

Extração de dados de arquivos semi-estruturados com auxílio do SAXES

É de praxe que os gerentes, assim que chegam ao trabalho, tenham a sua espera sobre a mesa, um material que irão ler com ávida atenção, e que poderá influenciar nas decisões que tomarão no decorrer do dia. Não está se falando do jornal, que será lido logo em seguida, mas de outras informações de extrema importância para a sobrevivência e boa administração de uma empresa. Tratam-se dos populares relatórios, que possuem o poder de sintetizar dados e analisar movimentações transcorridas durante um longo período de tempo, em poucas linhas de texto.

Os relatórios nos dão a indicação do rumo que a empresa está tomando, e permite que o trajeto seja corrigido, ou otimizado, quando houver necessidade. O mais interessante sobre eles, no entanto, é que os gerentes não são os únicos a consumi-los. Muitas vezes, os relatórios são gerados não para serem processados por humanos, mas sim, por máquinas. É dessa forma que sistemas diferentes conseguem se comunicar. Informações extraídas de um sistema são transformadas em um relatório, que por sua vez é enviado para o destino, onde é finalmente lido e interpretado, tudo contando com o auxílio de componentes de *software*. Notas fiscais eletrônicas, por exemplo, seguem esse princípio. Pode-se ver esses documentos como relatórios que precisam ser periodicamente enviados para a receita federal, onde um agente de *software* automatizado processa as informações neles contidos, e executa as ações apropriadas.

A formatação desse tipo de relatório costuma ser projetada de forma a comprimir os dados no menor espaço possível, seguindo regras fixas na disposição do conteúdo. Os componentes de *software*, por sua vez, são aptos a reconhecer essa estranha configuração. Para isso, é necessário desenvolver algoritmos que sejam capazes de varrer o relatório a procura de um estilo pré-definido, ou um *template*. Como os relatórios

costumam seguir um *template*, fica mais fácil para o algoritmo reconhecer essas estruturas de texto repetidas e identificá-las dentro do papel que elas desempenham. O grande desafio, portanto, está na especificação de tais algoritmos. Depois de entender a formatação interna do relatório, o desenvolvedor deve criar um código que consiga reconhecer esta formatação. Para isso, ferramentas como o SAXES podem auxiliar neste processo, agilizando a especificação da solução.

Com SAXES, é possível criar um esquema de dados de forma visual, como esquemas XML e de arquivo texto, bastante comuns em relatórios, e suportados pela ferramenta. Assim sendo, o desenvolvedor pode criar esquemas XML, definindo a relação hierárquica entre diferentes tipos de elementos, ou esquemas texto, em que determina os formatos de linhas que podem ser encontrados em um arquivo texto. Depois de criado o esquema, usa-se os recursos de mapeamento visual do SAXES para especificar como a informação extraída de um relatório deve ser processada. A flexibilidade do SAXES neste tipo de projeto é tanta, que mesmo relatórios para análise humana podem ser reconhecidos pelo *software*, até mesmo aqueles arquivos texto complexos, onde as informações são organizadas de forma agradável aos olhos, mas penosas ao computador. Para o desenvolvimento desse tipo de projeto, o SAXES conta com recursos avançados, como a configuração de delimitadores de colunas e expressões regulares.

Desta forma, relatórios complexos e semi-estruturados podem alimentar processos de BI, fornecendo informações mais ricas aos seus gerentes. Ao invés de dedicar tempo para a leitura e análise destes relatórios analíticos, os gerentes podem delegar estas compilações para um *software* de BI, e assim, contam com mais tempo para as demais atividades gerenciais, analisando relatórios sintéticos e relatórios resultantes de processos de BI.

Seus documentos contêm informações muito importantes. Use o SAXES para extraí-las.

Através de seu editor de esquemas para arquivos texto, o SAXES permite ao usuário criar layouts para a leitura de arquivos semi-estruturados. Dessa forma, é possível extrair informações de quase todos os tipos de documentos de texto, e utilizá-las em qualquer processamento eletrônico de dados.

Saiba mais sobre o SAXES:

55 (51) 3715.5523


IDEALOGIC
S O F T W A R E