Empresas lutam contra o caos de informações

Os dados não-estruturados, a maior parte sob a forma de arquivos de texto, representam 85% do armazenamento de conhecimento de uma organização, mas nem sempre são fáceis de encontrar, acessar, analisar ou usar. Essa situação, porém, tende a mudar com a nova geração de ferramentas de text mining, que permitem às empresas extraírem elementos essenciais de grandes conjuntos de dados não-estruturados, descubram relações e resumam a informação. Muitas organizações estão implementando ou cogitando a implantar esse tipo de software para lidar com suas montanhas de texto, apesar da exigência de conhecimento especializado para fazer as implementações funcionarem.

Desde 2002 a equipe de pesquisa da Dow Chemical, por exemplo, usa o software ClearResearch, da ClearForest, para extrair dados, armazenados há um século, de resumos de patentes químicas, papers de pesquisas publicados e arquivos da própria empresa. Com o gerenciamento melhor da informação e a eliminação do que é irrelevante, a empresa conseguiu reduzir o tempo que os pesquisadores levavam para encontrar o que precisam ler.

As ferramentas de text mining seguem diversas abordagens. O ClearResearch usa metodologia proprietária de correspondência de padrão para buscar a informação, classificá-la e mostrar geograficamente sua relação com outros dados. O software "vê", descobre e extrai conceitos, não apenas palavras, fornecendo uma representação pictórica fácil de entender do texto contido nos documentos.

Mesmo assim, os software de text mining disponíveis hoje ainda não alcançam a precisão das ferramentas de data mining, mas os fornecedores estão aprimorando a capacidade de seus produtos de entender o contexto, o que é fundamental para torná-los eficazes.

Devido a essas limitações, as ferramentas de text mining continuam sendo um nicho restrito a parcelas específicas de uma organização. Mas elas já começam a se disseminar. Na opinião do analista de pesquisa da Enterprise Storage Group, Brian Babineau, quando essa área estiver sob controle, as pessoas vão adaptar as aplicações a outros fins, como data warehousing e CRM.

Embora muitos sistemas de software analisem dados estruturados e não-estruturados, um grande número de empresas usa software de business intelligence (BI) tradicional para dados estruturados e emprega outras ferramentas para dados baseados em texto. Na EDS, por exemplo, todos os 130 mil funcionários preenchem um questionário online sobre seu trabalho uma vez por ano. Outras três vezes por ano, 20 mil funcionários respondem uma pesquisa adicional. Algumas perguntas da pesquisa são de múltipla escolha, o que facilita a inserção das respostas em software de BI da SAS Institute e SPSS, nos quais elas são agregadas, dissecadas e analisadas. Parte do feedback mais importante, porém, vem das respostas às perguntas dissertativas. Antigamente, estas respostas eram encaminhadas a gerentes de linha para que eles extraíssem conclusões, já que não se enquadravam em nenhuma estrutura fácil de gerenciar.

Três anos atrás, a EDS começou a procurar uma maneira melhor de interpretar estas respostas e usufruir as informações que elas continham. A empresa começou a usar o PolyAnalyst, da Megaputer Intelligence, que extrai inteligência de dados estruturados e não-estruturados. O sistema se baseia em uma implementação do dicionário semântico WordNet desenvolvido pelo Cognitive Science Laboratory na Universidade de Princeton. Entre outras funções, ele atribui palavras a categorias e fornece palavras relacionadas. O PolyAnalyst é usado para analisar as perguntas dissertativas, mas ainda utiliza software BI tradicional para as perguntas de múltipla escolha e combina a informação dos dois em relatórios consolidados.

Existem ferramentas distintas especializadas em analisar bancos de dados ou arquivos de texto, mas também há técnicas que permitem que os dois sejam correlacionados. Patricia Cerrito, bio-

estatística e professora de matemática da Universidade de Louisville, vasculha registros hospitalares para descobrir meios de melhorar o índice de alta de pacientes. Ela usa o SAS Text Miner em arquivos de texto, como as fichas dos pacientes. Mas também faz capturas de arquivos simples de bancos de dados de faturamento e farmacêutico e os analisa como texto, em vez de entradas em bancos de dados.

Ao examinar milhares de altas de pacientes com o Text Miner, Cerrito descobriu informações úteis – que a prescrição de determinados medicamentos pode prolongar a permanência de pacientes no hospital, por exemplo, e que os níveis de açúcar no sangue de pacientes diabéticos podem estar relacionados a risco de infecção após cirurgia cardíaca.

No geral, a instalação de um text miner é um processo simples. Cerrito conta que só precisou carregar seis CDs para fazer o SAS Text Miner rodar em sua workstation. O difícil é obter resultados significativos de um processo que depende da habilidade e do conhecimento da pessoa que está usando o software. É preciso ser um analista talentoso para interrogar repositórios de texto apropriadamente.

Além de ter conhecimento de análise, o usuário precisa estar suficientemente familiarizado com o conjunto de dados para entender o que os resultados significam. De todo modo, visando a disponibilizar mais funcionalidade enquanto resolvem problemas de usabilidade, os fornecedores estão incorporando ferramentas de text mining como uma função de background para aprimorar a eficácia de aplicações mais familiares de gerenciamento de documento ou busca.

Glossário

Os text miners empregam diversas abordagens para extrair e apresentar informações relevantes. Abaixo, listamos as definições de métodos comuns

. Classificação – Apresentar os resultados da busca em categorias, em vez de uma massa indiferenciada.

. Clustering – Agrupar documentos similares com base em seu conteúdo.

. Extração – Extrair informação relevante de um documento – por exemplo, extrair todos os nomes da empresa de um conjunto de dados.

. Busca por palavra-chave – Pesquisar documentos conforme a ocorrência de uma palavra ou conjunto de palavras em especial.

. Processamento de linguagem natural – Determinar o significado de palavras escritas levando em conta seu contexto, gramática, coloquialismo e assim por diante.

. Taxonomia – Classificar dados de acordo com uma estrutura predefinida, seja padrão da indústria ou personalizada. Algumas ferramentas podem gerar automaticamente uma taxonomia baseada em análise do data store.

. Visualização – Apresentar graficamente os dados extraídos para que fique mais fácil descobrir e entender as relações.

Ferramentas de Texto

As ferramentas de text mining são produtos autônomos ou embutidos em um sistema de software maior. Fornecedores de banco de dados como IBM e Oracle estão incorporando algoritmos de correspondência de modelo aos produtos de banco de dados, e de gerenciamento de conteúdo como a SAP estão usando taxonomias e classificação de texto para gerenciamento do ciclo de vida da informação.

Fornecedores de data mining como SAS Institute, SPSS e Science Applications International começam a acrescentar text mining aos seus portfólios, além de terem surgido empresas especializadas em text mining como Inxight Software e Stratify.

Computerworld online - acesso em 8/9/2004