

Fazendo uso da categorização de textos em atividades empresariais

Claudia Brandelero Rizzi^{1,2}, Leandro Krug Wives¹,
José Palazzo M. de Oliveira¹, Paulo Martins Engel¹

claudiab, wives, palazzo, engel@inf.ufrgs.br

1 - Universidade Federal do Rio Grande do Sul
(UFRGS)
Avenida Bento Gonçalves, 9500 -- Agronomia
91501-970 – Porto Alegre(RS) – Brasil

2 - Escola Estadual do Oeste do Paraná
(UNIOESTE)
Rua Universitária, 2069 – Caixa Postal 711
85814-110 - Jd. Universitário – Cascavel – Paraná

Resumo

O objetivo deste trabalho é apresentar a categorização de textos e algumas de suas aplicações, visando promover uma reflexão sobre suas funções no contexto de atividades empresariais. Na primeira parte do trabalho são feitas considerações teóricas sobre a categorização de textos, partindo de uma introdução, passando por aplicações, culminando com a apresentação de alguns softwares específicos e considerações a respeito do funcionamento genérico de tais programas. Na segunda parte são apresentados os resultados obtidos a partir da realização de um experimento em que uma rede neural do tipo Perceptron Multicamadas, treinada com o algoritmo Backpropagation, é utilizada na categorização de textos da Coleção Reuters-21578. Uma discussão final é feita considerando o uso potencial e estratégico da categorização de textos em atividades empresariais.

Palavras-chave

Categorização de Textos; Inteligência Competitiva; Disseminação de Informações;
Rede Neural; Backpropagation.

Introdução

A quantidade de dados e informações de que uma empresa necessita para suprir sua demanda de informação está diretamente relacionada com o nível de globalização de seu setor. Quanto mais disponível estiver a informação sobre sua área de atuação e seus mercados maiores são as chances de concorrentes utilizarem essas mesmas informações diminuindo, em consequência, a competitividade desta empresa. Não só a quantidade, mas também a qualidade e a facilidade de acesso à informação incitam o surgimento de novos concorrentes.

Nesse contexto, as empresas devem ser capazes de obter rapidamente (antes de seus concorrentes) a maior quantidade possível de informações relevantes. Para tanto, devem ser utilizadas técnicas de coleta de informação capazes de selecionar somente as informações pertinentes às empresas, filtrando as irrelevantes. Essas informações devem ser distribuídas ou repassadas a pessoas específicas da empresa, que expressaram sua necessidade e que sabem dar utilidade a ela, transformando-a em ação, produção e competitividade. Quanto melhor o acesso, a disponibilidade e a qualidade da informação, maior é a chance de acerto no processo de tomada de decisão e maior a competitividade das organizações capazes de utilizá-la.

Disponer e armazenar informação não são suficientes para manter a liderança em um mercado. A informação não é uma vantagem competitiva sustentável, já que todos possuem, em teoria, a possibilidade de obter as mesmas informações. Logo, esta deve ser transformada em ação rapidamente, geralmente com o desenvolvimento de uma nova tecnologia ou estratégia de ação. Esta nova tecnologia ou estratégia também não é sustentável como fator diferencial, segundo Alan Webber (editor da revista *Fast Company*) [1], já que, com o tempo, todos os concorrentes adotam a mesma tecnologia ou estratégia e ela acaba tornando-se um fator básico para a operação do setor.

Com isso, o ambiente externo à empresa deve ser constantemente monitorado para que uma empresa possa manter-se competitiva no mercado. Monitorar, localizar e adquirir informação no ambiente externo são atividades custosas. Assim, encontrar

informação relevante no universo da informação existente tem se mostrado uma área fértil de pesquisa. Uma das técnicas que pode ser utilizada para melhorar a qualidade dos serviços e sistemas de informações das empresas é a categorização de textos.

A categorização de textos é uma técnica utilizada para classificar um conjunto de documentos em uma ou mais categorias existentes. Ela é geralmente utilizada para classificar mensagens, notícias, resumos e publicações. A categorização também pode ser utilizada para organizar e filtrar informações. Essa capacidade faz com que esta técnica possa ser aplicada em empresas, contribuindo no processo de coleta, análise e distribuição de informações e, conseqüentemente, na gestão e na estratégia competitiva de uma empresa.

O objetivo deste trabalho é apresentar a categorização de textos e algumas de suas aplicações visando promover uma reflexão sobre suas funções. Neste sentido, são apresentados alguns exemplos que visam contribuir e ilustrar essas reflexões. O que se pretende é que, desta reflexão, possam emergir idéias e possíveis soluções para auxiliar na atividade empresarial.

1 Categorização de informações textuais

Grande parte das informações externas de que uma empresa necessita para satisfazer sua necessidade de informação (especialmente decorrente da competitividade) é encontrada em formato textual.

Esse tipo de informação é extremamente rico e complexo, o que dificulta sua análise computacional e até mesmo humana. A linguagem natural utilizada nos textos para expressar idéias é cheia de ambigüidades e de figuras de linguagem, exigindo muito da capacidade de interpretação humana. Isso, aliado a grande quantidade de dados, torna impossível a análise do conjunto da informação que uma empresa tem a disposição.

A categorização de textos é um ramo da computação (mais corretamente, uma técnica de descoberta de conhecimento) cujo objetivo é o de estudar métodos para que os textos possam ser manipulados com eficiência por computadores. O que a categorização faz, na verdade, é classificar documentos em relação a um conjunto de categorias

predefinidas. Essa técnica tem várias aplicações. O enfoque deste trabalho está centrado na discussão destas diferentes aplicações.

1.1 Algumas aplicações da categorização de textos

Seguindo uma sistemática proposta por Paul Jacobs [2], a categorização de textos pode ser utilizada para a *disseminação*, *recuperação* de informação e para a *navegação* na estrutura de conhecimento.

a) **Disseminação**: compreende o envio de textos para determinados usuários. Nesse caso, os funcionários de uma empresa definem suas necessidades de informação e o sistema, ao receber documentos externos ou internos, analisa-os e envia os selecionados para os usuários que solicitaram esta classe de dados. Os sistemas que utilizam essa técnica são denominados de sistemas de roteamento (routing), ou sistemas de recomendação [3]. Na disseminação, o sistema recebe constantemente informações e as repassa para os usuários (geralmente neste caso o volume de informações é pequeno e momentâneo e as informações mudam constantemente);

b) **Recuperação**: envolve as tarefas de obtenção de documentos que possam ser de interesse do usuário. A diferença principal dessa para a aplicação anterior é que o usuário deve acioná-la para que possa receber a informação de que necessita. Uma vez recuperada a informação relevante, o processo é finalizado. Os sistemas de recuperação de informação lidam com grandes volumes de dados, acumulados ao longo do tempo. A dificuldade portanto, consiste em localizar a informação desejada neste universo. Na gestão de documentos, a principal aplicação da categorização de textos dá-se na indexação de documentos [5] onde são identificadas as palavras ou frases que melhor representam o conteúdo de cada documento. Uma vez identificado o conteúdo de um documento, ele é armazenado em uma categoria correspondente ao seu conteúdo. O processo de recuperação dá-se pela recuperação de todos os documentos pertencentes à determinada categoria;

c) **Navegação**: A navegação (*browsing*) compreende a ação de organizar assuntos hierarquicamente, permitindo que o usuário navegue por essa hierarquia até

encontrar a informação de que necessita [6]. Esta organização é válida tanto para quem procura por informação quanto para quem deseja armazená-la. Neste caso, mais uma vez categorias (e provavelmente subcategorias) são definidas, e ligadas a elas estão os textos cujos temas estão relacionados. Um bom exemplo deste tipo de aplicação é o sistema Yahoo [7]. As categorias na hierarquia do Yahoo são definidas por especialistas humanos. Os textos são conectados a estas categorias por *links*. Assim, quando se quer consultar por exemplo, temas relacionados à venda de eletrodomésticos, deve-se inicialmente, acessar a categoria principal "Negócios e Economia". Após, deve-se ir selecionando as subcategorias que estejam mais de acordo com a classe de informação desejada, até chegar no nível de especificidade desejado. Assim, são percebidos dois momentos: busca e classificação. O primeiro caracteriza-se pela atividade em que o usuário está consultando algum assunto de interesse acessa *links* e *sub-links*. O segundo é aquele em que existe um documento que deve ser catalogado em determinado tema e necessita passar por um processo de triagem para que sejam definidas sua pertinência e sua relação com as categorias possíveis.

1.2 Alguns sistemas de categorização

Existem vários sistemas de cunho acadêmico que utilizam como base a técnica de categorização de textos. Analisando estes sistemas é possível desenvolver um modelo sobre como a categorização pode ser utilizada para suprir as necessidades de informação de uma empresa.

Um destes sistemas é o PHOAKS (*People Helping One Another Know Stuff*) [8], que utiliza a abordagem de filtragem colaborativa a respeito de mensagens do tipo *newsgroups* (grupo de discussão) da *Usenet*. A premissa do sistema é que os *newsgroups* são populadas por pessoas que recomendam recursos (objetos, páginas WEB, etc.) para outras pessoas. Essas recomendações nada mais são do que análises sobre o referido recurso indicando sua utilidade. Para tanto, o sistema vasculha as mensagens em busca de URLs (*links*). Quando uma URL é encontrada o sistema realiza uma série de testes predefinidos que identificam se ela faz parte de uma recomendação ou não (segundo os

autores ela obtém sucesso em 90% dos casos). Quanto mais mensagens recomendarem o mesmo URL, mais significantes elas são para o sistema. Naturalmente as recomendações devem ser de pessoas distintas para terem mais valor.

Um sistema muito parecido com o PHOAKS é o *GroupLens*. O *GroupLens* também é um sistema que analisa informações de *newsgroups* e faz filtragem colaborativa, porém seu objetivo é um pouco diferente. Os grupos de discussão movimentam diariamente uma quantidade muito grande de informações. Nesse contexto, torna-se muito difícil selecionar as mensagens de interesse. Um usuário avançado, por exemplo, pode não estar interessado em mensagens cujo tópico seja muito básico. Da mesma forma, um usuário inexperiente pode não estar interessado em mensagens muito complexas. Para resolver esses problemas o *GroupLens* cria grupos de usuários (de nível ou interesse similar) e pede para que eles atribuam uma nota entre zero (ruim) e cinco (excelente) para cada mensagem que recebem. Com o passar do tempo, o sistema consegue identificar as mensagens mais relevantes para grupo de usuários, colocando ao lado do cabeçalho da mensagem um valor correspondente ao quanto ela pode ser interessante para aquele grupo [9].

Outros sistemas, um pouco mais avançados, mas muito similares na aplicação são o FAB (que combina recomendação colaborativa com a baseada em conteúdo) [10], o *Referral Web* (que identifica co-relacionamentos entre pessoas e consegue recomendar quais os que mais conhecem sobre determinado tema ou assunto) [11].

Um sistema cuja aplicação é bem voltada a empresas é o *Grapevine*. O *Grapevine* é um *software* capaz de indexar, priorizar e disseminar informações de forma seletiva. Ele armazena todos os *e-mails*, mensagens de *newsgroups*, idéias e opiniões de uma empresa; classifica-os, prioriza e envia para as pessoas interessadas. Seu funcionamento é muito semelhante ao do PHOAKS, pois existem pessoas encarregadas de cada categoria (os assuntos) que devem avaliar as mensagens que são ali classificadas. Os outros usuários estabelecem um nível mínimo de importância ou prioridade para cada uma das categorias de assuntos existentes. Estes usuários só recebem as mensagens se as mesmas receberem um grau de prioridade igual ou maior do que o estabelecido. Assim,

determinado funcionário de um setor recebe todas as mensagens relativas a seu setor (o nível de prioridade deve ser mínimo), sem ter que ler todas as mensagens da empresa (mas apenas que lhe são relevantes). Somente quando uma mensagem pertencente a outro setor recebe uma prioridade muito alta, todos da empresa a recebem.

Outro sistema também voltado a empresas é o NLDB. Seu principal objetivo é auxiliar seus usuários no tratamento de informações em processos de disseminação, recuperação e navegação. O NLDB é formado por um mecanismo de categorização automática, uma base de dados segmentada com respectivas categorias, e uma interface para navegação e recuperação de informações. Seu funcionamento está baseado nas definições de preferências feitas pelos usuários (em *profiles*). Quando um documento trata de algum dos temas de interesse do usuário, este documento lhe é enviado via mensagem eletrônica, efetivando a disseminação. O componente de recuperação permite que o usuário elabore sua pesquisa informando palavras-chave relacionadas com seu interesse particular. O componente de navegação mostra ao usuário, de forma gráfica e hierárquica, grupos de temas relacionados em função de suas preferências. Assim, o próprio usuário pode consultar ou não esses grupos [2]. A grande vantagem deste tipo de ferramenta é a possibilidade do usuário definir e monitorar, a qualquer momento, seus interesses por informação.

Todos esses sistemas de categorização de textos têm em comum a possibilidade de permitirem um maior dinamismo na circulação de informações, internas ou externas. Em se tratando de empresas, este dinamismo pode contribuir de forma importante, especialmente em processos de tomada de decisão.

2 Metodologia dos sistemas de categorização de textos

Quando se pretende fazer uso ou desenvolver um Sistema de Categorização de Textos (SCT), é importante compreender ou definir seu processo interno de funcionamento. Nesta seção, são apontados alguns dos principais sub-processos que compõem um SCT, supondo uma implementação do tipo estatística. A FIGURA 1 ilustra estes sub-processos. Uma abordagem estatística compreende a adoção de métodos para

análise automática de conteúdos que incluem seleção e contagem de termos nos textos. Uma abordagem lingüística é aquela em que esta análise é feita aplicando-se métodos semânticos e sintáticos de tratamento de linguagem natural.

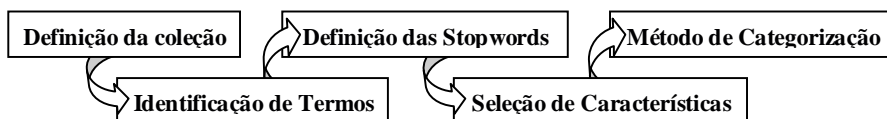


FIGURA 1 - SUB-PROCESSOS DE CATEGORIZAÇÃO

a) **Definição da Coleção de Textos**: definir a coleção de textos que será categorizada é de suma importância no processo. Esta coleção pode ser estática ou dinâmica, ou ainda, uma combinação de ambas;

b) **Identificação de Termos**: visa normalizar os textos de entrada [12], a exemplo de converter uma língua estrangeira em *Unicode*, excluir dos textos de entrada símbolos, figuras, caracteres especiais outros, com fins de facilitar a análise de conteúdos;

c) **Definição da lista de "Stopwords"**: são chamadas de *stopwords* aquelas palavras que contribuem pouco no significado geral de um texto, tais como artigos, preposições e advérbios. Estas palavras são bastante frequentes e sua eliminação pode prover uma redução de 40 a 50% dos textos dos documentos a serem analisados [13];

d) **Seleção de Termos ou Características**: visa identificar, aplicando-se métodos específicos (tais como frequência de termos, informação mútua, semântica latente, entre outros [14]) os termos que melhor representam o conteúdo dos textos [15]. Assim, pode-se também definir um conjunto de termos que melhor representam uma determinada categoria, ou um conjunto delas;

e) **Método de Categorização**: visa determinar para cada documento, uma ou mais categorias a que pertence, aplicando-se métodos (tais como árvores de decisão, modelos de regressão, *fuzzy*, entre outros [14]). Estes métodos diferem entre si pela abordagem com que realizam a categorização. Em [16] há uma excelente avaliação de um significativo conjunto desses métodos.

Na prática, como indicado por autores como Kowalski [12], podem ser utilizados outros sub-processos, tal como a redução de variações morfológicas através do uso de radicais (*stemming*) ou pelo uso de *thesaurus*. Porém, o mais importante é identificar as necessidades de informação, as bases de dados (fontes) que serão utilizadas e suas possibilidades de acesso. Feito isso, a inclusão ou retirada de algum desses sub-processos é uma consequência.

A principal particularidade, associada a uma maior dificuldade, na categorização de textos está na grande quantidade de termos a serem processados: a alta dimensionalidade do espaço de características [5]. Este espaço de características constitui-se de termos únicos ou compostos que são extraídos ou adaptados dos textos processados (ação realizada nos sub-processos *b*, *c* e *d*). O fato é que podem existir centenas ou milhares destes termos (até mesmo para uma coleção de textos de tamanho relativamente pequeno).

Assim, um bom STC é aquele que consegue lidar com este espaço de características, conseguindo reduzi-lo ao máximo sem sacrificar a identificação dos conteúdos dos documentos, o que implica no desempenho do processo de categorização.

Dentre as muitas técnicas possíveis a serem empregadas no processo efetivo da categorização, uma delas é a utilização de redes neurais. O experimento relatado neste artigo trata de um processo de categorização utilizando redes neurais. A próxima seção descreve este experimento e inicia com alguns conceitos básicos inerentes a redes neurais artificiais.

3 Classificação de textos utilizando RNA

As Redes Neurais Artificiais (RNA) são sistemas computacionais constituídos por neurônios artificiais, de implementação em hardware ou software, cujo funcionamento originalmente baseou-se naquele do sistema nervoso biológico.

O primeiro modelo matemático para uma rede neural foi proposto por McCulloch e Pitts, em 1943. Este modelo tratou o cérebro como um organismo computacional. A

FIGURA 2 mostra um neurônio artificial padrão, uma generalização do modelo de McCulloch e Pitts.

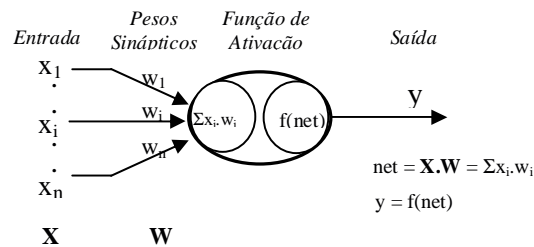


FIGURA 2 - NEURÔNIO ARTIFICIAL PADRÃO

As RNAs utilizam neurônios artificiais interconectados, que recebem informações de entrada ($x_1, \dots, x_i, \dots, x_n$) através de sensores ou de outros neurônios, executam operações sobre elas e as enviam, como saída (y), para outros neurônios ou estruturas responsáveis pela conclusão da operação. Cada conexão entre neurônios, também chamada de sinapse, possui uma intensidade associada, expressa por um valor numérico denominado peso ($w_1, \dots, w_i, \dots, w_n$), que pode ser modificado.

Cada neurônio artificial determina um valor de entrada (net) através da soma dos produtos dos valores de entrada pelos valores de peso [17]. Uma vez calculado este valor de entrada, ele se transforma no *valor de ativação* do respectivo neurônio. Este valor de ativação é uma função decorrente dos valores de entrada ($y=f(net)$). A ação da função de ativação, seja ela contínua ou descontínua, provoca uma reação ou saída de um neurônio em termos do nível de suas atividades internas, quando um certo *valor de limiar* é atingido.

Maiores considerações a respeito de algoritmos de aprendizagem, arquitetura de redes neurais, aplicações e funcionamento são encontradas em literatura específica, e nos clássicos [18] e [17]. A próxima seção apresenta as características gerais do experimento realizado com a rede neural.

3.1 Um experimento de classificação

Um experimento foi realizado para demonstrar que uma RNA pode ser utilizada para classificar documentos e ser, portanto, de alguma forma útil para empresas. Para esse experimento foi utilizado um *corpus* (base de documentos) desenvolvido especificamente para fins de comparação e avaliação de técnicas de categorização. O corpus utilizado foi a coleção Reuters-21578 [19]. Esta coleção contém 21.578 textos de notícias divulgadas pela agência Reuters no período entre 1987 e 1991. Nessa coleção os documentos estão organizados cronologicamente em cinco grupos (*topics, places, people, orgs, e exchanges*) e suas respectivas sub-categorias.

Seguindo as etapas do processo de categorização proposto na seção 2 deste artigo, necessitou-se de um *software* capaz de analisar os documentos, retirar os termos irrelevantes (*stopwords*) e indicar aqueles mais relevantes para a representação de cada classe. Para isso, desenvolveu-se a ferramenta *Analyzer*, que foi implementada a partir do sistema *Eurekha* [15]. Com ela foi realizada a eliminação de *stopwords* e seleção das características representantes de cada uma das classes, através do método "Termo Seleção" [14].

A lista de 266 *stopwords* utilizada foi aquela elaborada pelo Laboratório de Recuperação de Informações da Universidade de Massachusetts em Amherst, que pode ser obtida no trabalho de [5].

O algoritmo utilizado para conversão de termos em seus respectivos radicais (*stemming*) foi o de Martin Porter. Seu algoritmo realiza um processo de remoção de letras do final de palavras da língua Inglesa, que possuem mesma variação morfológica e de flexão. Em [20] há uma descrição completa do algoritmo, bem como suas regras e etapas de processamento.

O método de avaliação dos resultados utilizado foi aquele proposto por David Lewis [21]. Ele sugere um modelo em que uma "tabela de contingência" é inicialmente preenchida. A organização desta tabela parte do princípio de que um sistema realiza n decisões binárias, sendo que cada uma das quais tem uma resposta correta, ou sim ou não.

A partir desta tabela, podem ser extraídas importantes medidas de eficácia para STCs como *Abrangência (ou Revocação)*, *Precisão*, *Falha*, *Acurácia* e *Erro*.

A rede neural implementada é uma *Perceptron* Multicamadas, com três camadas e alimentação para frente. O aprendizado supervisionado foi realizado com o algoritmo *Backpropagation*, tendo como função de ativação a tangente hiperbólica. A próxima seção descreve a sistemática de organização das entradas, utilizada no experimento.

3.2 Organização dos termos

Uma vez selecionadas as características (os termos) que melhor representam cada uma das categorias da coleção de documentos, elas foram numeradas em ordem crescente e dispostas de forma seqüencial. Esta relação de termos resultante tornou-se o *vetor base*.

Um *vetor base* de tamanho c_n representa, em cada uma de suas posições, uma das características selecionadas. Quando um *vetor de treinamento* ou *de teste* é construído, este *vetor base* é consultado para que sejam obtidas as posições correspondentes que representam as características desejadas.

Deste modo, tanto o *vetor base* quanto os vetores *de treinamento* e *de teste* utilizados pela rede, são formados por um número n de posições. Estas posições já definem o número de neurônios da camada de entrada da rede. Por exemplo, neste experimento em que existem 6 categorias, e decidiu-se por representar cada categoria com suas 20 características mais importantes, totalizando 120 características ($n = 120$). Uma vez concluída a etapa de construção dos vetores de treinamento e de testes, tomando como referência o vetor base da Coleção, a rede já pode ser treinada.

3.3 Treinamento e teste da RNA

Para explicar o processo de treinamento da rede, um vetor de treinamento pertencente à categoria 1 é tomado como exemplo. Este vetor foi construído a partir da leitura de um dos textos do conjunto de treinamento, que pertence a esta categoria.

Este vetor de treinamento é apresentado à rede juntamente com a informação de que ele representa a categoria 1. O algoritmo *Backpropagation* encarrega-se de receber estas entradas, assim como todas as outras entradas que representam as demais categorias, e as processa. Ele também calcula os valores de erro e ajusta os valores dos pesos. Este processo se repete até que uma taxa de erro desejada seja alcançada. Quando isto ocorre, a rede é dita treinada e o processo de teste (a classificação em si) pode ser iniciado.

O processo de teste da rede é exemplificado tomando-se um vetor de teste que, ao contrário do vetor de treinamento, não contém informação a respeito de qual categoria pertence (pois é exatamente isso que se deseja descobrir). Da mesma forma que o vetor de treinamento, este vetor de teste é construído a partir da leitura de um dos textos que se deseja categorizar.

Deste texto, também são extraídas as características nele encontradas. Para isso, identificam-se os termos, são eliminadas as *stopwords* e, finalmente, são extraídas suas características. No entanto, as características consideradas como válidas são apenas aquelas que correspondem às definidas como representantes da área a categorizar e que compõem o vetor base. Do exemplo, apenas 120 termos são considerados válidos.

Na fase de seleção de características do texto a ser categorizado, os termos são comparados com a relação dos 120 termos que representam a área a ser categorizada, e que compõem o vetor base. Em existindo equivalência, o vetor base que possui esta correspondência, informa as posições em que uma marca “+1” deve ser colocada no vetor de teste, indicando a existência daquelas características. Igualmente, uma marca “0” é colocada nas demais posições do vetor, indicando o contrário.

Este vetor de teste é apresentado à rede já treinada. Ela o processa e emite como saída uma lista com 6 valores reais, indicando o "grau de pertinência" daquele vetor (portando daquele documento que ele representa) em relação às seis categorias existentes.

O maior valor real apresentado (que varia entre -1 e 1), e que necessariamente deve atingir um valor mínimo de $0,5$, indica que o documento pertence à categoria que lhe é correspondente. A opção pelo limiar $0,5$ permite que se tenha no mínimo este valor de segurança para aceitação do rótulo de classe.

3.4 Resultados do experimento

Os textos extraídos da Reuters-21578 foram aqueles cujo tema estava relacionado com metais (mais especificamente, com alumínio, cobre, ouro, ferro, prata e zinco). Ao todo, 6 categorias e 312 textos foram selecionados para a realização do experimento. Este conjunto foi chamado de “Sub-Coleção Metais”.

Dos 312 textos disponíveis, foram utilizados 107 para treinamento (além de vetores heurísticos) e 205 para testes. A taxa de aprendizagem utilizada foi de 0.009, demandando 129 épocas (iterações completas) para a conclusão do processo. A TABELA 1 abaixo mostra a Tabela de Contingência deste experimento.

Tabela de Contingência				
Categoria	a	b	c	d
Alumínio	23	3	10	169
Cobre	40	6	4	155
Ouro	57	14	18	116
Ferro	28	8	5	164
Prata	4	5	6	190
Zinco	10	7	0	188
Totais:	162	43	43	

TABELA 1: CONTINGÊNCIA DA SUB-COLEÇÃO METAIS

A coluna “a” mostra a quantidade de documentos que foram corretamente atribuídos a cada categoria. A coluna “b” mostra a quantidade de documentos incorretamente atribuídos a cada categoria. A coluna “c” mostra a quantidade de documentos incorretamente rejeitados para as categorias. A coluna “d” mostra os documentos corretamente rejeitados para elas.

Categoria	$a/(a+c)$ Abrangência	$a/(a+b)$ Precisão	$b/(b+d)$ Falha	$(a+d)/n$ Acurácia	$(b+c)/n$ Erro
Alumínio	0,69	0,88	0,017	0,93	0,063
Cobre	0,90	0,86	0,037	0,95	0,048
Ouro	0,76	0,80	0,107	0,84	0,156
Ferro	0,84	0,77	0,046	0,93	0,063
Prata	0,40	0,44	0,025	0,94	0,053
Zinco	1,00	0,58	0,035	0,96	0,034

TABELA 2: EFICÁCIA DA CATEGORIZAÇÃO DA SUB-COLEÇÃO METAIS

A partir da *Tabela de Contingência*, medidas individuais de eficácia (para cada categoria) podem ser calculadas. Essas medidas são mostradas na TABELA 2.

O gráfico da FIGURA 3 ilustra as relações entre a abrangência e a precisão obtidas em cada uma das 6 categorias. Calculando-se a média geral de abrangência ela atinge 79%, assim como a precisão.

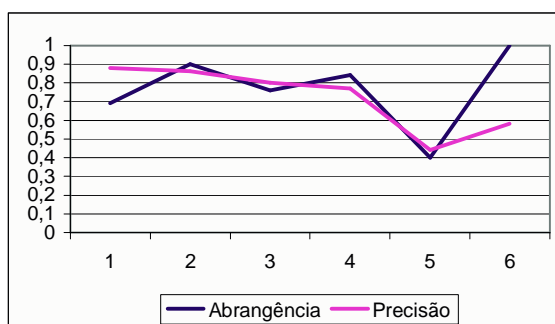


FIGURA 3 - GRÁFICO COMPARATIVO ENTRE "ABRANGÊNCIA E PRECISÃO"

4 Conclusões

Nesse artigo foram expostos os problemas relacionados com a coleta e a análise de informações que as empresas estão começando a enfrentar. Esses problemas são decorrentes do novo ambiente globalizado que, aliado às facilidades oferecidas pelos meios atuais de comunicação, aumenta enormemente a quantidade de informações que circulam em um ambiente empresarial. Essa grande quantidade dificulta a análise e a utilização efetiva das informações.

Nesse cenário, a quantidade de informações que um profissional recebe deve ser a mínima possível. Apesar de mínima, essa informação deve conter todos os aspectos necessários ao seu trabalho, sem que dados relevantes sejam deixados de lado. É somente assim que uma empresa consegue manter-se competitiva no mercado, monitorando e recebendo todas as informações relevantes ao seu ambiente de negócio (e somente essas). Em um ambiente dinâmico, os processos de coleta, disseminação e análise de informações

não são facilmente realizados. Torna-se necessária a utilização de ferramentas computacionais de apoio, que facilitem essas tarefas.

Considerando esse argumento, esse artigo abordou a metodologia e as técnicas envolvidas na categorização de textos. Também foram tecidos comentários teóricos, ilustrados com alguns exemplos de sistemas de categorização de textos existentes. Esses comentários demonstram que a categorização de textos pode ser útil nos processos de disseminação e análise de informações para empresas, já que os profissionais podem definir as categorias de textos que mais lhes interessam e a categorização é capaz de lhes identificar os textos que pertencem a essas categorias.

A fim de demonstrar uma aplicação prática da categorização, foram apresentados os resultados de um experimento em que foi feita a categorização de uma sub-coleção do corpus de documentos da coleção Reuters-21578. Estes, especificamente, foram obtidos através da utilização de uma rede neural Perceptron Multicamadas, treinada com algoritmo *Backpropagation*, obtendo médias de abrangência e precisão de 79%.

Se cada uma das seis sub-categorias fosse atribuída a um funcionário de uma empresa, eles passariam a receber praticamente somente as informações da categoria que lhes foi definida (com um índice de acerto de 79%). Dependendo da categoria, isso pode significar uma grande redução na quantidade de informações que o usuário recebe diariamente. Além disso, mesmo com um índice de erro de 21%, as informações recebidas tornam-se mais específicas (da categoria), facilitando sua leitura e análise e evitando que o usuário fique sobrecarregado. De outra forma, muito possivelmente, o usuário não consiga identificar e analisar sequer os 79% de informações que lhe são relevantes.

Esses dados mostram que a categorização de textos é uma técnica extremamente promissora no âmbito de atividades empresariais. Por outro lado, a categorização não resolve todos os problemas. Por ser utilizada como filtro que minimiza a quantidade de informações recebidas, a categorização não age em problemas relacionados à busca (coleta) e a análise de informações.

Além disso, devido a problemas pertinentes à própria linguagem natural, que é ambígua e, muitas vezes, nem mesmo as pessoas conseguem identificar corretamente o

contexto de um documento, frase ou palavra, existem diversos níveis de complexidade que podem ou não ser tratados pelas rotinas de classificação. Quanto maior o nível de complexidade envolvido no processo de classificação, melhores tendem a ser os resultados. Porém, a dificuldade de implementação e o tempo de processamento também tendem a aumentar com a complexidade.

Com isso, dependendo do que a empresa espera do processo de categorização (e isso pode depender da sua necessidade de informação), o processo pode não precisar ser muito complexo. Nos casos em que há uma necessidade mais complexa, torna-se necessário combinar técnicas de análise de linguagem natural, recuperação de informações e métodos de análise de dados qualitativos.

Apesar desses fatores, a existência de um SCT empresarial, mesmo que simples e pouco complexo, pode representar uma importante ferramenta de apoio, especialmente para a tomada de decisões.

Referências bibliográficas

- [1] DAVENPORT, Thomas H.; PRUSAK, Lawrence. Working knowledge: how organizations manage what they know. ACM Ubiquity, Reprinted by permission of harvard business school press. Excerpt of Working Knowledge: how organizations manage what they know, by Thomas H. Davenport and Lawrence Prusak. 2000. n.24, p8-14, 2000. Disponível por W3 em http://www.acm.org/ubiquity/book/t_davenport_1.html (20/07/2000).
- [2] JACOBS, Paul. Using Statistical Methods to Improve Knowledge-Based News Categorization. **IEEE Expert Intelligent Systems and their Applications**. Los Alamitos, v.8 n.2, p.13-23, april, 1993.
- [3] RESNICK, Paul; VARIAN, Hal R. Recommender Systems. **Communications of the ACM**, New York: ACM Press. v.40, n.3, p.56-58, 1997.

- [5] LEWIS, D. D. **Representation and Learning in Information Retrieval**. Massachusetts: Department of Computer and Information Science. University of Massachusetts, 1992. PhD Thesis.
- [6] GROBELNIK, Marko; MLADENIC, Dunja. **Efficient text categorization**. Disponível por W³ em <http://www.cs.cmu.edu/afs/cs/user/dunja/www/pww.html> (20/01/2000).
- [7] <http://www.yahoo.com>.
- [8] TERVEEN, Loren et al. PHOAKS: A System for Sharing Recommendations. **Communications of the ACM**, New York: ACM Press. v.40, n.3, p.59-62, 1997.
- [9] KONSTAN, Joseph a. et al. GroupLens: Applying Collaborative Filtering to Usenet News. **Communications of the ACM**, New York: ACM Press. v.40, n.3, p.77-87, 1997. Disponível por WWW em <http://www.cs.umn.edu/Research/GroupLens> (em 23/05/2000).
- [10] BALABONIVIC, Marko; SHOHAM, Yoav. FAB: Content-Based, Collaborative Recommendation. **Communications of the ACM**, New York: ACM Press. v.40, n.3, p.66-72, 1997.
- [11] KAUTZ, Henry et al. Referral Web: Combining Social Networks and Collaborative Filtering. **Communications of the ACM**, New York: ACM Press. v.40, n.3, p.63-65, 1997.
- [12] KOWALSKI, G. **Information Retrieval Systems : Theory and Implementation**. Boston: Kluwer Academic Publishers, 1997.
- [14] RIZZI, Claudia Brandelero. **Categorização de Textos por Rede Neural - Estudo de Caso**. Porto Alegre: PGCC da UFRGS, 2000. Dissertação de mestrado (ainda não publicada).
- [13] SALTON, Gerard. **Introduction to Modern Information Retrieval**. New York: MacGraw-Hill, 1983.

- [14] WIENER, E.; PEDERSEN, J.; WEIGEND A. **A Neural Network Approach to Topic Spotting.** Disponível por WWW em http://cora.jpcc.com/Information_Retrieval/Retrieval/index.html (17/12/1999).
- [15] WIVES, Leandro. **Um Estudo sobre Agrupamento de Documentos Textuais em Processamento de Informações não Estruturadas usando Técnicas de Clustering.** Porto Alegre: PPGC da UFRGS, 1999. Dissertação de Mestrado.
- [16] YANG, Yiming. **An Evaluation of Statistical Approaches to Text Categorization.** Disponível por WWW em <http://www.cs.cmu.edu/~yiming/publications.html> (15/01/2000).
- [17] FREEMAN, J.A., SKAPURA, D.M. **Neural Networks, Algorithms, Applications, and Programming Techniques.** Massachusetts: Addison-Wesley Publishing Company, Inc., 1991.
- [18] HAYKIN, Simon. **Neural Networks - A Comprehensive Foundation.** New York: Macmillan College Publishing Company, 1994.
- [19] LEWIS, D. D. **Reuters-21578 Text Categorization Test Collection.** AT&T Labs Research. Disponível por WWW em <http://www.research.att.com/~lewis> (04/03/2000).
- [20] PORTER, M. *An Algorithm for suffix stripping.* **Program.** v.14, n.3, p.130-137, 1980.
- [21] LEWIS, D. D. **Evaluating Text Categorization.** Disponível por WWW em <http://www.research.att.com/~lewis/chronobib.html> (25/11/99).