EA876 – Introdução Software de Sistema

Análise de dados em notas fiscais

Rodrigo Caus (186807) e Victor Ferrão Santolim (187888).

Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação, Campinas-SP.

Introdução e Objetivo

Notas fiscais emitidas no país não seguem um único formato padrão. Ao observar um pequeno conjunto delas, é rápida a percepção que diferentes municípios geram notas fiscais que divergem em estrutura, mas que contém as mesmas informações. O objetivo do trabalho é implementar um programa de computador capaz de receber como entrada uma Nota Fiscal de serviços em formato XML e automatizar a coleta de informações de interesse administrativo, mais especificamente município do tomador e do prestador, valor dos serviços e valor do ISS retido. De posse desses dados, é desejada a confecção uma tabela em formato CSV para melhor visualização das informações coletadas.

Metodologia

Para atingir o objetivo, partimos da constatação que os arquivos XML das notas fiscais seguem um padrão estrutural de formação, onde cada informação útil, seja ela numérica ou textual, está encapsulada entre tags na forma <info>informação</info>[1]. Um conjunto dessas estruturas encapsuladoras podem ser também ser encapsuladas por outro par de tags <a> e e assim em diante.

A solução assim se divide em duas partes: A primeira é identificar e coletar as strings correspondentes às tags e informações no arquivo fonte; a segunda é associar as informações aos campos de interesse, através do tipo de tag <info> encapsuladora, ignorando os campos restantes.

Na primeira parte, um analisador léxico é responsável por ler o arquivo XML da nota fiscal e gerar tokens conforme a Tabela 1. Para cada um dos tipos de tokens de 1A a 7A mostrados na tabela, existe também a versão dual B que é o token correspondente à tag de fim da região encapsulada iniciada por A.

Tabela 1 - Geração de tokens conforme tags e informações do XML.

Tipo de Token	Exemplos no XML	Significado
1A	<ValorISS>; <tsVlrISSRet>	Antecede o valor de ISS
2A	<ValorServicos>; <tsVlrSvc>	Antecede o valor de Serviços
3A	<TOMADOR_CIDADE>; <tsMunTmd>	Antecede a cidade do tomador
4A	<PRESTADOR_CIDADE>; <tsMunPtd>	Antecede a cidade do prestador
5A	<Tomador>; <ns3:Tomador>	Antecede a seção de dados do tomador.
6A	<PrestadorServico>; <prestador>	Antecede a seção de dados do prestador.
7A	<Cep>	Antecede o CEP (sendo ele do tomador ou prestador)
8	“1.234,56” ; “CANAA DOS CARAJAS”; ”12345-678”	String alfanumérica com espaços, vírgulas, pontos, hífens

A segunda parte consiste em um analisador sintático, que recebe a stream de tokens gerados e as strings capturadas, sendo responsável por identificar sequências correspondentes de tokens no formato das informações buscadas. A Tabela 2 mostra os padrões de tokens detectados para retornar a informação requerida. Strings avulsas, encontradas fora dos padrões estabelecidos, são descartadas. O CEP é utilizado em casos em que o XML não deixa explícito o nome da cidade do tomador ou do prestador na forma de tags específicas.

Tabela 2 - Padrões de detecção de informação requerida

Padrão Identificado	Informação retornada
1A 8 1B	Valor total do ISS associado ao token 8
2A 8 2B	Valor total do serviço associado ao token 8
3A 8 3B	Cidade do tomador, associado ao token 8
4A 8 4B	Cidade do prestador, associado ao token 8
5A 7A 8 7B 5B	CEP do tomador, associado ao token 8
6A 7A 8 7B 6B	CEP do prestador, associado ao token 8

Resultado

O programa desenvolvido é capaz de filtrar os arquivos XML recebidos e preencher corretamente a tabela CSV com as informações de interesse em 70 de 72 testes realizados. Foi considerado aceitável o preenchimento da coluna de cidade do tomador e do prestador com os seus respectivos CEPs de instalação. Entre as entradas que geraram saídas indesejadas, foi observada uma quebra do padrão em comparação com os demais XML de NFs da mesma cidade. Em um dos casos, o arquivo de entrada não continha a informação de ISS retido e em outro não havia CEP nem município do prestador.

O êxito do resultado depende, porém, que as diferentes tags, de modelos de NF distintos, que delimitam as informações alvo tenham sido previamente inseridas no padrão buscado pelo analisador léxico na forma de expressões regulares, caso sigam um padrão diferente dos já presentes para outras cidades. Nesse processo, as regras estabelecidas pelo analisador sintático podem permanecer inalteradas, pois as regras de formação e os tokens serão os mesmos.

Entre os possíveis usos do programa implementado, está a automatização de coleta de dados para contabilidade de empresas. Com um script apropriado, é possível gerar uma tabela onde as colunas são as cidades e valores e as linhas são correspondentes às informações coletadas de cada NF a ser contabilizada, reduzindo assim o esforço braçal em copiar os valores manualmente.

Referências

[1] Microsoft Developer Network. **XML Standards**. Disponível em <https://msdn.microsoft.com/pt-br/library/bb399529(v=vs.120).aspx>.