EXTRACÇÃO DE RELAÇÕES SEMÂNTICAS EM INFORMAÇÃO NÃO ESTRUTURADA PARA ENRIQUECIMENTO DE ONTOLOGIAS DE DOMÍNIO

SEMANTIC RELATIONS EXTRACTION IN UNSTRUCTURED INFORMATION FOR DOMAIN ONTOLOGIES ENRICHMENT

Abstract

Since semantic web appearance, several domain ontologies were developed and delivered in open access repositories. The existing domain ontologies describe semantic elements specific to a particular domain. These elements can be used complementing existing document information. This articulation can be duly empowered if, new methods could be created that in a semi automatic way, could help the ontologic precision. Specifically, the new patterns originating the building of new knowledge will be extracted not only from domain ontologies, but also from unstructured information sources.

One of the greatest challenges related to domain ontology enrichment, known in scientific community as “ontology learning”, is the fact in which “pure” automated processes that could make the above mentioned enrichment from unstructured information sources does not exist. In scientific community there are several contributions in this area, namely in the development of methods to quantify the way that existent ontology concepts inside domain ontology are related. These approaches only take advantage of the information included in ontologies and do not consider exterior information to quantify these relations.

The main goal of this dissertation is to use data mining techniques as a way of extracting patterns (here presented as semantic associations) from unstructured information sources. The idea to develop in this work is based on the statistic analysis of co-occurrence between, the more relevant terms from a document corpus and to quantify this analysis through semantic relations between concepts from domain ontology. The domain of the information sources presented in this project is focused on Civil Construction.

Keywords: Artificial Intelligence, Semantic Web, Data Mining, Knowledge Discovery, Machine Learning, Ontology Enrichment, Association Rules, Frequent Pattern

Sumário

Desde o aparecimento da *web* semântica, várias ontologias de domínio foram desenvolvidas e disponibilizadas em repositórios de acesso aberto. As ontologias de domínio existentes descrevem elementos semânticos, que são específicos a um determinado domínio, elementos esses que podem ser usados com forma de complemento à informação existente em documentos. Esta complementaridade poderá ser devidamente potenciada, se forem criados novos métodos que de forma semi-automática auxiliarem o refinamento ontológico. Mais especificamente, os novos padrões que dão origem à geração novo conhecimento poderão ser extraídos não só de ontologias de domínio, mas também de fontes de informação não estruturada.

Um dos grandes desafios relacionados com o enriquecimento de ontologias de domínio designado na comunidade científica por *ontology learning*, prende-se com o facto não existirem automatismos “puros” que permitam esse mesmo enriquecimento a partir de fontes de informação não estruturadas. Existem bastantes contribuições científicas nesta área, nomeadamente no desenvolvimento de métodos que permitam quantificar a forma como os conceitos existentes numa ontologia de domínio estão relacionados. Estas abordagens utilizam apenas a informação contida nas ontologias e não fazem uso de informação externa à ontologia para quantificar essas mesmas relações.

Esta dissertação tem como principal objectivo, o uso de técnicas de *data mining* como forma de extracção de padrões (aqui definidos para associações semânticas) em fontes de informação não estruturada. A ideia a ser desenvolvida no âmbito desta dissertação, tem por base a análise estatística da co-ocorrência entre os termos mais relevantes de um corpus de documentos e, quantificar essa análise sob a forma de relações semânticas entre conceitos de uma ontologia de domínio. O domínio das fontes de informação aqui a serem tratadas, é focado no sector da construção civil.

Palavras-Chave: Inteligência Artificial, Web Semântica, Data Mining, Descoberta de Conhecimento, *Machine Learning*, Enriquecimento de Ontologias, Regras de Associação, Padrões Frequentes

*Dedico a concretização desta etapa, finalizada por esta dissertação aos meus Pais, Mário Luiz e Maria Edite…*

*“Always look on the bright side of life!”*

*Monty Python, in “Life of Brian”*

Contents

[1 Introduction 2](#_Toc397462220)

[1.1 Motivation 2](#_Toc397462221)

[1.2 Development context 4](#_Toc397462222)

[1.3 Dissertation Structure 4](#_Toc397462223)

[2 State of the Art / Related Work 5](#_Toc397462224)

[2.1 Ontology Learning 5](#_Toc397462225)

[2.2 Association rules 5](#_Toc397462226)

[2.3 Concept Relation Quantification 8](#_Toc397462227)

[2.4 Building & Construction 8](#_Toc397462228)

[3 Theoretical and Technical Foundation 9](#_Toc397462229)

[3.1 Association Rules 9](#_Toc397462230)

[3.1.1 Frequent Pattern Growth 12](#_Toc397462231)

[3.1.2 Association Rules Measurement 20](#_Toc397462232)

[3.1.3 Association Rules Concurrency 26](#_Toc397462233)

[3.2 Vector Space Model 26](#_Toc397462234)

[3.2.1 Term Weighting – The TF-IDF 27](#_Toc397462235)

[3.3 Similarity Measure in Information Retrieval 28](#_Toc397462236)

[3.3.1 Cosine Similarity Measure Algorithm 29](#_Toc397462237)

[3.3.2 Other Similarity Measures 30](#_Toc397462238)

[4 Building & Construction Domain Ontology 31](#_Toc397462239)

[4.1 Ontology 31](#_Toc397462240)

[4.1.1 Construction Methodology 31](#_Toc397462241)

[4.1.2 Ontologic enrichment Dynamics 31](#_Toc397462242)

[4.1.3 The E-COGNOS Framework (– Ontology in Building and Construction) 32](#_Toc397462243)

[5 Design and Implementation 33](#_Toc397462244)

[5.1 Tools and Technologies 33](#_Toc397462245)

[5.2 Requirements 34](#_Toc397462246)

[5.3 Conceptual & Technical Architectures 36](#_Toc397462247)

[5.4 Implementation and Development 37](#_Toc397462248)

[5.5 Conceptual & Technical Architectures 38](#_Toc397462249)

[5.5.1 Document Analysis 39](#_Toc397462250)

[5.5.2 FP-Growth 41](#_Toc397462251)

[5.5.3 Association Rules 41](#_Toc397462252)

[5.5.4 Frequent Itemset Mapping 41](#_Toc397462253)

[5.5.5 Ontology Enrichment 42](#_Toc397462254)

[5.6 Front end 42](#_Toc397462255)

[6 Evaluation 44](#_Toc397462256)

[6.1 Use cases 44](#_Toc397462257)

[6.2 Scientific publications 44](#_Toc397462258)

[7 Conclusion and Future Work 47](#_Toc397462259)

[8 Bibliography 49](#_Toc397462260)

[9 Appendices 53](#_Toc397462261)

[Figure 2.1 – Study fields examples 6](#_Toc397462262)

[Figure 3.1 – An FP-Tree example for the items in the transaction Table 3‑1. 15](file:///F:\Os%20meus%20documentos\Universidade\Dissertação\Escrita\working\inprocess\Dissertação%20v4%20(Tudo).docx#_Toc397462263)

[Figure 3.2 – The three initial trees at the end of the first three transactions 16](file:///F:\Os%20meus%20documentos\Universidade\Dissertação\Escrita\working\inprocess\Dissertação%20v4%20(Tudo).docx#_Toc397462264)

[Figure 3.3 – Prefix sub-paths for all frequent items 17](file:///F:\Os%20meus%20documentos\Universidade\Dissertação\Escrita\working\inprocess\Dissertação%20v4%20(Tudo).docx#_Toc397462265)

[Figure 3.4 – Conditional FP-Tree for item *professor* 19](file:///F:\Os%20meus%20documentos\Universidade\Dissertação\Escrita\working\inprocess\Dissertação%20v4%20(Tudo).docx#_Toc397462266)

[Figure 3.5 – Interestingness measures types tree (adapted from (Silberschatz and Tuzhilin, 1995)) 21](#_Toc397462267)

[Figure 5.1 - Tools and Technologies adopted 33](file:///F:\Os%20meus%20documentos\Universidade\Dissertação\Escrita\working\inprocess\Dissertação%20v4%20(Tudo).docx#_Toc397462268)

[Figure 5.2 - System requirements 35](#_Toc397462269)

[Figure 5.3 – Knowledge layer architecture 36](#_Toc397462270)

[Figure 5.4 – System Architecture – MVC Methodology 37](#_Toc397462271)

[Figure 5.5 – Rapidminer Main Process 38](#_Toc397462272)

[Figure 5.6 – Conceptual Architecture 38](#_Toc397462273)

[Figure 5.7 – Vector Creation Rapidminer Process 39](#_Toc397462274)

[Figure 5.8 – Document Analysis Pipeline 40](#_Toc397462275)

[Figure 5.9 – Entity Relation Model 42](#_Toc397462276)

[Figure 6.1 - Main requirements UUC 44](#_Toc397462277)

[Figure 6.2 - Front-End UUC 44](#_Toc397462278)

[Table 2‑1 - Examples of Fields of Study with Association Rules Application 7](#_Toc397462279)

[Table 3‑1 – Transaction table for frequent items in database 13](#_Toc397462280)

[Table 3‑2 – Paths table for frequent items 17](#_Toc397462281)

[Table 3‑3 – Frequent itemsets discovered for all items 19](#_Toc397462282)

[Table 5‑1 – Numerical to Binomial regulation 40](#_Toc397462283)

[Table 5‑2 – Association Rules Database Structure 41](#_Toc397462284)

**SYMBOLS AND NOTATION**

|  |  |  |
| --- | --- | --- |
|  | **AEC** | **A**rchitecture, **E**ngineering and **C**onstruction |
|  | **AI** | **A**rtificial **I**ntelligence |
|  | **API** | **A**pplication **P**rogramming **I**nterface |
|  | **AR** | **A**ssociation **R**ules |
|  | **ASP** | **A**ctive **S**erver **P**ages |
|  | **CSS** | **C**ascading **S**tyle **S**heet |
|  | **ECLAT** | **E**quivalent **CLA**ss **T**ranformation |
|  | **ERD** | **E**ntity **R**elation **D**iagram |
|  | **FI** | **F**requent **I**tem |
|  | **FP** | **F**requent **P**attern |
|  | **HTML** | **H**yper **T**ext **M**arkup **L**anguage |
|  | **IR** | **I**nformation **R**etrieval |
|  | **IT** | **I**nformation **T**echnology |
|  | **JB** | **J**ava**N**eans |
|  | **JDBC** | **J**ava **D**ata**B**ase **C**onnection |
|  | **JDOM** | **J**ava **D**ocument **O**bject **M**odel |
|  | **JSP** | **J**ava **S**erver **P**age |
|  | **JVM** | **J**ava **V**irtual **M**achine |
|  | **KD** | **K**nowledge **D**iscovery |
|  | **MVC** | **M**odel **V**iew **C**ontroller |
|  | **OWL** | **W**eb **O**ntology **L**anguage |
|  | **RDF** | **R**esource **D**escription **F**ramework |
|  | **SQL** | **S**tructured **Q**uery **L**anguage |
|  | **TF-IDF** | **T**erm **F**requency – **I**nverse **D**ocument **F**requency |
|  | **TID** | **T**ransaction **ID**entification |
|  | **UML** | **U**nified **M**odelling **L**anguage |
|  | **USD** | **U**ML **S**tate **D**iagram |
|  | **USQD** | **U**ML **S**e**Q**uence **D**iagram |
|  | **UUC** | **U**ML **U**se **C**ase |
|  | **VSM** | **V**ector **S**pace **M**odel |
|  | **W3C** | **W**orld **W**ide **W**eb **C**onsortium |
|  | **XML** | e**X**tensible **M**arkup **L**anguage |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |