Chapter 1 – Introduction

* How to represent and structure knowledge. (Talk about controlled vocabularies).
* Ontologies as a method to represent and structure information.
* Pattern Extraction and Data Mining Techniques
* Objective and approach of the work presented in this document.

Section 1.1 – Challenges

* Lack of existence of a pure approach to quantify relations discovered from unstructured information in documents, without help of an ontology.

Ontology learning is a problem because there are no pure automatic mechanisms. (Explain ontology learning??)

* What can be done to measure a relation and find its meaning?
* This document presents an approach to help discover relations in unstructured information in documents, knowing that there are no real methods to help measure a relation between two or more concepts.

Research question:

How to quantify semantic relations between concepts in a domain ontology, using external sources of non-structured information.

Hypothesis:

Semantic relations between concepts from a domain ontology, can be quantified by applying data mining techniques for pattern extraction into non-structured sources of information.

* Having a set of documents with unstructured information, how could meaning be discovered, in the way of relations between its concepts?
* How to discover the domain of a set of words?

Section 1.2 – Expected outcomes

Present the way that I will propose solutions to research questions.

* How to address the problems?
* What techniques to use?
* Why are these techniques used to solve the problems, and not others?
* Develop a system, proof of concept, to present the results to domain experts.

Section 1.3 – Context of work

* Falar onde foi desenvolvido o trabalho
* A sua ligação com os projectos europeus (e-Cognos e CoSPaces)
* Enquadramento da tese de doutoramento do Ruben e a minha contribuição para a mesma.

Section 1.4 – Document Structure

Chapter 2 – Controlled Vocabularies

* (What are they? What do they represent?)
* What forms of representation of information exist?
* Ontologies (Definition, Construction, relations, concepts)

What is an ontology? What is it utility? How to construct one? Languages to represent it.

* Relations (meaning)
* Concepts
* Ontology learning
* E-cognos (European project for the creation of an ontology in B&C domain).
* Application domain. (Practical cases in building and construction domain)

Chapter 3 – Pattern Extraction from unstructured information sources

* Data mining. (What is DM? Techniques used today?)
* Association Rules (Definition, Rules)
  + Algorithms to discover [ECLAT, APRIORI, FP-GROWTH]
  + Weaknesses/Strengths between them
  + Why FP-Growth?
* Application domain. (Practical cases where association rules are used)

Chapter 4 – Concept Model

- Explain conceptual model/solution

- Describe an application example

From unstructured information to knowledge representation and ontology structure

- Dimensions included in the model???

- Enrichment process

FP-Growth how to build and FP-Tree

Association rule evaluation

- DER / MVC / UML Diagrams

Chapter 5 – Model Design and Development (Proof of concept)

- Method proposal to address the question.

- What were the technologies used for the solution.

Technologies used,

- Implementation description.

(Present the server / front end solution)

- Include use cases (Relations discovered, new concepts discovered, etc.)

(Discover a relation between two concepts, update a relation between two concepts, and discover new concepts)

- Front end

Brief explanation of the functionality of the front work. Explain in a form of manual??

Chapter 6 – Assessment

* Present list of relations discovered and discuss them
* Present new concepts discovered

Chapter 7 – Conclusion and Future directions

- Evaluate if the goals reached success.

- Evaluate the achievement of the hypothesis

- Present the paper

Chapter 1 – Introduction

(Controlled vocabularies)

In many domains communication is the key to success. However, if the language is not common between all, the communication process could be in danger. One approach to address this question is to create mechanisms to define meaning and standardize the elements of communication. Since the appearance of information systems novel approaches were created to define meaning and represent knowledge. Controlled Vocabularies (CV) provide mechanisms to represent knowledge in a domain and facilitate a standardization for the communication process. These mechanisms are used to provide a clear and uniform meaning to the concepts used by everyone accessing the knowledge in a specific domain. For instance, the definition “*a piece of furniture with a flat top and one or more legs, providing a level surface for eating, writing or working at*” (Oxford University, 2006) must represent the same meaning to everyone. The reader can easily understand, that the former definition is referring to a *table*, and not to a *car*, or a *house*. Hence, when a community is talking about tables, everyone can refer to its definition.

Other responsibility of a CV is to provide precision to thoughts.

Why CV?? One of the main reasons of the use of CVs is the capacity to provide mechanisms

Why CV are important and what do they try to address/solve?

(Ontologies)

Ontologies are a type of CV that address the problems of representation and standardization of knowledge. Specifically, Gruber provides a definition for ontology as “*(...) a specification of a conceptualization. (misses something yet)*” (Gruber, 1993) In other words, an ontology represents a formal agreement in some domain for the representation of concepts with similar meaning. Ontologies provide meaning through relations between the concepts. These relations are measured with a numeric value that quantifies the strength of the relation. However, although ontologies provide a structure on concept representation, they are static. The improvement or update of the meaning in an ontology And this means that, there are

do not solve all the problems.

CV can be presented in different forms; among these are Thesaurus, Taxonomies, Folksonomies and Ontologies, just to name some of them. Specifically, Gruber provides a definition for ontology as “*(...) a specification of a conceptualization.*” (Gruber, 1993) In other words, ontology represents a formal agreement in some domain for the representation of words with similar meaning.

(Pattern extraction & Data Mining)

A Pattern can be seen as a recognizable repetition in information. For a system to be able to recognize and further extract these patterns, several processes can be used. Data Mining is one of them, and is a process to analyse and discover patterns and knowledge in data.

(This document approach)

The main objective of the work presented in this document is to discover and extract patterns in the form of knowledge from a set of documents with unstructured information. This is going to be made with the use of data mining techniques. This document will also propose an approach to help maintain and update controlled vocabularies, namely ontologies.

Section 1.1 – Challenges