Chapter 1 – Introduction (planning)

[Flow: Communication Process->Vocabularies->Controlled Vocabularies->Goals of CV->Ontology->Ontology problems->Ontology learning->Pattern Extraction/Knowledge discovery->Approach to enrich an ontology based in data mining techniques.]

1P - Communication process/Languages/Vocabularies

2P - Controlled vocabularies, what are they?

3P - What problems CVs address?

* words with similar meaning and different spelling (Synonym),
* words with the same spelling and different meaning (Homograph)
* Standardization of the vocabulary of a community, one term represent only one concept

4P - Ontology as a form of CV, what is an ontology?

Why use an ontology?

5P - Problems of ontologies.

Static

Require previous agreement on the vocabulary

Require high maintenance

6P - (Automatic maintenance) Ontology learning, to help on maintenance of an ontology

What processes/techniques exist?

7P - Pattern Extraction and Data Mining Techniques to help on ontology learning and knowledge discovery

8P - What will I propose? Approach to discover knowledge in unstructured documents.

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 projetos 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.

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

Chapter 3 – Pattern Extraction from unstructured information sources

* Data mining / Knowledge Discovery. (What is DM/KD? 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 end. 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

1

# Introduction

With the exponential growth of data in the web is getting harder to search through it and retrieve the intended goal. In order to easy and fasten the search of relevant data, this data needs organization. Although

accessibility is harder to achieve. rises the necessity for computer systems organize the data in a way understandable by computers and humans. In order to do this, the systems first need to recognize the contents that are represented in the data. In the present days, computers can recognize a set of terms or words, a vocabulary.

translate the data and recognize its contents. For instance, to a computer, a text document is just a set terms each separated by a space or some sort of punctuation.

In many domains communication is the key to reach success. Therefore, the language used must be common between and understood by all community members, otherwise the communication process could be ineffective. In this sense, each language has its own vocabulary. Vocabulary is a set of terms that represent concepts and the connection between them. Vocabulary is a representation in the form of words in order to provide meaning to human thoughts. Nevertheless, vocabularies itself present several weaknesses. Word ambiguity is one of them. Vocabularies have several words that represent the same concept. Similarly, other weakness of vocabularies is the representation of several different concepts by a single word.

In order to address these weaknesses and ease the communication process, one approach can be considered. This approach is the creation of mechanisms to define meaning, to structure and to standardize the elements of communication, designated as Controlled Vocabularies (CV). CVs provide mechanisms to represent concepts consistently in a domain. CVs also provide a standardization for the communication process, providing a structure for all elements included. This structure have to be agreed beforehand by the community.

The main goal of CVs is to provide a clear and uniform meaning to the concepts used by a community accessing the knowledge in a specific domain. They are one support to help every community member understand the concepts of the community language itself in the same way. In a shared working environment, all members should speak the same language, therefore CVs provide this precision, giving to each concept the same meaning. When a concept is referred by a member, every member knows exactly what its meaning is.

Specifically, CVs try to solve some issues in the vocabulary, in order to remove ambiguity and provide precision and consistency of the concepts in each community. Synonym words, different words with the same meaning, is an issue that CVs address with its mechanisms. They define the concept meaning through a relation between two or more words, and gather in a list of terms every word with the same meaning. Other issue that CV address is a word with the same spelling but different meanings, namely homograph words. CVs deal with this issue by having each concept described by only one authorized or controlled term.

To create a CV, there are several paths to do it. Amongst others, Ontologies is one of the paths. Ontology is a type of CV that addresses the problems of consistent representation and standardization of knowledge. Specifically, Gruber provides a definition for ontology as “*(...) a formal specification of a shared conceptualization of a domain of interest.*” (Gruber, 1993) In other words, an ontology represents a formal agreement, where formal implies that it has to be machine readable, and agreement implies an arrangement by the community members for the representation of concepts with similar meaning related to a domain by the same term.

Ontologies provide a structure of terms to represent concepts, and also provide a hierarchy in order to define the relations between the concepts. These relations provide means to quantify a relation and its strength. In this sense a concept can be more related to a term A than to other term B, if the numeric value of the relation between the concept and term A is higher. Although ontologies provide structures for concept representation, they are static. Additionally, the maintenance necessary in an ontology consumes a lot of time and other resources, such as a specialized administrator to do it. Other requirement for an ontology that could be a problem, is the agreement of the concepts inside it by the community. This process also requires time to do it, and this could bring delay to the creation of an Ontology.

One of the areas that deal with the creation and maintenance of an ontology, is referred to as Ontology Learning. Ontology Learning (OL) is the area that studies the mechanisms and processes to transform heavy tasks like the creation and maintenance of an Ontology, into a semi or complete automatic process. This area still has a long path ahead. Although semi-automatic methods were found that still requires the help of an ontology expert, for example, to validate a new concept, the idea of a pure automatic method that could enrich an ontology is not yet in a near future.

One of the motors that drive OL itself is the recognition of patterns in the data that could gather interesting information (knowledge) to further evaluation. A pattern, specifically in semantic domain, can be seen as a predictable occurrence that repeats itself along some text data. Knowledge itself, can be seen as information discovered in some data that could be interesting or useful for the domain applied. For instance, this could be information not yet known or unpredictable in the domain.

For a system be able to recognize patterns and further extract useful and valuable information (knowledge) from data, several processes can be used. Data Mining (also referred in literature as Knowledge Discovery in Databases or KDD) is one of them, and is a process to analyse and discover patterns and knowledge in data. Data mining allows experts to find knowledge in new data or data they already have. Additionally, with data mining techniques, decision makers can use the new knowledge that otherwise could be unknown or unavailable, to make better decisions.

The main objective of the work developed in this dissertation is to present an approach that could help businesses undertake this decision making process more easily. This approach is made through a technique aimed to discover and extract patterns and discover knowledge from unstructured information in a set of documents. This is going to be made taking advantage of data mining techniques. Additionally, is also proposed an approach to help maintain and update CVs, namely ontologies. Furthermore, a proof of concept to represent this approach, referred as DOKS (Dynamic Ontology learning with Knowledge sources from unstructured text System), was developed by the author of the present dissertation.

## Challenges

One of the biggest challenges in information systems when constructing a controlled vocabulary is to find the meaning and relations between concepts and ideas. This dissertation propose an approach to solve this issue based in the following research question:

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

This question raises the following hypothesis that this project aims to prove:

**Semantic relations between concepts from a domain ontology, can be quantified by applying data mining techniques for pattern extraction and knowledge discovery into unstructured sources of information.**

## Expected Outcomes

With this work, the author expects to achieve the following goals.

Starting with a set of documents with unstructured data as content, the first goal will be to discover patterns in this data. This will be made by the use of an algorithm, namely FP-Growth (FP-G). FP-G is an algorithm that discovers frequent patterns in sets of data. After research, FP-G presented as the one that provide better performance.

The second goal will be, from a set of structured information, with frequent patterns, represented by concepts, already discovered and extracted, that one can discover relations (or associations) between them. Association Rules (AR) is the technique that will be used to execute this task.

The third goal is to discover new concepts from the set of unstructured data, to enrich a domain ontology. This will take advantage of FP-G and AR techniques also, in two steps. Step one will be to discover the frequent words in text data. Step two will be to compare this words with all terms in an ontology and discover the ones that are not present in the ontology.

The fourth goal is the discovering of knowledge in the set of data that could be useful to help on OL process. This is made taking advantage of a set of metrics applied in the AR algorithm. This will provide the possible context and domain of a concept.

The fifth and last proposed goal is to develop a proof of concept, a software system, based in the previous challenges in order to show the results in an understandable form. Additionally, is to turn this process of OL into a pure automatic OL process. Starting from a set of documents received to knowledge discovered for OL.

Moreover, another goal is to write a scientific document with the development of the present work that could be approved by the academic community.

## Section – Context of work

The context of the present work arises from the Innovative Collaborative Work Environments for Individuals and Teams in Design and Engineering project (CoSPaces). CoSPaces (EC-IST FP6) was a project with funds from European Community (EC) with the goal to provide digital solutions in a collaborative workspace between individuals, teams and enterprises. The project expected to achieve the former by improving collaboration methods, like human communication and knowledge sharing support, taking advantage and improving existing IT systems.

The contribution of this dissertation to CoSPaces is related to the Component of Collaboration Tools, specifically the Knowledge Management sub-component. This sub-component is responsible to provide knowledge capable functionalities to CoSPaces main goal.

The present work described in this dissertation was also part of a PhD Thesis, namely “*Semantic enrichment of knowledge sources supported by domain ontologies*”, whose main goal was to “*introduce a novel conceptual framework to support the creation of knowledge representations based on enriched Semantic Vectors, using the classical vector space model approach extended with ontological support*” (Costa, 2014). The contribution for this thesis was the proposal of an ontology learning method based in knowledge discovery techniques.

This work from this dissertation also adapt some resources from, SEKS (Figueiras, 2012), namely the ontology adopted.

## Section – Document Structure

Following this brief introduction in Chapter 1 with the presentation of the problem, the goals that the author of the present document expects to achieve and the contextualization of the work, this dissertation will be guided by the following structure.

In Chapter 2, the domain of study is Controlled Vocabularies. Ontology will be the selected CV to present. It will be explained how to build one. The existent formalisms to represent them and where are they used.

Chapter 3 will explain what is data mining and knowledge discovery, and present techniques to discover patterns in unstructured data. One of them, Association Rules will be explained in more detail. FP-Growth, and the concurrent algorithms to discover patterns will be compared, and explained why the former was chosen.

In the next chapter can be observed the explanation for the solution proposed. Thus, Chapter 4 will present the concept model, an application example describing the steps from the unstructured information to knowledge representation and ontology learning. This chapter also includes the methodology behind FP-Growth and the evaluation of an Association Rule.

With Chapter 5, one can expect to read about the development of a proof of concept. The design and development of a model, with the proposal method to address the question. This will be described with the technologies used, following a description of the implementation and use cases. The framework developed will also be presented in this chapter.

Chapter 6 will be the assessment of the solution proposal, and Chapter 7 will present some conclusions from the author, and some possible future directions in this area.

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)

# Controlled Vocabularies

2