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

# Contents

[Contents I](#_Toc414948269)

[Figures III](#_Toc414948270)

[Tables IV](#_Toc414948271)

[Symbols and Notation V](#_Toc414948272)

[1 Introduction 1](#_Toc414948273)

[1.1 Challenges 4](#_Toc414948274)

[1.2 Expected Outcomes 4](#_Toc414948275)

[1.3 Context of work 5](#_Toc414948276)

[1.4 Document Structure 6](#_Toc414948277)

[2 Controlled Vocabularies 9](#_Toc414948278)

[2.1 Controlled Vocabularies – Definition 9](#_Toc414948279)

[2.1.1 Problems Address by a CV 11](#_Toc414948280)

[2.1.2 Advantages / Disadvantages of CV 12](#_Toc414948281)

[2.1.3 Types of CVs - Differences and strengths (/ and characteristics) 13](#_Toc414948282)

[2.2 Ontology – What is it? (Definition) 16](#_Toc414948283)

[2.2.1 Ontology – What is it for? 16](#_Toc414948284)

[2.2.2 Concepts and ideas 17](#_Toc414948285)

[2.2.3 Relations / meaning 17](#_Toc414948286)

[2.2.4 Ontology Building 17](#_Toc414948287)

[2.2.5 Ontology Languages 17](#_Toc414948288)

[2.3 Maintenance of ontologies - Ontology Learning 19](#_Toc414948289)

[3 Pattern extraction from non-structured sources of information 21](#_Toc414948290)

[3.1 Data mining / Knowledge Discovery. 21](#_Toc414948291)

[3.1.1 Definition - What is DM/KD? 21](#_Toc414948292)

[3.1.2 Techniques used today? 21](#_Toc414948293)

[3.2 Pattern Discovery 21](#_Toc414948294)

[3.2.1 FP-Growth – definition 21](#_Toc414948295)

[3.2.2 Algorithms to discover patterns Weaknesses/Strengths between them 21](#_Toc414948296)

[3.2.3 Association Rules (Definition, Rules) 21](#_Toc414948297)

[4 Concept Model 23](#_Toc414948298)

[5 Model Design and Development 25](#_Toc414948299)

[6 Assessment 27](#_Toc414948300)

[7 Conclusion and Future directions 29](#_Toc414948301)

[7.1 Work overview 29](#_Toc414948302)

[7.2 Research Contributions 30](#_Toc414948303)

[7.3 Future Directions 33](#_Toc414948304)

[8 Bibliography 37](#_Toc414948305)

# Figures

[Figure 1 - Word search example (Yahoo, 2015) 10](#_Toc414948248)

[Figure 2 - Vocabulary Example 14](#_Toc414948249)

[Figure 3 - Domain Ontology example (Innovation Ontology adapted from (Stick-iSchool, 2013) ) 15](#_Toc414948250)

# Tables

[Table 1 - Examples of a) Homophone, b) Homograph and c) Synonym words 11](#_Toc414948231)

# Symbols and Notation

|  |  |  |
| --- | --- | --- |
|  | **API** | **A**pplication **P**rogramming **I**nterface |
|  | **AR** | **A**ssociation **R**ules |
|  | **B&C** | **B**uilding and **C**onstruction |
|  | **CV** | **C**ontrolled **V**ocabulary |
|  | **DB** | **D**ata**b**ase |
|  | **DM** | **D**ata **M**ining |
|  | **DOKS** | **D**ynamic **O**ntology learning with **K**nowledge sources from unstructured data **S**ystem |
|  | **ECLAT** | **E**quivalent **CLA**ss **T**ranformation |
|  | **FI** | **F**requent **I**tem |
|  | **FIM** | **F**requent **I**temset **M**apping |
|  | **FP** | **F**requent **P**atterns |
|  | **HTTP** | **H**yper**T**ext **T**ransfer **P**rotocol |
|  | **IT** | **I**nformation **T**echnology |
|  | **KDD** | **K**nowledge **D**iscovery in **D**atabases |
|  | **OL** | **O**ntology **L**earning |
|  | **OWL** | **W**eb **O**ntology **L**anguage |
|  | **PHP** | **H**ypertext **P**re**P**rocessor |
|  | **RDBMS** | **R**elational **D**ata**b**ase **M**anagement **S**ystem |
|  | **SEKS** | **S**emantic **E**nrichment of **K**nowledge **S**ources |
|  | **XML** | e**X**tended **M**arkup **L**anguage |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

1

# Introduction

The exponential growth of available information in digital format created the need to discover ways to organize it, in order to be easily accessible. First search engines were essentially word-based, meaning that the results provided by the search process could only be achieved if documents had in their bodies exactly the same words being searched for (Lei et al., 2006). The evolution of search engines motivated by the fact that a simple search by term for the information could not be enough, as the set of terms, or vocabulary available in information being searched could be different from the vocabulary being used. Therefore, it was of great importance to discover approaches for the representation of ideas (concepts), and not just the representation of terms, aiming at getting better results for queries (Almeida and Souza, 2011).

Nowadays, computers systems can represent sets of terms or words (also referred to as vocabularies). However, vocabularies themselves, do not represent ideas or concepts, they just represent words. In order to represent concepts and ideas, one approach can be considered. This approach is the use of mechanisms to represent more than pure words, to represent concepts. These mechanisms are referred to as Controlled Vocabularies (CV) (Lima et al., 2007). CVs are defined subsets of terms from a natural language (e.g. Esperanto), or can be pure symbols of any sort (e.g. sequence of digits) used to represent concepts, with some sort of organization. CVs represent the concepts by assigning to each, one or more words, or phrases and some describing properties that both translates its meaning. CVs also describe if or how a concept is related to other concept.

Natural languages are very rich in their vocabulary properties. They can have different meanings represented by the same word (Homograph words), in several contexts. Also, there are words that can be pronounced in the same way, however have different spelling and meaning (Homophone words). Homograph and Homophone words can lead to ambiguity and confusion when using the terms by people. CVs address the problems of Homograph and Homophone words solving them by assigning each term to just one concept, and adding properties to explain and provide a better meaning to each concept. For instance, the word “board” can represent a base used in a classroom to write with chalk, or can represent a platform to use in snow sports to ride on top of a mountain hill covered with snow. The way CVs deal with this, is by adding some properties that will increase the precision of the meaning of each term, reducing the ambiguity when these words are used. (N.I.S.O. (US) and others, 2005)

An Ontology is a type of CV that addresses problems like the consistent representation or word ambiguity in information. According to Gruber (Gruber, 1993) an ontology is “*(...) a formal specification of a shared conceptualization of a domain of interest.*” In other words, an ontology represents a formal agreement, where *formal* implies that it has to be machine readable, and *agreement* implies a shared understanding of meaning on the ontological concepts. An Ontology is used when there is the need to share or exchange knowledge within a given domain. Ontologies can be represented as a hierarchically structured set of concepts describing a specific domain of knowledge.

Although ontologies provide structures for concept representation, they face some challenges (Uschold and Gruninger, 1996). So why use an ontology? Inside an organization people from different domains can have different points of view and different words to communicate. In this sense the benefits of using an ontology is to be able to provide a common ground that can lead to a shared understanding for the same concepts. Additionally, when two IT systems need to exchange knowledge, ontologies provides them inter-operability features in order to ease the integration between them (Pouchard et al., 2000). Furthermore, ontologies are useful when there is the need to reuse its contents and features. There is no need to re-invent the wheel (Gangemi and Presutti, 2009).

Ontology Learning (OL) deals with the creation and maintenance of an ontology, and studies the mechanisms and processes to transform heavy tasks like creation and maintenance of Ontologies, into a semi or complete automatic process. IT is worth noting that relevant literature already presents first results on automatic maintenance of ontologies, but still in a very early stage. Human-based processes are still the current way to update and maintain ontology growth (Zhou, 2007).

One of the motors that drive OL itself is the recognition of patterns in the data that could originate new knowledge to further evaluation. For instance, this could be learned from some information not yet known or unpredictable in a specific domain. A pattern, in the area of information retrieval and text mining, can be defined as a predictable occurrence that repeats itself along some text data. Furthermore, Knowledge is defined as “*awareness, familiarity, or understanding of someone or something (e.g. facts, information, descriptions or skills), acquired through experience or education by perceiving, discovering or learning*.” (Oxford University, 2006) Therefore, OL provides techniques to discover knowledge.

Several processes can be used for a system be able to recognize patterns and further extract knowledge from data and information. Data Mining (also referred in literature as Knowledge Discovery in Databases or KDD) is one of them (Hand et al., 2001). Data mining allows experts to find knowledge in new data or data they already have. Additionally, by adopting data mining techniques, it is expected that decision makers can use new knowledge that otherwise could be unknown, unavailable or difficult to discover, to make better decisions. (Witten et al., 2011)

Having settled the context, urge to say that this dissertation aims at proposing an approach to support part of the process of ontology learning. Specifically, the proposed approach adopts a mechanism suitable for the use of data mining techniques for pattern discovery and extraction, and knowledge discovery from unstructured sources of information from a document corpus. Additionally, it is also proposed an approach to help maintain and update CVs, namely domain ontologies, with the previous discovered knowledge. This means: (i) to discover concepts and relations between them; (ii) to propose an approach to quantify these relations; (iii) to discover new concepts; And finally, (iv) to take advantage of (i), (ii) and (iii) results to update a domain ontology. Furthermore, a proof of concept to characterize this approach, referred as DOKS (Dynamic Ontology learning with Knowledge sources from unstructured text System), is also part of the results produced.

## Challenges

One of the biggest challenges in information systems when constructing a CV is to find both meaning and relations among concepts and ideas. Furthermore, how to say that a concept is more related to one, than it is to other concept? How to quantify this relation? Similarly, other challenge is to discover knowledge in sources of information that could be later used, for instance, to update a CV. Moreover, is it possible to fully automate this process? Still, other challenge identified relates to the limited amount of information that is inside a single document. This dissertation proposes an approach to help solving these challenges based in the following guiding question:

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

That question highlights the research path leading the development of this work, as follows:

**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

When a study is made, there is a need to consider its contribution and applicability that can arise from it. In this sense, the expected outputs to be provided by this work are the following ones:

* To develop a method to describe how to extract concepts and recognize relations between them from a data document corpus, and to find new knowledge sources in order to update a domain ontology.
* To develop a proof of concept, a software platform, based in the previous method in order to reflect the application of the studied techniques.
* Present results of the semi-automatic OL process. Results composed by patterns discovered in the documents, their relations and the new concepts discovered. They should be presented in an understandable way to the user.
* Finally, publication of scientific documents about the work, to be assessed by the academic community.

## Context of work

The context of the present work arisen from three MSc. Dissertations (Antunes, 2010; Figueiras, 2012; Parada, 2010) in the area of Data Mining and Knowledge Sources. These studies provided the background and inspiration for the reasoned path choice of the present work. The setting made through these studies was provided by CoSPaces. CoSPaces was an European Research project aiming 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.

EU research project E-Cognos was an inspiration in CV domain. Specifically, it provided the insight and methodology needed to build a domain ontology. Also, provided the ground for the structure representation of the semantics in an ontology applied in the B&C sector.

This work takes advantage of the application domain background based in the Building & Construction sector, which provided the knowledge sources, specifically technical documents (e.g. reports and papers) to be used. They were adopted from (Costa, 2014), a PhD Thesis, that also received a contribution from this study. 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*”. The respective contribution was the proposal of an ontology learning method based in knowledge discovery techniques.

SEKS (Figueiras, 2012) also provided some resources which were adopted in this work, namely the domain ontology manipulation libraries.

The applicability context of the present work relied in B&C sector, as it was the domain that provided the resources and inspiration. However, in a more abstract sense, the contribution made here can be further used wherever there is the need of a shared communication and understanding of concepts, and in all the fields where knowledge and domain ontologies can be used.

## Document Structure

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

In Chapter 2, Controlled Vocabularies are the domain of study. Ontology will be the selected CV discussed. It will be explained in more detail what is an Ontology and how to build one. Additionally, it will be presented some existent formalisms to represent them and where are they used.

Chapter 3 will explain what is data mining and knowledge discovery, and describe techniques to discover patterns from non-structured 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 following chapter, can be observed the explanation for the solution proposed. Thus, Chapter 4 will present the concept model, an application example describing how to reach from non-structured 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 proposed 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 the evaluation of the results. Chapter 7 will present some conclusions from the author, namely an overview of this dissertation, the achievement of the proposed outcomes, some possible future directions in this area and some scenarios where this work could be an asset.

2

# Controlled Vocabularies

In this chapter it will be presented an introduction to some concepts and definitions about Controlled Vocabularies. Moreover, it will be described forms of representation of information. In particular it will be given special attention to Ontologies. Furthermore, it will be explained how to represent a relation and a concept or idea into an information system and how ontologies use them. Additionally, a more in depth overview of Ontology Learning will be presented in order to understand better what is it and how does it works. Lastly, a brief insight to the project that inspired the idea of Ontology use in the present work, the E-Cognos European project.

## Controlled Vocabularies – Definition

The exponential growth of available information in digital format created the need to discover ways to organize it, in order to be easily accessible. First search engines were essentially word-based, meaning that the results provided by the search process could only be achieved if documents had in their bodies exactly the same words being searched for (Lei et al., 2006). For instance, if one wanted to do a query on a common search engine (e.g. Yahoo, Google, Bing) for the word “*car*”, each result would need to explicitly contain the word searched for. (Figure 1)

The evolution of search engines was motivated by the fact that a simple search by term for the information could not be enough, as the set of terms, or vocabulary available in information being searched could be different from the vocabulary being used. Referring to the example from Figure 1, it is shown that if a user could query a search engine for the concept “*road vehicle, typically with four wheels, powered by an internal-combustion engine and able to carry a small number of people”*(Oxford University, 2006) represented by the word “*car*”, consequently the results would include the ones containing just the search term. Although, the terms “*automobile*” and “*vehicle*” could also describe the same concept. Alternatively, if the term used to search the same concept was “*automobile*”, the results would be different. Therefore, it was of great importance to discover approaches for the representation of ideas (concepts), and not just the representation of terms, aiming at getting better results for queries (Almeida and Souza, 2011). In this sense, the results provided by the query of Figure 1 example would be an overlap of the terms *“car”*, *“automobile”* and *“vehicle”*.



Figure 1 - Word search example (Yahoo, 2015)

Nowadays, sets of terms or words (also referred to as vocabularies) can be represented in computers systems. However, vocabularies themselves, do not represent ideas or concepts, they just represent words. Vocabularies are just a word lists with no specific organization. Also, words *per se* are just units of a language, they have the responsibility to be the carriers of meaning. One can only understand an idea represented by a word when a meaning is associated to the word, as the meaning is itself the idea that a person wants to express when using that word.

One approach can be considered, in order to represent concepts and ideas. This approach is the use of mechanisms to represent more than pure words, to represent concepts. These mechanisms are referred to as Controlled Vocabularies (CV) (Lima et al., 2007). CVs are defined subsets of terms from a natural language (e.g. Esperanto), or can be pure symbols of any sort (e.g. sequence of digits) used to represent concepts, with some sort of organization. CVs represent the concepts by assigning to each, one or more words, or phrases and some describing properties that both translates its meaning. CVs also describe if or how a concept is related to other concept.

*Controlled* refers to (i) the fact that the vocabulary used needs to be organized based in some logic structure (e.g. Alphabetically, Hierarchically, etc.) and defined based in the association of each term to just one meaning, reducing the ambiguity and improving the consistency of a CV; (ii) the fact that the constraints in its use are bigger than in a natural language; (iii) the fact that the access control to the maintenance of the terms of a CV by the people is restricted. It can have different restrictions to different users (e.g. Normal user, Expert user, Admin user). For example, just expert users can propose new words for a CV and just admin users can add new words.

### Problems Address by a CV

Natural languages are very rich in their vocabulary properties. They can have different meanings represented by the same word (Homograph words), in several contexts. Also, there are words that can be pronounced in the same way, however have different spelling and meaning (Homophone words). Homograph and Homophone words can lead to ambiguity and confusion when using the terms by people. (Refer to Table 1 a) and b) for examples of Homophone and Homograph words respectively)

Table 1 - Examples of a) Homophone, b) Homograph and c) Synonym words

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Homophone |  | Homograph | |  | Synonym | |
| **Words** |  | **Word** | **Meaning** |  | **Word** | **Synonyms** |
| Board, Bored |  | Advocate | Speak or write in support of/Person who supports cause of another person |  | Car | Vehicle, Automobile |
| Dual, Duel |  | Board | Base to write/Platform to ride in snowboard |  | Couch | Sofa, Divan, Chair |
| Loan, Lone |  | Desert | A hot, arid region/To leave |  | Garbage | Trash, Junk, Waste |
| Made, Maid |  | Evening | Late afternoon/Making more even |  | Honest | Honourable, Fair, Sincere, Trustworthy |
| Sloe, Slow |  | Match | Wood stick to ignite fire/Sporting event |  | Intelligent | Smart, Bright, Brilliant, Sharp |
| Peak, Peek, Peke, Pique, Pick |  | Object | Thing to see or touch/Goal |  | Vocabulary | Dictionary, Terminology, Glossary |
| Rain, Reign, Rein |  | Tear | To rip/A drop of water from the eye |  | Woman | Lady, female, girl |
| a) |  | b) | |  | c) | |

CVs address the ambiguity problems of Homograph and Homophone words solving them by assigning each term to just one concept, and adding properties to explain and provide a better meaning to each concept. For instance, the homograph word “*board*” (Table 1-b)) can represent a base used in a classroom to write with chalk, or can represent a platform to use in snow sports to ride on top of a mountain hill covered with snow. Likewise, the homophone words “*made*” and “*maid*” have the same pronunciation, although the first refers to the conjugation of the verb make in the simple past tense and past participle, and the second refers to a female servant. How can a search system (e.g. Yahoo!, Bing from Microsoft or Google) deal with homograph words by being able to distinguish two different meanings that can be represented by the same word? How can a voice recognition system (e.g. Cortana from Microsoft, Siri from Apple or Google Voice Search) deal with homophone words by being able to recognize accordingly two words that sounds the same? CVs deal with this by adding some properties that will increase the precision of the meaning of each term, reducing the ambiguity when these words are used. (N.I.S.O. (US) and others, 2005).

Additionally, natural languages have more properties that must be addressed by a CV. There are words that have the same or similar meaning and have different spelling. These words are referred to as Synonym words (refer to Table 1c). As a result, a concept can be represented by more than one word. Referring to Figure 1 example, one can infer from it that the concept represented by the word “*car*” have also some other words that can represent them, as any of its synonyms “*vehicle*” or “*automobile*”. A CV must allow the use of synonym words, as different people can use different vocabulary for the same concept. And in this case, a query for any of the words from the same concept must return the same or similar results.

### Advantages / Disadvantages of CV

When looking at the advantages from the use of CVs, the following can be enumerated:

* CVs improve the efficiency and precision of retrieval systems;

By providing more than one possible term to search for a concept, it allows for results that do not explicitly contain the search term and still are somehow related to it. For instance, a search by cars would provide results containing the word *“cars”* and also the results containing its synonyms words, like *“automobile”* and *“vehicle”*. Likewise, by limiting the terms that can be used, providing a more objective search through the terms used. For example, if one searches documents about *“football”* would not get documents about *“cars”*.

* Removes the ambiguities from natural languages;

Natural languages associate a word to more than one meaning. Consequently, it is hard for an information system to know what the user wants to search. As a result, each term is associated to a specific and unambiguous meaning.

* Activates the semantic search, meaning that the search will be made by idea and not by word. Through the use of its associated terms. In other words, means that the terms used do not need to explicitly be in the data searched for.
* Improve communication through peers in a community (eg. Between two civil engineers). When everybody that uses a CV know the terms to use when referring to a concept, it allows better communication through all people involved in CV use.

In the Disadvantages side, the following can be found:

* CVs cost time & money to build;

Building a CV takes time. First to gather all concepts necessary and related to a specific domain; Second, to find and associate each term to a specific concept. These tasks takes its time. And companies want fast revenue from its investments ant short term, meaning that CV can be an asset that does not yet provide good return to a company.

* Allow Human/Expert error;

As the concepts are gathered by humans, most of the times experts in a domain. This suggests that humans are prone to errors, even experts. And these error can lead to an imprecise and bad formed CV.

### Types of CVs - Differences and strengths (/ and characteristics)

Nowadays there are several ways to represent information in retrieval systems. One of them are CVs. CVs can be divided by complexity, use and level of control.

The simplest form of CV is a **Vocabulary**, a list of words or terms without any specific organization or logic (Figure 2). Although a vocabulary can have some uses, in a retrieval system, most of the times just the words are not enough for semantic retrieval purposes.



Figure 2 - Vocabulary Example

When a definition is added to each word from a vocabulary, it becomes a Dictionary. **Dictionary** is a vocabulary, or a list of words alphabetically ordered which contains the source of all meaning. Each word has its own meaning associated.

A **Taxonomy** is a structured vocabulary that adds a hierarchical layer to the dictionary. Each term is in a parent-child-based structure. It is through a taxonomy that the association between words can be introduced through the parent-child hierarchy. It provides a structured classification mechanism for each term from a vocabulary.

Adding another type of layer between terms, a sibling-based (on the same hierarchically level) structure, to a taxonomy, result in a **Thesaurus**. A Thesaurus takes advantage of the taxonomy structure and its hierarchy associations, adding semantic relations to its terms. These associative relations are in the form of synonyms.



Figure 3 - Domain Ontology example (Innovation Ontology adapted from (Stick-iSchool, 2013) )

**Ontology** is the most complex form of a CV (Figure 3). An Ontology includes a vocabulary, vertical or horizontal relations, and properties in each word. (Please refer to section 2.2 for more information about ontologies)

In brief, one can find some arguments that are common to every type of CV defined above:

CVs are mechanisms to structure, classify and represent terms or concepts;

CVs allow a community to agree and use the same terms in the same way;

#### Uncontrolled Vocabularies

It is worth mentioning another kind of representation structure that is not a controlled vocabulary, however is still a managed vocabulary. It is referred to as **Folksonomy**, and is an **uncontrolled vocabulary**. Folksonomy provides a user the possibility to associate any word he considers adequate to any information element (e.g. documents). This could be understood as the possibility to custom the information from each entity (e.g. person, company, etc) and adapt to the context of such entity. One of differences between a CV and a folksonomy is the control factor. For a folksonomy there is less control over the vocabulary, meaning that the control is not made by experts as in a CV; on the contrary, the control is made with the people that uses it every day. Although, a folksonomy can be customized by each user, it can be dangerous in the sense that the information is not validated by someone who has the expertise knowledge.

A folksonomy relies on what is referred to as Social Tagging (INSERIR REF), which is relied in the users from It is a nouvelle representation mechanism that takes advantage of users and social networks to help classify words and building a vocabulary. This new form of representation is more user-oriented in contrast to the CVs which are more standard-oriented.

## Ontology – What is it? (Definition)

An Ontology is a type of CV that addresses problems like the consistent representation or word ambiguity in information. In the

According to Gruber (Gruber, 1993) an ontology is “*(...) a formal specification of a shared conceptualization of a domain of interest.*” In other words, an ontology represents a formal agreement, where *formal* implies that it has to be machine readable, and *agreement* implies a shared understanding of meaning on the ontological concepts. An Ontology is used when there is the need to share or exchange knowledge within a given domain. Ontologies can be represented as a hierarchically structured set of concepts describing a specific domain of knowledge.

### Ontology – What is it for?

Although ontologies provide structures for concept representation, they face some challenges (Uschold and Gruninger, 1996). So why use an ontology? Inside an organization people from different domains can have different points of view and different words to communicate. In this sense the benefits of using an ontology is to be able to provide a common ground that can lead to a shared understanding for the same concepts. Additionally, when two IT systems need to exchange knowledge, ontologies provides them inter-operability features in order to ease the integration between them (Pouchard et al., 2000). Furthermore, ontologies are useful when there is the need to reuse its contents and features. There is no need to re-invent the wheel (Gangemi and Presutti, 2009).

### Concepts and ideas

A concept is defined as “An abstract idea; a general notion; an idea formed by mentally combining all its characteristics or particular” (Oxford University, 2006). In other words, a concept is an idea that can be difficult to understand and is constructed in someone mind. It can be an object, it can be a place, it can be whatever can be on a mind.

### Relations / meaning

Meaning is the concept that is represented by a word, phrase. Is the idea that a person desires to express through the use of words, signs, pictures, etc.

A relation in an ontology is a connection between two or more concepts.

### Ontology Building

Taxonomy/hierarchy, concepts, relations, properties

### Ontology Languages

There are several formalisms defined that can provide representation of information in an ontology. Table X will provide a list just for demonstration purposes.

((Lima, 2004) cwa4)

|  |  |  |
| --- | --- | --- |
| Language | Description | URL |
| DAML+OIL | DAML+OIL is a semantic markup language for Web resources. It builds on earlier W3C standards such as RDF and RDF Schema, and extends these languages with richer modelling primitives. DAML+OIL provides modelling primitives commonly found in frame-based languages. It is important to emphasise that this language was the basis of OWL. | http://www.w3.org/TR/daml+oil-reference |
| EXPRESS / EXPRESS-G | EXPRESS-G is a standard graphical notation for information models. It is a useful companion to the EXPRESS language for displaying entity and type definitions, relationships and cardinality. Used by the ISO DIS 12006-3. | http://www.steptools.com/support/stdev\_docs/devtools/devtools-8.html |
| OIL | OILS stands for Ontology Inference Layer, a language that was developed in the context of the European IST Ontoknowledge project. It is built on top of RDF(S), using as much as possible RDF(S) constructs in order to maintain backward compatibility. | http://www.ontoknowledge.org/oil/ |
| OWL | The OWL Web Ontology Language is designed for use by applications that need to process the content of information instead of just presenting information to humans. OWL facilitates greater machine interpretability of Web content than that supported by XML, RDF, and RDF Schema (RDF-S) by providing additional vocabulary along with a formal semantics. | http://www.w3.org/TR/owl-features/ |
| RDF(S) | Resource Description Framework (RDF) defines a language for describing relationships among Web resources in terms of named properties and values. It is particularly intended for representing metadata about Web resources, such as the title, author, copyright and licensing information about a Web document, or the availability schedule for some shared resource. | http://www.w3.org/TR/rdf-schema/ |
| XML | Extensible Markup Language (XML) is a simple, very flexible text format derived from SGML. Originally designed to meet the challenges of large-scale electronic publishing, XML is also playing an increasingly important role in the exchange of a wide variety of data on the Web and elsewhere. XML has been largely used to represent "semantics" in the Web, here including taxonomies, classification systems, etc.. | http://www.w3.org/XML/ |
| Topic Maps | Topic Maps (ISO/IEC **13250)** define a model for the semantic structuring of knowledge networks and are a solution for organising and accessing large and continuously growing information pools. They provide a ‘bridge’ between the domains of knowledge management and information management. They can also be used to generate navigation for a website, and lots of other metadata tasks. A topic map is a collection of topics (a topic is a resource that acts as a proxy for some subject; the topic map system's representation of that subject), associations, and scopes that may exist in one of two forms: (i) a serialized interchange format (e.g. as a topic map document expressed in XTM syntax); or (ii) Some application-internal form, as constrained by the XTM (XML Topic Maps) Processing Requirements. A topic in a topic Map represents a subject inside the computer. | http://www.topicmap.com/  http://www.topicmaps.org |
| KIF | Knowledge Interchange Format (KIF) is a language designed for use in the interchange of knowledge among disparate computer systems. KIF, a particular logic language, has been proposed as a standard to use to describe things within computer systems, e.g. expert systems, databases, intelligent agents, etc.. Moreover, it was specifically designed to make it useful as an "interlingua". This means a language useful as a mediator in the translation of other languages. KIF has declarative semantics; it is logically comprehensive (i.e. it provides for the expression of arbitrary sentences in the first-order predicate calculus); it provides for the representation of knowledge about the representation of knowledge; it provides for the representation of non-monotonic reasoning rules; and it provides for the definition of objects, functions, and relations. When the computer system needs to communicate with another computer system, it maps its internal data structures into KIF. KIF is a programmer-readable language and thereby facilitates the independent development of knowledge-manipulation programs. | http://logic.stanford.edu/kif/kif.html |

## Maintenance of ontologies - Ontology Learning

* (What are they? What do they represent?)

Vocabulary, words/terms, concepts/ideas, meaning.

* 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)

3

# Pattern extraction from non-structured sources of information

Chapter 3 – Pattern Extraction from unstructured information sources

## Data mining / Knowledge Discovery.

### Definition - What is DM/KD?

### Techniques used today?

## Pattern Discovery

### FP-Growth – definition

### Algorithms to discover patterns Weaknesses/Strengths between them

[ECLAT, APRIORI, FP-GROWTH]

### Association Rules (Definition, Rules)

* + Algorithms to discover patterns [ECLAT, APRIORI, FP-GROWTH]
  + Weaknesses/Strengths between them
  + Why FP-Growth?
* Application domain. (Practical cases where association rules are used)
* measures

4

# Concept Model

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 / Ontology learning process

FP-Growth how to build and FP-Tree

Association rule evaluation

- DER / MVC / UML Diagrams

5

# Model Design and Development

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??

6

# Assessment

Chapter 6 – Assessment

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

7

# Conclusion and Future directions

Through the following lines will be presented an overview of the work developed in this thesis. The objectives were defined in Chapter 1 that intended to guide the path of the study. For these objectives it will be described which ones were achieved and which ones were not, describing also the problems and difficulties found during the development and research, and also, how were these difficulties solved. Similarly, this will also bring to attention some possible future research topics, where achievements addressed by this work can constitute a solid basis.

## Work overview

As presented through this dissertation, it was described the importance of concept representation in contrast to word representation, in the Semantic Web area. It was highlighted the advantage of the use of mechanisms (e.g.. Controlled Vocabularies), as these provides means for semantic representation, which allows more than just simple word representation. These mechanisms provide the possibility to make semantic search in contrast to the currently used search technology, providing results more adapted to what users want.

Similarly, frequent pattern discovery in texts may enhance the recognition of semantic relations between words. As a result this recognition can help discover the meaning associated to a word. The Data Mining techniques adopted to achieve this task were FP-Growth to discover frequent patterns and Association Rules to provide more than the just recognition of relations between the words (refer to chapter XX.XX). Based in the AR algorithm, this work demonstrated that it is possible to measure the strength of a relation.

Pattern recognition by itself can be an indicator of relations between words, however this can be enhanced through the use of a domain ontology. In this sense, this work proposed Frequent Itemset Mapping, a process to match frequent items discovered in a document corpus and keywords associated to concepts from a domain ontology related to B&C (refer to chapter 4).

As explained in Chapter 2.3, Ontology Learning is the area related to the automatic or semi-automatic maintenance (meaning without human supervision) of an ontology. Through newly discovered knowledge sources it is possible to learn a domain ontology, in the sense that one can use this new knowledge that could otherwise be unknown, be difficult to discover or be unavailable to improve concept relations inside the ontology. The method proposed in this dissertation, based in the AR algorithm, provides metrics in the form of numeric values to evaluate the strength of semantic relations between concepts. Through these value is possible to know if a concept A is more related to a concept B than to a concept C, therefore learn or maintain the concepts from a domain ontology, as the one used in this work related to B&C.

## 7.2 Research Contributions

The development of this work proposed four expected outcomes in Chapter 1 as follows:

* To develop a method to describe how to extract concepts and recognize relations between them from a data document corpus, and to find new knowledge sources in order to update a domain ontology.

The proposed method relies on applying Data Mining techniques to discover knowledge in documents that could be useful to update a domain ontology. Knowledge, meant the discovery of new concepts, relations or the improvement of the relations between the concepts already in a domain ontology (e.g. the ontology used in this work adopted from the B&C domain). The initial resources were a set of documents from the ICONDA[[1]](#footnote-1) database and a domain ontology adopted from SEKS framework developed under a MSc. Dissertation (Figueiras, 2012) both related to B&C domain. The documents were initially processed in the Rapidminer software tool. Rapidminer proved to be a satisfactory tool, as it also allowed to apply the algorithms FP-Growth to discover frequent patterns and Association Rules to discover the relations. The process created for the matching between the frequent items discovered in the documents and the keywords associated to the concepts from the domain ontology was the Frequent Itemset Matching (refer to chapter XX.XX). This process allowed to search through the ontology in order to verify if the frequent items discovered in the documents were associated to any concept inside the ontology, or if it originated new knowledge.

This work tries to develop a method for Ontology Learning where it is possible to turn a domain ontology more up to date. Even with a small sample, this process provided some good results (refer to chapter XX.XX), as it discovered new concepts, and also provided some interesting relations between the concepts. However, OL relies on automatic methods, this work did not intended to provide a full automatic method to learn an ontology. Alternatively, it was intended to develop a semi-automatic method that relies in human interaction to complete the OL task with the knowledge discovered through all the results.

* To develop a proof of concept, a software platform, based in the previous method in order to reflect the application of the studied techniques.

In order to execute all the steps from the method proposed it was developed a software tool, DOKS (refer to chapter XX.XX). DOKS is a client-server application developed using Java technology to implement all the processes and components in this tool. To interact with the ontology, it was used Jena API. The communication to the database was made by JavaBeans technology. The ontology was developed in OWL. Rapidminer provided an API to access its results, and they were exported through a script represented in Groovy. To hold the results for later access, it was created an XML message. Both DBs for the ontology and for the AR results were saved in a MySQL RDBMS.

* Present results of the semi-automatic OL process. Results composed by patterns discovered in the documents, their relations and the new concepts discovered. They should be presented in an understandable way to the user.

To present the results from DOKS, a FrontEnd was implemented in web technology. Here the set of technologies used were: (i) Html5+CSS3 as a base to support the layout; (ii) The communication with the server was made through HTTP requests based on PHP technology to send the results; (iii) To present the results in the web page, the technology chosen was PHP + XPath. The results were presented in a first page, in which the user could choose two concepts, based on the Frequent Itemset Mapping, and the values of the metrics from each association rule presented by the FrontEnd. This way, a relation between two concepts could be chosen for later processing. It is worth mentioning, the creation of a colour scheme for the Frequent Itemset Mapping process, in order to help the user choose the concept from the domain ontology that best matches the frequent item.

* Finally, publication of scientific documents about the work, to be assessed by the academic community.

The following scientific documents were published after assessment by the academic community during the development of this work:

* Luis Paiva, Ruben Costa, Paulo Figueiras, Celson Lima, “Discovering Semantic Relations from Unstructured Data for Ontology Enrichment - Association rules based approach”, 8ª Conferência Ibérica de Sistemas e Tecnologias de Informação: CISTI'2013, pp 579-584, 2013
* Ruben Costa, Paulo Figueiras, Luis Paiva, Ricardo Jardim-Gonçalves, Celson Lima, “Capturing Knowledge Representations Using Semantic Relationships An Ontology-based Approach”, Sixth International Conference on Advances in Semantic Processing: SEMAPRO 2012, pp 75-81, 2012
* Paulo Figueiras, Ruben Costa, Luis Paiva, Ricardo Jardim-Gonçalves, Celson Lima, “Information Retrieval in Collaborative Engineering Projects-A Vector Space Model Approach”, International Conference on Knowledge Engineering and Ontology Development: KEOD2012, pp 233-238, 2012

## 7.3 Future Directions

As this work relates to some areas from Semantic Web and Ontology Engineering, some possible directions can be identified for further work and improvement. Two paths are proposed, one related to the improvement of the presented method, the second related to its applicability and reuse.

Sometimes, the knowledge that results from the method proposed herein can be huge, and if the process is not fully automated it can be an exhaustive task to analyse these results. This suggests further research related to DOKS ability to deal with the size growth of data used in the Ontology Learning process, can be identified in three areas: (i) speed to process large sets of data as it can be really slow. Research can be taken in methods to, for instance, take advantage of multi-core processor technology in order to use parallelization techniques to improve the speed of the matching process; (ii) way to present results for evaluation by an expert, although this work provided a colour scheme to represent the strength of the matching process (refer to chapter XX.XX). This means to improve the way in which the results are shown, by using more graphics (e.g. graphs to represent relations). This will provide a better efficiency of the method itself and allow for faster reasoning of the results; (iii) method to process large/huge and complex sets of data, also known as Big Data. Big Data is the nouvelle sub domain of Data Mining that studies solutions to the problem of big and complex sets of data.

Searching for patterns in a document, was proved by this work that it is not an easy task, although it is possible. The relation between words in a document can lead to the discovery of a central concept or idea that could represent its context or domain, for instance a document including the words “*bridges*” and “*buildings*”. The central concept from this document could be identified as “*Civil Engineering*”. However, how can one discover the central idea in a document? Is it even possible? Can this discovery be done? How to find the central idea? Syntactic Context[[2]](#footnote-2) or Latent Semantic Analysis[[3]](#footnote-3) are areas that tries to address this questions, and can be a promising future direction.

It is worth mentioning that the intention of this research was not to develop a fully functional model to deal with data mining. However, the author thinks that it could be a good contribution to the following areas:

* Information Systems: Search engines like Google, Bing or Yahoo, just to name a few, could use semantic search capabilities to improve its results instead of just statistical ones. For instance, if one would like to search for a car, the search engine could provide the pages where the word “*car”* appears, and also the pages where the word *“automobile”* appears based on the relation between these two words. This means that one could search for a concept, instead of having to know every word that represents it. Additionally, it could also provide suggestions, for instance, related to their brands, turning the search results more close of what the user really searched for.
* Cybersecurity: This is an area of great interest in the present, based on several world events related to terrorism. The method proposed in this work, could help in this area, for instance, if one could use a search engine to search in the web for a person A that could be known as related to Al-Qaeda. After using the method proposed, one could also discover a person B that appears frequently in some pages related to person A, although not directly related to Al-Qaeda. This could be a proof of the relation between both people, and the discovery of the relation of person B to Al-Qaeda.
* Cybersecurity and human rights: MEMEX is a project from DARPA with an initial goal of using search technology to help fight human trafficking, as they identified this as a serious problem to solve. The secondary goal of this project was identified as to improve the search mechanisms and tools that are used today. Semantic search could help in the sense that it could discover pages with terms related to human traffic, for example *human trade* or *modern slavery*, which could represent the same idea.
* Team Sports: GloballCoach is a software tool idealized by former Liverpool and Chelsea coach, Rafael Benitez, that targets Football Teams and their coaches, providing them data analysis capabilities. Amongst others, this software aims to recognize patterns in game data to show to team players and improve their tactical and technical behaviour. This kind of software systems are focused in the individual behaviours of each player. In this sense, the method proposed in this work could be a great aid in the recognition of relations between the players, augmenting the analysis from an individual up to team perspective analysis capabilities. For instance, the coach could analyse which is the best relation in their left side. Meaning that if he wants to select player A for a game, he could know if the relation between players A and B provides more goals than with A and C. Other example could be, to whom a player A provides more assistances[[4]](#footnote-4), meaning that the relation between player A and B provides more goals to the team than a relation between player A and C.

Summing up, the semantic search is here to stay and is spreading along all research areas. In this sense, controlled vocabularies are useful tools to enhance semantic search capabilities in information systems. Ontologies themselves are great mechanisms to provide search capabilities to users, experts or not, in theirs daily search quests. Consequently, knowing how to provide or obtain the best results will throw companies or entities one technological step ahead of their competitors.

8

# Bibliography

Almeida, T., Souza, R.F. de, 2011. O vocabulário controlado como instrumento de organização e representação da informação na FINEP [WWW Document]. URL http://repositorio.ibict.br/handle/123456789/88 (accessed 2.6.15).

Antunes, J.P.D., 2010. Design and implementation of an autonomous, proactive, and reactive software infrastructure to help improving the management level of projects.

Costa, R., 2014. Semantic Enrichment of Knowledge Sources Supported by Domain Ontologies. Faculty of Science and Technology - New University of Lisbon, Lisbon.

Figueiras, P.A., 2012. A framework for supporting knowledge representation – an ontological based approach.

Gangemi, A., Presutti, V., 2009. Ontology Design Patterns, in: Staab, S., Studer, R. (Eds.), Handbook on Ontologies, International Handbooks on Information Systems. Springer Berlin Heidelberg, pp. 221–243.

Gruber, T.R., 1993. A translation approach to portable ontology specifications. Knowl. Acquis. 5, 199–220. doi:10.1006/knac.1993.1008

Hand, D.J., Mannila, H., Smyth, P., 2001. Principles of Data Mining. MIT Press.

Lei, Y., Uren, V., Motta, E., 2006. SemSearch: A Search Engine for the Semantic Web, in: Staab, S., Svátek, V. (Eds.), Managing Knowledge in a World of Networks, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 238–245.

Lima, C., 2004. Final draft CWA4 proposal “European eConstruction Ontology “version 2004–03–26, in: Workshop on eConstruction N.

Lima, C., Zarli, A., Storer, G., 2007. Controlled Vocabularies in the European Construction Sector: Evolution, Current Developments, and Future Trends, in: BSc, G.L., MSc, BEng, R.C. (Eds.), Complex Systems Concurrent Engineering. Springer London, pp. 565–574.

N.I.S.O. (US), others, 2005. Guidelines for the construction, format, and management of monolingual controlled vocabularies. NISO Press.

Oxford University, 2006. Oxford Dictionary of English. Oxford University Press, London.

Parada, V.M.M., 2010. Desenho e implementação de um sistema computacional para apoiar a gestão de projectos utilizando técnicas de data mining.

Pouchard, L., Ivezic, N., Schlenoff, C., 2000. Ontology engineering for distributed collaboration in manufacturing, in: Proceedings of the AIS2000 Conference. Citeseer.

Stick-iSchool, 2013. Innovation Ontolgy [WWW Document]. URL http://stick.ischool.umd.edu/newsite/innovation\_ontolgy (accessed 3.24.15).

Uschold, M., Gruninger, M., 1996. Ontologies: Principles, methods and applications. Knowl. Eng. Rev. 11, 93–136.

Witten, I.H., Frank, E., Hall, M.A., 2011. Data Mining: Practical Machine Learning Tools and Techniques, 3rd ed. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Yahoo, 2015. Yahoo [WWW Document]. Yahoo. URL https://www.yahoo.com/ (accessed 3.22.15).

Zhou, L., 2007. Ontology learning: state of the art and open issues. Inf. Technol. Manag. 8, 241–252.

1. ICONDA is a large database of technical documents related to B&C domain. [↑](#footnote-ref-1)
2. Syntatic Context relates to the order of the words in a sentence, and states that through language rules, one can infer the context of a sentence. [↑](#footnote-ref-2)
3. Latent Semantic Analysis is the area that analyse the relations between documents, trying to find correspondence between its terms and concepts in order to infer its context. [↑](#footnote-ref-3)
4. Assistance is the word used in a football game and represents the moment when a player passes the ball to a teammate, and this teammate scores a goal. [↑](#footnote-ref-4)