002). Dentre os algoritmos existentes, destacamos o algoritmo de Rocchio, discutido a seguir.

3.2.1. Algoritmo de Rocchio

O algoritmo de Rocchio (Rocchio 1971, Harman 1992) é um algoritmo de lote, que produz um novo vetor de pesos w a partir de um vetor de pesos existente w_1 e de um conjunto de exemplos de treinamento. O j^{th} componente w_j do novo vetor de componentes é (Lewis 1996):

$$w_j = \alpha w_{1,j} + \beta \frac{\sum_{i \in C} x_{i,j}}{n_C} - \gamma \frac{\sum_{i \notin C} x_{i,j}}{n - n_C}$$

Onde n é o número de exemplos de treinamento, $C = \{1 \le i \le n : y_i = 1\}$ é o conjunto de exemplos de treinamento positivos e n_C é número de exemplos positivos de treinamento. Os parâmetros α, β eϒ controlam o impacto relativo do vetor de pesos original, dos exemplos positivos e dos negativos respectivamente.

Para utilizar o algoritmo de Rocchio na alteração da definição de categorias, basta realizar a classificação de um conjunto de documentos nessas categorias e solicitar o *feedback* do usuário. Assim, as definições das categorias serão os vetores a serem modificados e a coleção de documentos, os exemplos de treinamentos. Daí, aplicando-se a equação anterior, de acordo com o *feedback* do usuário, serão obtidas definições mais apuradas para as categorias envolvidas.

4. GECA – um Ambiente para Apoiar a Criação de Categorias

A materialização da proposta se deu através da especificação e implementação de um ambiente computacional, satisfazendo a arquitetura descrita na subseção 4.1. a partir da qual foi construído um protótipo apresentado na subseção 4.2

4.1 Arquitetura do Sistema

O sistema atuará em duas fases: a primeira será responsável pela geração de agrupamentos e identificação de seus termos relevantes, e os seus respectivos pesos, que darão origem aos assuntos e, a outra, é a fase de atualização da definição dos assuntos na medida em que novos documentos forem a eles apresentados. Na primeira fase utiliza-se o *Self -Organizing Map* (SOM) e a segunda é baseada no algoritmo de Rocchio.

A arquitetura proposta contempla um módulo para a construção semi-automática de categoria (assuntos*)*, um módulo para a atualização semi-automática de categoria (assuntos*)*, além de um ambiente de autoria. A figura 2 apresenta a arquitetura do Ambiente Construtor de Categorias. A seguir, fazemos uma descrição detalhada de seus componentes.

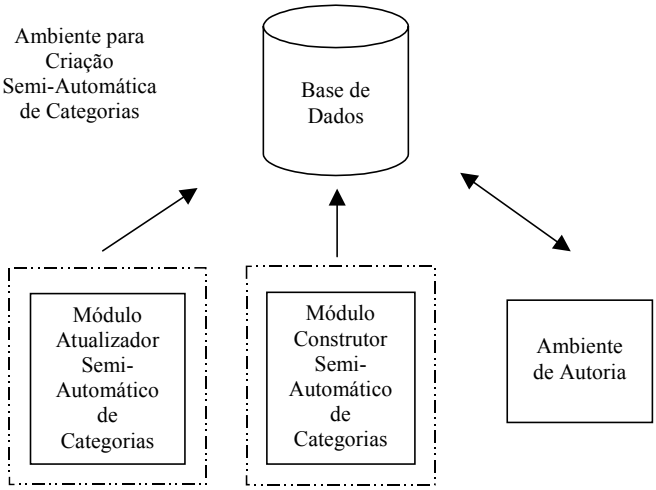

Figura 2 – Arquitetura do Ambiente Construtor de Categorias

Construtor de Semi-automático de Categorias. Este módulo é o responsável por agrupar documentos semelhantes pertencentes a uma coleção de documentos. Nele, os assuntos e seus termos relevantes são definidos, baseados nos agrupamentos gerados. As suas funcionalidades são as seguintes:

- Indexação automática dos documentos de uma coleção;
- Listagem dos índices/palavras de uma coleção de documentos;
- Listagem dos documentos da coleção;
- Geração automática de agrupamentos para uma coleção de documentos;
- Exibição gráfica dos agrupamentos gerados para uma dada coleção;
- Listagem dos documentos correspondentes a cada agrupamento;
- Identificação automática e listagem dos termos relevantes para um agrupamento.

Módulo Atualizador Semi-automático de Categorias. Este módulo é responsável por atualizar as definições das categorias (assuntos) a partir de novos documentos incluídos na coleção. As suas funcionalidades são as seguintes:

- Classificação dos documentos incluídos de acordo com os assuntos já existentes;
- Atualização da definição do assunto a partir dos termos relevantes encontrados no documento e do *feedback* do usuário em relação a classificação obtida para tal documento.

Ambiente de autoria. É nesse ambiente, que o usuário fornece uma coleção de documentos sobre um tema e tem a possibilidade de rotular os assuntos (dar nomes) e manipular os termos relevantes para a definição dos mesmos. As suas funcionalidades são as seguintes:

- Inserção de coleções de documentos sobre os temas desejados para que estes sirvam como base para a identificação de assuntos pertencentes a tais temas;
- Inclusão e exclusão de documentos de uma coleção;
- Exclusão de índices/palavras relevantes de uma coleção de documentos;
- Exclusão de palavras relevantes de um dado agrupamento;
- Exclusão de um agrupamento;
- Rotular um agrupamento dando origem a um assunto.

4.2 Implementação

O sistema foi desenvolvido na ferramenta Delphi 6.0, cuja linguagem é o pascal object, e adotou-se, como banco de dados o Interbase. O paradigma adotado foi o orientado a objetos e o sistema foi desenvolvido utilizando-se três camadas: Interface com usuário, regras de negócio e banco de dados.

A arquitetura interna do sistema é apresentada na Figura 3.

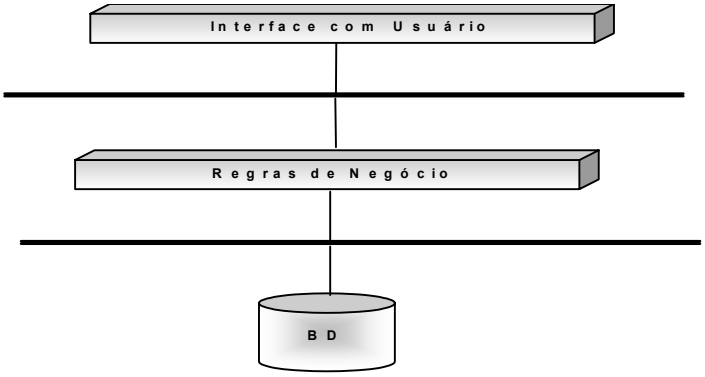

Figura 3 – Arquitetura interna do Sistema

Na versão corrente foram implementadas as seguintes funcionalidades:

- Criar coleção
- Inserir documentos em uma coleção
- Editar os índices de uma coleção
- Visualizar os documentos de uma coleção
- Agrupar documentos de uma coleção
- Nomear Assuntos
- Excluir Assuntos
- Editar os índices de Assuntos
- Classificar novos documentos
- Alterar Definição de Assuntos
- Incluir Palavras Irrelevantes
- Incluir Novo Idioma

O processo para a criação das categorias segue as seguintes fases:

1. O usuário informa a coleção de documentos que resultará em categorias
2. O sistema processa cada documento, retirando as palavras irrelevantes.
3. Neste ponto, o usuário pode verificar os termos utilizados pela coleção e eliminar aqueles que desejar;
4. O peso dos termos relevantes é calculado, gerando vetores de representação para cada documento.
5. Os vetores são apresentados ao *self-organizng map*, que agrupa documentos cujos assuntos são semelhantes.

6. O usuário pode verificar os documentos presentes em um agrupamento e nomeá-lo adequadamente.
7. As categorias são, então, geradas, extraindo-se o conjunto de termos que define cada agrupamento.

Para a manutenção de categorias, os seguintes passos são executados:

1. O usuário apresenta uma nova coleção de documentos.;
2. Os documentos são classificados nas categorias geradas;
3. O usuário envia o seu *feedback*, informando se a classificação de cada documento está correta, ou não, quando a categoria correta deverá ser informada.
4. Com as informações fornecidas, o algoritmo de Rocchio é aplicado, melhorando a definição das categorias criadas.

A Figura 4, a seguir, ilustra o uso do sistema. Após o sistema agrupar as palavras, o usuário pode nomear esses agrupamentos. No exemplo apresentado destacamos 3 assuntos: Agentes, Redes Neurais e Web Search. O sistema apresenta as palavras agrupadas por assunto e o usuário pode fazer modificações através da exclusão de palavras.

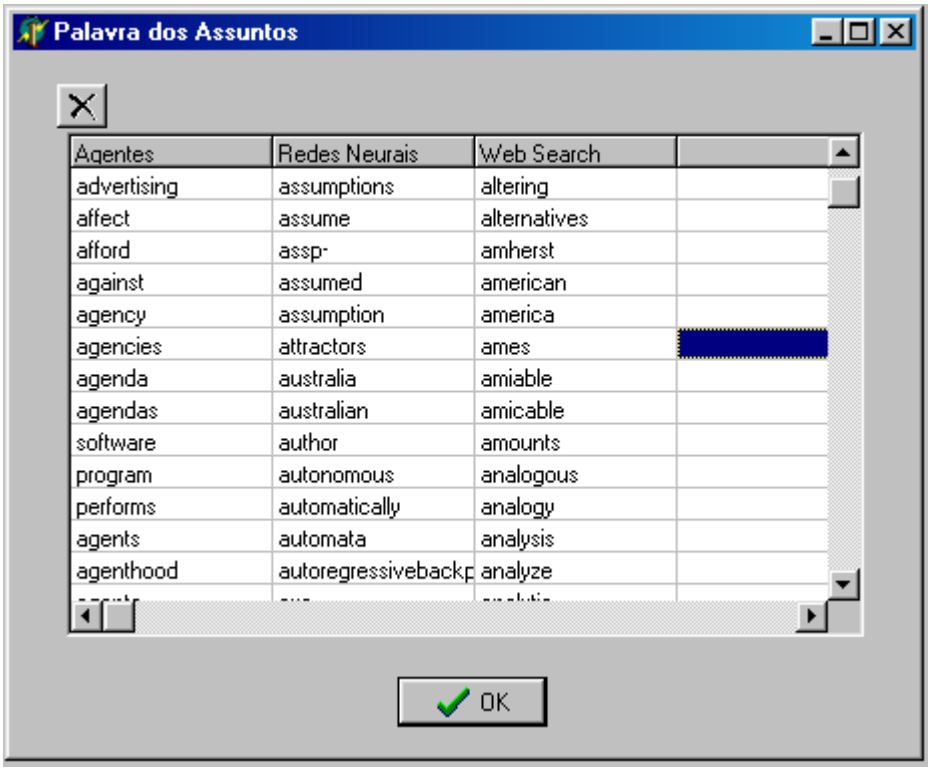

Figura 4 – Palavras dos Assuntos

5. Experimentos Realizados

Nesta seção, serão descritos alguns dos testes realizados com o GeCa, o protótipo do ambiente propostos para criação de categorias e classificação de documentos nas categorias geradas. Para testar a geração de categorias, foram feitos testes com três coleções distintas, descritos a seguir:

O primeiro teste foi realizado com documentos em português pertencentes a assuntos bastantes distintos, são eles: Gestão do Conhecimento, Direito, Informática e Medicina.

Os documentos utilizados tratam-se de artigos que foram obtidos por meio de sites específicos. Os 17 artigos sobre Gestão do conhecimento foram obtidos a partir do endereço http://www.informal.com.br/artigos/artigos.htm, os 15 artigos sobre Direito foram obtidos a partir do endereço http://www.teiajuridica.com/artigos.htm, os 18 artigos sobre Medicina foram obtidos a partir do endereço http://www.ibemol.com.br/artigos/default.asp e, por fim, os 15 artigos sobre informática foram encontrados no endereço http://www.guiadohardware.net/artigos/index.asp.

Ao se fornecer a coleção de documentos descrita acima ao GeCa e solicitar o agrupamento de documentos, os seguintes resultados foram alcançados:

No Assunto 0, foram agrupados 16 documentos sobre Gestão de Conhecimento e 1 documento sobre Direito;

No Assunto 1, foram agrupados 14 documentos sobre Direito e 4 sobre Medicina;

No Assunto 2, foram agrupados 15 documentos sobre Informática e 1 sobre Gestão do Conhecimento;

No Assunto 3, foram agrupados 14 documentos sobre Medicina.

O segundo teste foi realizado com documentos em inglês sobre a área de informática nas seguintes categorias Redes Neurais - 13 documentos, *Data Warehousing* - 6 documentos, Reconhecimento de Face - 14 documentos, Áudio - 10 documentos e Agentes - 16 documentos.

Os documentos sobre agentes, redes neurais foram retirados das primeiras páginas de documentos disponíveis nos diretórios do site da Nec Research Institute (http://citeseer.nj.nec.com/directory.html), já, os sobre Áudio, Reconhecimentos de Face e *Data warehousing* forma retirados da lista de discussão do http://www.google.com.

Ao se fornecer a coleção de documentos descrita acima ao GeCa e solicitar o agrupamento de documentos, os seguintes resultados foram alcançados:

No Assunto 0, foram agrupados 13 documentos sobre Agentes e 1 documento sobre Redes Neurais;

No Assunto 1, foram agrupados 3 documentos sobre Agentes, 1 sobre Áudio, 1 sobre Reconhecimento de Face e 2 sobre Redes Neurais. Este assunto foi descartado;

No Assunto 2, foram agrupados 7 documentos sobre Redes Neurais e 2 sobre Áudio;

No Assunto 3, foram agrupados 5 documentos sobre *Data Warehousing* e 3 sobre Redes Neurais;

No assunto 4, foram agrupados 7 documentos sobre Redes Neurais e 2 sobre Áudio;

No assunto 5, forma agrupados 9 documentos sobre Reconhecimento de Faces.

Por fim, foi realizados um último teste contendo documentos com assuntos bastante semelhantes. Os documentos tratavam de fisiculturismo e se dividiam em três categorias: Anabolizantes (14 documentos), Nutrição (15 documentos) e Treinamento (14 documentos). Tais documentos tratam-se de artigos existentes no seguinte endereço: http://www.fisioculturismo.hpg.com.br.

Ao se fornecer a coleção de documentos descrita acima ao GeCa, e solicitar o agrupamento de documentos, os seguintes resultados foram alcançados:

No Assunto 0, foram agrupados 10 documentos sobre Anabolizantes, 10 sobre Nutrição e 2 sobre Treinamento;

No Assunto 1, foram agrupados 3 documentos sobre Anabolizantes e 4 sobre Nutrição;

No Assunto 2, foram agrupados 12 documentos sobre Treinamento, 1 sobre Nutrição e 1 sobre Anabolizantes.

O sistema, portanto, não foi capaz de distinguir documentos sobre nutrição e anabolizantes, que possuem uma relação estreita.

Para realizar o teste de classificação de novos documentos em categorias geradas pelo GeCa, foram criadas categorias a partir de uma coleção contendo documentos sobre medicina, que foram encontrados

no endereço http://boasaude.uol.com.br/lib/. Foram utilizados documentos das seguintes categorias: Aids (16 documentos), Cardiologia (15 documentos), Câncer (16 documentos).

Ao se fornecer a coleção de documentos descrita acima ao GeCa e solicitar o agrupamento de documentos, os seguintes resultados foram alcançados:

No Assunto 0, foram agrupados 8 documentos sobre Cardiologia e documento sobre Aids, dando origem à categoria Cardiologia;

No Assunto 1, foram agrupados 10 documentos sobre Aids, 1 sobre Câncer e 2 sobre Cardiologia, dando origem à categoria Aids;

No Assunto 2, foram agrupados 1 documento sobre Aids e 14 sobre Câncer, dando origem à categoria Câncer;

No Assunto 3, foram agrupados 2 documentos sobre Aids, 1 sobre Câncer e 5 sobre Cardiologia. Este assunto foi descartado.

O próximo passo foi classificar novos documentos nos assuntos gerados. Para isso, utilizou-se uma nova coleção, cujos documentos foram obtidos do mesmo local. Esta coleção era composta da seguinte forma: 8 documentos sobre Aids, 7 sobre Câncer e 5 sobre Cardiologia. Após solicitar a classificação, os seguintes resultados foram obtidos: 17 documentos foram atribuídos às categorias corretas e 3 documentos, todos eles sobre Aids, forma classificados incorretamente.

Após o envio do *feedback* e a aplicação do algoritmo de Rocchio, quando as categorias foram ajustadas, a classificação desses mesmos documentos resultou em 100% de acerto.

6. Considerações Finais

Nossos estudos indicaram que a construção de um ambiente de apoio à criação de categorias é viável, sendo a sua utilização muito útil em ambientes onde o volume de documentos é considerável. Dentre esses ambientes, destacam-se aqueles para educação à distância, onde discussões sobre diversos tópicos são realizadas, serviços de perguntas e respostas são oferecidos e documentos, tais como, artigos e apostilas são disponibilizados. A organização da informação neste tipo de ambiente, além de reduzir o tempo gasto em busca do conhecimento e a duplicação de trabalho, no caso dos sistemas de esclarecimento, fornece, através dos tópicos, uma visão geral do conteúdo das informações existentes, permitindo o despertar da curiosidade de seus usuários sobre novos assuntos.

Há, no entanto, algumas melhorias que podem ser realizadas para que o ambiente apresente resultados mais satisfatórios. Em primeiro lugar, seria interessante disponibilizar esta ferramenta na Internet, aumentando, assim, as suas possibilidades de utilização. Além disso, a seleção dos termos relevantes dos documentos pode ser mais apurada, se algumas das técnicas de recuperação de informação forem aplicadas, tais como, *steamming* e a seleção de termos por freqüência. Por fim, há a possibilidade de organizar as categorias em uma hierarquia, aumentando, desta forma, a eficiência no uso desta ferramenta, uma vez que resultará em uma maior organização dos documentos, facilitando, ainda mais, o processo de busca realizado pelos usuários.

Referências

Aggarwal, C. C.; Gates, S. C.; Yu, P. S. On the merits of building categorization systems by supervised clustering. In: Proceedings of {KDD}-99, 5th ACM, 1999.

Bollacker, K. D.; Lawrence, S.; Giles, C. L. An autonomous web agent for automatic retrieval and identification of interesting publications. Proceedings of the Second International Conference on Autonomous Agents. New York. p. 116-123, 1998.

Goren-Bar, D.;KUFLIK, T.; LEV, D. Supervised learning for automatic classification of documents using self-organizing maps, Workshop: Information Seeking, Searching and Querying in Digital Libraries, 2000.

Han, E.; Karypis, G. Centroid based document classification: Anlysis and Experimental results.Principles of Data Mining and Knowledge Discovery, p. 424-431, 2000.

Harman, D.. Relevance feedback and other query modification techniques. In William B. Frakes and Ricardo Baeza-Yates, editors, Information Retrieval: Data Structures and Algorithms, p. 241-263. Prentice Hall, 1992.

Kohonen, T. Self-organized formation of topologically correct feature maps, Biological Cybernetics, v. 43, p. 59-69, 1982.

Kohonen, T. Sef-organizing maps. Springer, 1995.

Labrou, Y.; Finin, T. W. Yahoo! As an ontology: Using Yahoo! categories to describe documents. CIKM, p. 180-187, 1999.

Lewis, D. D.; Schapire, R. E.; Callan, J. Papka, R. Training algoritms for linear text classifiers. In: Proceedings of SIGIR-96, 19th ACM International Conference on Research and Development in Information Retrieval, 1996.

Lin, X.; Soergel, D.; Marchionini, G. A self-organizing semantic map for information retrieval. In: Proceedings of 14th Ann. International ACM/SIGR Conference on Research & Development in Information Retrieval, p. 262-269, 1991.

McCallum, A. K., et al. Automating de Construction of Internet Portals with Machine Learning Information Retrieval, v.2, n. 3, p. 127-163, 2000.

Menezes, C. S., Cury, D, Tavares, O., Campos, G., Castro, A., an architecture of an environment for cooperative learning (AmCorA), ICECE - International Conference on Engineering and Computer Education, São Paulo, Brasil 2000.Rocchio Jr., J. J. Relevance feedback in information retrieval. The SMART Retrieval System: Experiments in automatic document Processing, 313-32. Prentice-Hall, 1971.

Pessoa, J M.. Desenvolvimento de software orientado a agentes: uma experiência com agentes de interface. Dissertação de Mestrado (Mestrado em Informática), Universidade Federal do Espirito Santo, 1997.

Rheingold, H. The virtual community. 1994 Disponível em <http://www.rheingold.com/vc/book/>. Acesso em set. 2001.

Roussinov , D. G; Chen, H. Document clustering for electronic meetings: An experimental comparison of two techniques. Decision Support Systems,. v. 27, n. 1, p. 67-79, 1999.

Roussinov, D. G.; Chen, H. A scalable self-organizin map algorithm for textual classification: a neural network approach to thesaurus generation. ICC-AI Communication , Cognition and Artificial Intelligence, v.15, n.1-2, p. 81-111, 1998.

Uden, M. v. Rocchio: relevance feedback in learning classification algorithms, citeseer.ny.nec.com/57872.html, capturado em 10/2002.