Chapter 1 – Introduction

* How to represent and structure information. (Talk about controlled vocabularies).
* Ontologies as a method to represent and structure information.
* Pattern Extraction and Data Mining Techniques
* Objective and approach of 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.
* 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 (new way of ontology creation).
* Application domain. (Practical cases where association rules are used)

Chapter 3 – Pattern Extraction from 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

- Dimensions included in the model ???

- Enrichment process

- 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.

- Implementation description.

(present the server / front end solution)

- Include use cases (Relations discovered, New concepts discovered, etc)

- Front end

Chapter 6 – Assessment

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

Chapter 7 – Conclusão e perspectivas futuras

- Evaluate if the goals reached success.

- Evaluate the achievement of the hypothesis

- Present the paper

Chapter 1 – Introduction

(Controlled vocabularies)

Since the appearance of information systems, novel approaches were created to represent knowledge in the form of words. One of the main difficulties of this task was to find a method to represent the words in such form, that the same word could have the same meaning in the same domain amongst every expert and non-expert person. Controlled Vocabularies (CV) provide mechanisms to represent words in a domain and standardize a method of word representation that could bring the same meaning to everybody that needs to access the knowledge in a specific domain.

(Ontologies)

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 techniques can be used. Data Mining is one of them, and is a process to analyse and discover patterns and knowledge inside a database.

The convergence of computing and communication has produced a society that feeds on information. Yet most of the information is in its raw form: data. If data is characterized as recorded facts, then information is the set of patterns, or expectations, that underlie the data. There is a huge amount of information locked up in databases-information that is potentially important but has not yet been discovered or articulated. Our mission is to bring it forth.

Data mining is the extraction of implicit, previously unknown, and potentially useful information from 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. 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 structures of knowledge, namely ontologies.

Section 1.1 – Challenges