

Extração de informação de texto formal para a modelação expedita de edifícios monumentais

Ricardo Baptista
DEI - FEUP/ LIACC
Rua Dr. Roberto Frias, s/n
4200-465 Porto
ricardjose@fe.up.pt

Luís Paulo Reis
Universidade do Minho/LIACC
Guimarães/Porto
lpreis@dsi.uminho.pt

Luís Magalhães
Universidade de Trás-os-Montes e
Alto Douro/INESC TEC
Vila Real/Porto
lmagalha@utad.pt

Roberto Rodrigues António Coelho Augusto Sousa
INESC TEC / DEI - Faculdade de Engenharia, Universidade do Porto
Rua Dr. Roberto Frias, s/n, 4200-465 Porto
{roberto.rodrigues, acoelho}@fe.up.pt

Resumo

A extração de informação a partir de descrições textuais para a modelação procedimental de ambientes urbanos é apresentada com solução para os edifícios antigos. No entanto, este tipo de edifício carece de maior cuidado com os detalhes de alto nível. Este artigo descreve uma plataforma para a geração expedita de modelos 3D de edifícios monumentais, cuja arquitetura é modular. O primeiro módulo permite a extração de informação a partir de textos formais, pela integração do NooJ num Web Service. No segundo módulo, toda a informação extraída é mapeada para uma ontologia que define os objetos a contemplar na modelação procedimental, processo esse realizado pelo módulo final que gera os modelos 3D em CityGML, também como um Web Service.

A partir desta plataforma, desenvolveu-se um protótipo Web para o caso de estudo da modelação das igrejas da cidade do Porto. Os resultados obtidos deram indicações positivas sobre o modelo de dados definidos e a flexibilidade de representação de estruturas diversificadas, como portas, janelas e outras características de igrejas.

Palavras-chaves

Procedural modelling, natural language processing, information extraction.

1. INTRODUÇÃO

A modelação expedita apresenta-se como uma solução para a criação de ambientes urbanos com redução de custos e de tempo de execução. Contudo, para manter uma fidelidade visual aceitável é preciso aceder a fontes de informação, que no caso dos edifícios antigos, existe em grande quantidade em formato textual. No trabalho apresentado, damos o enfoque à extração do texto formal para um formato intermédio (XML) e, posterior conversão para o formato CityGML [Kolbe05]. Esta tecnologia é aplicada na modelação das igrejas do Porto.

2. TRABALHO RELACIONADO

Este trabalho segue uma linha de investigação relacionada com a extração de informação para edifícios monumentais [Rodrigues10]. A automatização do processo de modelação de edifícios baseia-se nas ferramentas automáticas de tradução e geração de resumos [Baeza04]. A utilização do NooJ neste trabalho revelou-se importante pelas seguintes características: multilingue, análise gramatical, sintática e morfológica de múltiplos textos em vários formatos, inserção de anotações ao texto e aplicação de gramáticas personalizadas [Silberztein05].

Para a visualização dos modelos 3D gerados, utilizou-se o CityGML, [Kolbe05].

3. ARQUITETURA DO SISTEMA

O sistema proposto designa-se por MEXEM (Modelação EXpedita de Edifícios – Monumentais) e é composto por três blocos: modelo de dados, extração e modelação3D.

3.1 Modelo de Dados

Entende-se por edifícios monumentais as catedrais, palácios e castelos cujas estruturas são complexas, precisando dum elevado nível de detalhe quando modelados.

A Figura 1 apresenta uma representação simplificada do modelo de dados desenvolvido, que faz corresponder os nós do ficheiro XML com os atributos das classes criadas para a criação do modelo em cityGML.

3.2 Extração de Informação com NooJ

A extração de informação tem como ponto de partida a extração de palavras-chave ou padrões de informação do ficheiro de texto. Para garantir a interoperacionalidade entre os vários módulos assume-se que o formato XML é o garante necessário, obedecendo às especificações do *schema* que suporta o modelo de dados.

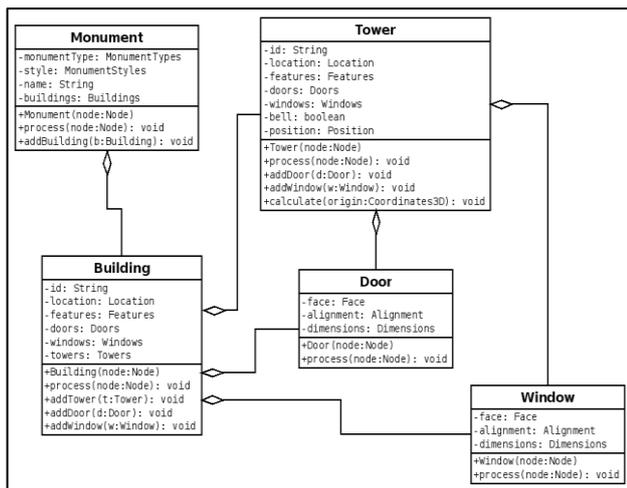


Figura 1: Modelo de dados para edifício monumental

Nas cerca de 109 gramáticas criadas, são definidos padrões conforme a sequência de texto pretendida e adicionadas anotações para enquadrar a informação extraída. Na Figura 2 apresenta-se um exemplo de gramática, com os padrões e anotações que se pretende no *output*.

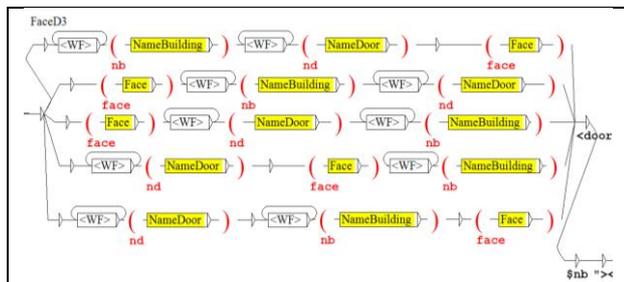


Figura 2: Exemplo de mapeamento da gramática

A partir da sintaxe própria da ferramenta conseguem-se inserir as anotações em formato de *tags* para delimitar a informação de forma coerente com o modelo de dados.

3.3 Modelação 3D em CityGML

A escolha do CityGML para converter para o modelo 3D deveu-se à possibilidade de combinação da informação semântica com a geométrica, os diversos tipos de detalhe e possuir uma biblioteca de criação de ficheiros gml.

O ficheiro XML é convertido para um modelo interno, baseado na ontologia, através de uma biblioteca gratuita fornecida pelo CityGML, *cityGML4j*.

4. DESENVOLVIMENTO

O desenvolvimento do protótipo Web teve como suporte a linguagem C# através do *Visual Studio 2010*. Com a utilização do *Web Service*, conseguiu-se que a extração de informação (Nooj) e a modelação 3D (*cityGML4j*) sejam independentes da plataforma, para potenciar outras utilizações e desenvolvimentos futuros.

5. RESULTADOS

Para os testes foi utilizado um conjunto de igrejas da zona do Porto, cujos textos estavam disponíveis em http://pt.wikipedia.org/wiki/Categoria:Igrejas_do_Porto.

Apresentamos na figura 3, o resultado da extração da informação relativa à igreja do Bonfim, onde a comparação com a imagem real permite verificar uma grande aproximação visual ao monumento arquitetónico.



Figura 3: Modelo 3D e imagem real da igreja do Bonfim

O resultado obtido com as 30 igrejas ficou um pouco aquém do esperado, pois só após a sua reescrita foi possível representar 11 delas com grande proximidade e realismo. Este facto deveu-se ao não alinhamento dos textos com os padrões das gramáticas, bem como a ausência de detalhe suficiente para o modelo de dados.

6. CONCLUSÕES

O objetivo de extrair informação de uma descrição textual para obter a representação 3D foi conseguido para diversas igrejas, bem como confirmada a adequação do modelo de dados à modelação pretendida a partir das gramáticas do NooJ. Desta forma, tornou-se possível a criação de modelos 3D com um nível aceitável de fidelidade visual. Também se constatou a dificuldade inerente à extração de informação quando as descrições não foram escritas com orientação às gramáticas personalizadas.

Como trabalho futuro, propomos a extensão da extração de informação para outros tipos de edifícios arquitetónicos e a apresentação de um exemplo do formato textual mais ajustado às gramáticas utilizadas. Pretende-se também promover a edição interativa de texto, bem como colmatar algumas lacunas de informação relativa às dimensões dos edifícios, integrando o sistema com outras fontes de dados geográficos (ex. GIS).

7. AGRADECIMENTOS

Este trabalho é financiado por Fundos FEDER através do Programa Operacional Fatores de Competitividade – COMPETE e por Fundos Nacionais através da FCT – Fundação para a Ciência e a Tecnologia no âmbito do projeto ERAS - Expeditious Reconstruction of Virtual Cultural Heritage Sites (PTDC/EIA-EIA/114868/2009).

8. REFERÊNCIAS

[Baeza04] Ricardo, Baeza-Yates. Challenges in the Interaction of Information Retrieval and Natural Language Processing. *5th Int. Conf. on Computational Linguistics and Intelligent Text Processing*, 2004.

[Kolbe05] T. H. Kolbe, Gröger G., et al. CityGML - Interoperable Access to 3D City Models. *Proc. Int. Symp. on Geoinformation for Disaster Management*, 2005.

[Rodrigues10] Roberto Rodrigues, António Coelho e Luís Paulo Reis. Procedural modelling of monumental buildings from textual descriptions. In *2010 IEEE Cong. on Evolutionary Computation (CEC)*, Barcelona, 18-23 July, pp. 1-8, 2010.

[Silberstein05] M. Silberstein. NooJ: A Linguistic Annotation System For Corpus Processing. *HLT/EMNLP on Interactive Demonstrations*, Stroudsburg, USA, 2005.