Aplicação de Inteligência Artificial em categorização de informações

Maria Carolina Carlos Pinto;Knowtec;Robson Garcia Formoso;Knowtec.

Resumo: O objetivo do presente trabalho consiste em utilizar técnicas de inteligência artificial e aplicá-las no processo de categorização de informação, utilizando a técnica de RBC (Raciocínio baseado em casos).

Com essa técnica é feito um calculo de similaridade entre textos livres, mais especificamente entre notícias. Existe uma base de dados com notícias categorizadas por grupos temáticos e é essa base que é utilizada como fonte de informação para o processo de categorização. A partir de um software de captura de notícias o sistema de categorização recebe uma ou mais notícia nova como entrada e cada notícia dessa é submetida ao processo de calculo de similaridade com um conjunto de notícias que está na base de dados.

Existem para esse processo de similaridade duas variáveis importantes que regulam o resultado do processo. As variáveis são: o grau de similaridade, que determina quanto dois textos livres são semelhantes, e o intervalo de tempo de notícias, que vai servir como base de informações para o calculo de similaridade.

O processo consiste em, com a entrada de uma notícia nova a mesma é submetida ao calculo de similaridade com cada notícia contida no intervalo de tempo. Quando o calculo obter um resultado maior ou igual ao grau de similaridade previamente determinado pelo especialista, automaticamente essa nova notícia é associada aos grupos temáticos da noticia que obteve o grau de similaridade aceitável.

Ao final desse processo cada notícia nova estará associada aos respectivos novo grupos temáticos sem a necessidade da intervenção humana de um especialista. Após testes de campo, o trabalho aqui apresentado teve um aproveitamento de 89% de acerto. Resultado esse satisfatório.

Porém, com a utilização das variáveis de controle o sistema pode ser configurado para servir como um sistema especialista de apóio a decisão. Não decidindo, mas sim indicando um conjunto de decisões, onde o especialista na área tem a função de aprovar ou não o resultado do processo.

Palavras Chave: Organização de informação. Inteligência artificial. Raciocínio baseado em casos. Categorização

1 INTRODUÇÃO

O surgimento das tecnologias da informação e comunicação (TIC's) e sua aplicação no ambiente corporativo foram inevitáveis, promovendo a adaptação de todos os indivíduos a essas transformações. Com a modernização das instituições, os processos de comunicação têm gerado informações num fluxo de tempo acelerado, o que dificulta o armazenamento e organização dos conteúdos. Nos últimos meses, a implementação e a criação de novos softwares nas empresas cresceu de maneira acentuada. Com isso, ficou demonstrado que, além de facilitar, agilizar processos, proporcionou o aumento da produtividade destas na maioria das tarefas.

Categorização de informações é uma tarefa que pode ser feita manualmente, mas dependendo do número de fontes agregadas, vai exigir automatização do processo. Com a utilização de um sistema especialista em categorização é possível facilitar o trabalho e antecipar o resultado final da armazenagem. O resultado é o compartilhamento das informações com a mesma qualidade.

Para testar o processo de categorização de informações, foi criado um categorizador de notícias num sistema desenvolvido por uma equipe de tecnologia da informação (TI), o Auto SI. Nele, as principais mídias nacionais impressas e com versões disponíveis eletronicamente são parseadas (capturadas) e suas notícias armazenadas em uma base de dados que é alimentada diariamente. Assim que o sistema realiza a busca, as notícias ficam disponíveis para que uma equipe de revisão realize manualmente o processo de categorização.

O processo de categorizar notícias nada mais é do que a classificação destas conforme o assunto. Hoje, é um revisor quem faz a leitura, seleção e distribuição da notícia num grupo temático adequado ao tema. Mais tarde, ele a direciona aos clientes. Os grupos temáticos já estão previamente cadastrados na base de dados, atualizada pelos próprios revisores sempre que necessário.

Sem um software de automatização os revisores têm como obrigação conhecer todos os grupos temáticos. Por isso, eles passam por um treinamento, haja vista que o número de grupos é extenso. Isso não impede que surjam dúvidas e que se cometam pequenos erros de associações a grupos. Consequentemente, o clipping é enviado com atrasos.

O clipping é o resultado final de todo o processo de categorização. Consiste num e-mail personalizado para o cliente em que todas as notícias de seu interesse encontram-se disponíveis e separadas por assunto. O processo costuma levar cerca de duas horas e meia.

No entanto, com o auxílio dos métodos de inteligência artificial, pôde-se criar um módulo para automatizar o processo de categorização, resultando em maior produtividade, menor tempo e menor chance de erros. Do ponto de vista comercial, trata-se de ferramenta imprescindível, pois é capaz de gerar clipping sem a interferência do revisor.

2 RACIOCÍNIO BASEADO EM CASOS

O Raciocínio Baseado em Casos (RBC) teve a sua origem na ciência cognitiva e na inteligência artificial. Devido a essa interdisciplinaridade, tem sido aplicado em diversos domínios, como na aprendizagem e na resolução de problemas atuais, sempre a partir de conhecimentos extraídos de experiências similares, anteriormente vivenciadas. Segundo LEAKE (1996), problemas parecidos têm soluções parecidas e os tipos de problemas se repetem.

Para MATLIN (1998), o processo de resolução de problemas é muito influenciado

pelo uso de conhecimento prévio. Porém, este não é obrigatoriamente aplicado de forma rotineira e reprodutiva. Existe a capacidade de alterar e moldar seletivamente experiências anteriores, de forma totalmente conceitual, para torná-las aplicáveis em situações novas e inesperadas (*apud* EYSENCK & KEANE, 1990).

WANGENHEIM (2003) leciona que o "Raciocínio Baseado em Casos é uma abordagem para a solução de problemas e para o aprendizado com base em experiência passada". Assim, podemos dizer que o RBC utiliza soluções já conhecidas para solucionar novos problemas.

"O RBC utiliza o mesmo sistema de resolução de problemas dos seres humanos, onde através de raciocínio analógico as pessoas verificam as suas experiências passadas para resolver os seus novos problemas" (COSTA, 1999).

Existem características básicas em sistemas de RBC, que segundo WANGENHEIM (2003) seriam:

- Representação do conhecimento: o conhecimento é representado em forma de casos e armazenado em casos passados.
- Similaridade: é o grau de semelhança entre dois casos, que será usado para a recuperação de casos similares.
- Adaptação: é o processo de alteração de um caso ou de sua solução. Geralmente um caso nunca é igual ao outro. Por isso, muitas vezes torna-se necessário a modificação do caso para solucionar o problema proposto.
- Aprendizado: está uma das principais características dos sistemas de RBC. Após o sistema de RBC resolver um problema com sucesso, o caso pode ser armazenado na base de casos, a fim de que o sistema possa se manter sempre atualizado.

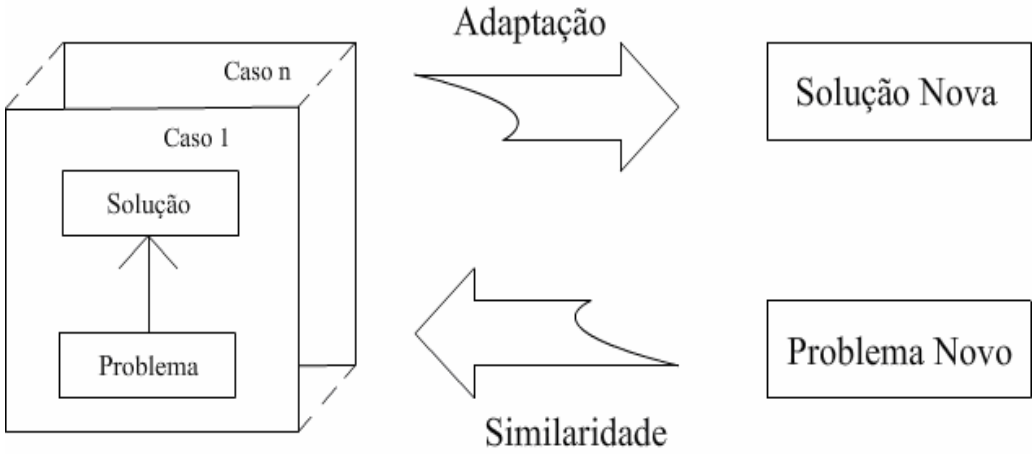

Figura 1. Modelo básico do enfoque RBC
Fonte: WANGENHEIM, 2003

Um sistema de RBC pode funcionar da seguinte maneira: surge um novo problema, o sistema de RBC procura pelos casos mais similares. Em seguida, adapta o(s) caso(s) similar(es) para a solução do problema e armazena essa nova solução (COSTA, 1999). Esse processo é chamado de Ciclo de RBC.

AAMODT e PLAZA (1994) sugerem um modelo tradicional de ciclo RBC (figura 2) que também é conhecido como 4R, formado por quatro tarefas principais, quais sejam:

- recuperar: após o problema ser identificado, o sistema recupera os casos similares;
- reutilizar: com posse dos casos similares, o sistema reutiliza os casos para

resolver os problemas;

- revisar: depois de uma solução definida pelo sistema, ela é revisada;
- reter: após o problema ser solucionado, o sistema retém a solução do caso na base de casos para o aprendizado do sistema.

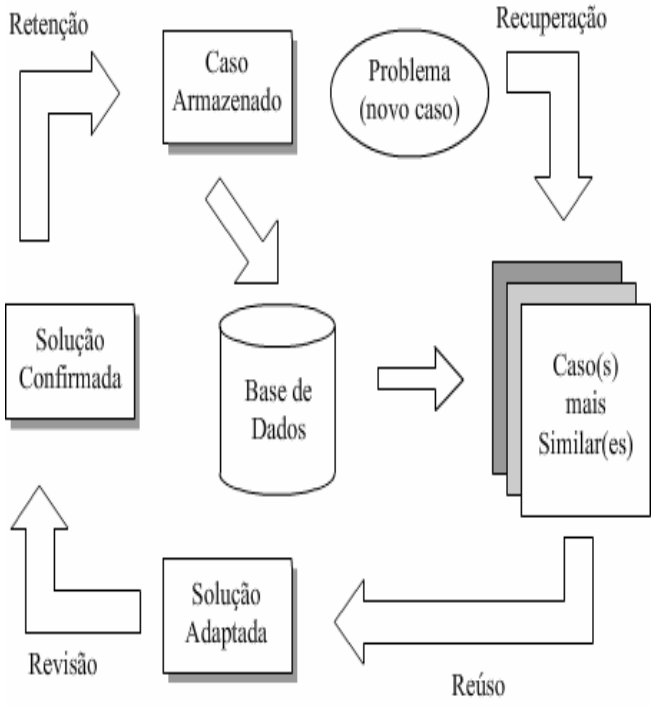

Figura 2. Ciclo de RBC
Fonte: WANGENHEIM (2003)

3 SIMILARIDADE

Para a recuperação dos casos é utilizado um cálculo de similaridade. É este que vai definir o grau de semelhança entre dois casos diferentes.

Problemas similares possuem soluções similares e essa é uma idéia de RBC. “A identificação das características relevantes de um caso vai determinar a sua similaridade com outros casos” (WANGENHEIM, 2003).

Em uma ferramenta de RBC podem ser utilizadas várias medidas de similaridade. A escolha da medida mais adequada depende muito do domínio de aplicação. Pode-se utilizar também medidas em conjunto para obter uma medida de similaridade mais eficaz (WANGENHEIM, 2003).

É na base do domínio da aplicação que é calculada a similaridade do sistema. Dessa forma, o contexto do sistema RBC vai estar atrelado à similaridade. Uma definição da medida de similaridade tem que ser baseada nas metas de recuperação. As metas de recuperação são os objetivos da recuperação. As metas variam conforme a área de atuação do sistema WANGENHEIM (2003).

3.1 Métrica de similaridade

Devido à dificuldade em comparar textos livres, onde nesse caso seria necessária uma complexa comparação de semântica entre as palavras, sugere-se substituir as palavras por valores simbólicos, o que facilitaria o cálculo da similaridade com ajuda de técnicas como a do Modelo de Vetor (WANGENHEIM, 2003).

A métrica de similaridade utilizada seria então a do Modelo de Espaço Vetorial. Nesse modelo, a similaridade é a expressão de igualdade entre os termos. Pode-se adotar esse termo como forma de representação do número de propriedades que determinados objetos possuem em comum, assim como o número de propriedades incomuns entre objetos (WIVES, 1997). As propriedades podem ser representadas por características e os objetos por documentos ou termos.

Por meio dessa técnica conseguimos gerar um grau de similaridade entre as notícias armazenadas sempre que uma nova notícia entrar na base de dados e for agrupada aos grupos temáticos das notícias similares. O grau de similaridade, nesse caso, é variável e um usuário responsável pode alterar conforme a necessidade. É possível fazer uma alteração dentro de um intervalo de data das notícias para efeitos de comparação.

Com a entrada de uma nova notícia para comparação, o processo é iniciado com a remoção de palavras sem significado. Palavras que se encontram em uma lista chamada StopWord, da qual o sistema utiliza com referência para o primeiro processo. Logo em seguida é realizado um processo de limpeza no texto com a remoção de caracteres especiais que também não têm significância.

As notícias são armazenadas em dois locais diferentes. Num local ficam as notícias originais e no outro as notícias limpas pelo processo inicial. As notícias livres de palavras sem significância para o processo de similaridade é que são utilizadas no cálculo de similaridade, tanto as novas quanto às antigas.

O processo de cálculo da similaridade do Modelo de Espaço Vetorial é iniciado com a aplicação da seguinte fórmula:

$$sim(x,y) = \frac{\sum_{i=1}^{t}(w_{i,x} \times w_{i,y})}{\sqrt{\sum_{i=1}^{t}(w_{i,x})^2} \times \sqrt{\sum_{i=1}^{t}(w_{i,y})^2}}$$

Para aplicar fórmula precisamos seguir os passos:

a) Definir a freqüência das palavras na notícia nova e nas notícias que vão ser comparadas. Nesse caso, são todas as notícias em um intervalo de datas previamente definido.

$$tf_{t,d} = freq_{t,d}$$

b) Calcular a “inverser document frequency” que é utilizada na fórmula para determinar o peso de cada palavra. O cálculo deve ser feito dividindo o número total de notícias encontradas no intervalo de data definido pelo número de notícias que foi encontrado cada palavra no processo de definição das freqüências.

$$idf_t = \frac{N}{n_t}$$

c) Obter o peso de cada palavra na notícia de origem (notícia nova) e obter o peso de cada palavra da notícia que já está na base de dados. Um vetor de palavras e seus pesos são criados. Aqui, as comparações são somente entre duas notícias: a notícia de entrada e a notícia que já se encontrava na base de dados com seus respectivos grupos temáticos definidos.

$$w_{t,d} = tf_{t,d} \times idf_t$$

d) É aplicada a fórmula da similaridade entre a nova notícia e cada uma das que foram encontradas no intervalo das datas definidas.

Um relatório informando o grau de similaridade entre a notícia nova e todas as outras é gerado após aplicação do cálculo da similaridade entre notícias. O próximo passo é agrupar todos os grupos temáticos encontrados nas notícias com o valor da similaridade maior ou igual ao valor definido previamente para o grau de similaridade. O valor somado ao intervalo de datas é o que define o grau de acerto de todo o processo.

4 RESULTADOS

O processo foi aplicado utilizando 30% para o grau de similaridade e utilizado um intervalo de tempo de 30 dias para 15 notícias novas que entraram na base de dados. No período de 30 dias foram obtidas 8.003 notícias e foram feitas aproximadamente 120.000 comparações entre elas.

Do processo obteve-se um resultado de 89% de acerto dos grupos temáticos. O resultado é satisfatório para um sistema que tem como objetivo sugerir grupos temáticos. Porém, um problema foi encontrado durante a execução. Um processo com 976 mil comparações levou aproximadamente 34 minutos para ser concluído, o que faz com que a próxima etapa de desenvolvimento seja a otimização de processos para aumento do desempenho.

5 CONCLUSÃO

O presente artigo procurou demonstrar um modelo de processo utilizando inteligência artificial para categorizar automaticamente notícias por grupos temáticos, visando aplicá-lo na organização de informações por assunto, e a conseqüente disseminação seletiva da informação.

O processo da categorização de notícias hoje é feito manualmente por pessoas

habilitadas. Todavia, essas habilidades não se encontram por completo em apenas uma pessoa e sim em determinado grupo, em que cada um tem um papel importante dentro do processo.

O fato das bases estarem divididas por cliente é uma vantagem, pois cada cliente apresenta determinado perfil e, consequentemente, é possível direcionar com maior precisão o processo de categorização. Isso quer dizer que cada notícia na base de dados é categorizada de acordo com o cliente.

De posse dessas bases de dados e com todas as informações e os resultados obtidos podemos afirmar que parte do conhecimento dos especialistas em categorização está contida nessa base de dados.

Ao término do processo verificamos que o resultado obtido foi importante. Para um cliente com poucos grupos temáticos e um número elevado de notícias, o protótipo aqui descrito mostrou plena capacidade de categorização das notícias sem a intervenção humana. A geração de clippings automáticos resulta num produto de rápido desenvolvimento e alto grau de assertividade. Diferentemente dos clippings temáticos comuns que utilizam apenas palavras-chave para categorizar notícias.

Além disso, em clientes com um grande número de grupos temáticos o protótipo apresentou ser um grande facilitador, identificando os grupos temáticos com um acerto de 89%, como foi observado nos resultados. Como ainda há intervenção humana para correção do resultado obtido, existe a possibilidade, futuramente, de se criar processos utilizando a inteligência artificial, a fim de que essa intervenção sirva de aprendizado para o protótipo.

Podemos ver que a partir de uma base de dados com textos livres e previamente associados a grupos temáticos temos um processo de categorização automática que simula o conhecimento e a experiência de um especialista em categorização de notícias.

Assim, mostra-se possível criar um software que faça a categorização de informações automaticamente e proporcione aos especialistas uma seleção e uma disseminação dessas mais rapidamente.

6 REFERÊNCIAS

AAMODT, A.; PLAZA, E. Case-based reasoing: foundational issues, methodological variations, and system approaches. **AI Communications**, v.7, n.1, p.39-59, mar. 1994.

COSTA, Marcello Thiry Comicholi da; BARCIA, Ricardo Miranda. **Uma arquitetura baseada em agentes para suporte ao ensino à distância.** 1999. 90f. Tese (Doutorado em Engenharia de produção) - Universidade Federal de Santa Catarina, Florianópolis, 1999. Disponível em: <http://www.eps.ufsc.br/teses99/thiry/>

EYSENCK, M. & Keane, M. (1990). **Cognitive Psychology,** Hove (UK), p.179-219.

LEAKE, David B. **Case-Based Reasoning – Experiences, Lessons & Future Directions.** 1 ed. Massachusetts, EUA : The MIT Press, 1996. 420p. ISBN 0-262-62110-X.

MATLIN, M. W. **Cognition**. Orlando, Florida (EUA): Harcourt Brace & Company, 1998.

WANGENHEIM, Christiane Gresse von; WANGENHEIM, Aldo von. **Raciocínio Baseado em Casos.** 1ed. Barueri, SP : Manoele, 2003. 294p. ISBN 85-204-1459-1.

WIVES, Leandro Krug. **Um estudo sobre técnicas de recuperação de informações com ênfase em informações textuais.** Programa de Pós-Graduação em Computação (Universidade Federal do Rio Grande do Sul), Rio Grande do Sul, 1997.