1

# Introduction

Hoje em dia a quantidade de documentos digitais disponível é abismal. Para descobrir o que se pretende nesta quantidade de documentação é necessário que esta apresente uma organização adequada.

Com o crescimento exponencial de informação em formato digital disponível nos dias de hoje, surge a necessidade de a organizar de forma a esteja facilmente acessível. Para que uma pesquisa de informação seja bem sucedida, ou seja, o seu resultado seja exactamente o que se procura, é necessário que os termos pelos quais se efectua a pesquisa, transportem a mesma ideia presente nessa mesma informação. Desta forma, poderá não ser suficiente uma simples pesquisa pelo termo, já que o conjunto de termos, ou vocabulário presente nesse motor de pesquisa, poderá não ser igual ao introduzido para pesquisa. Por isto, torna-se imperativo descobrir mecanismos de representação digital de ideias ou conceitos, em vez de apenas a mera representação de vocábulos, para que a informação seja facilmente descoberta.

With the exponential growth of available information in digital systems nowadays, rises the necessity to discover ways to organize it, in order to be easily accessible. In order for an information search succeed, meaning that the results provided are exactly the ones being searched for, this search has to include the same terms of each document. This suggests that a simple term search for the information could not be enough, as the set of terms, or vocabulary available could be different from the one being used. Therefore, is important to discover mechanisms of representation of ideas or concepts, and not just the representation of vocabs, with the purpose of access the information intended.

A representação de termos em sistemas de informação, como um computador, é um meio para transmitir informação. As palavras são meios de transporte de ideias e conceitos numa linguagem.

For a computer system (CS) a word is just a set of characters. Computer systems do not understand the meaning that each word represents.

can represent bits. Although to represent the idea that each or several terms represent, techniques must be used to be able to represent the ideas behind those terms.

For a computer system represent a word from any language, it must represent its combination of letters, the term, and the meaning associated to that word.

A concept is a general notion or idea, normally formed in humans mind. To transmit a concept present in the mind of someone, humans must be able to use some sort of representation that translates the concepts. In natural languages, like English, this representation is done using a term or a set of terms. When meaning is associated to each term, it translates into words. Words are the units of a language that have the responsibility of providing meaning to terms. When one is using words to speak or to write, is communicating an idea that is in its mind. Sometimes, one word is not enough to express an idea. Consequently, a concept can be represented by one or more words.

Computer systems can only represent sets of bits and bytes that are not understandable for humans. In order to represent a text in a computer

Com o crescimento das tecnologias de informação, a quantidade de informação disponível cresceu exponencialmente. Esta informação é disponibilizada através de textos, imagens ou outro tipo de documentos. A sua forma de representação num sistema informático é através de termos. O conjunto destes termos designa-se por vocabulário. Cada documento, possui o seu próprio vocabulário. Uma palavra é um termo que apresenta uma ideia. No entanto, uma única palavra pode não ser suficiente para apresentar uma ideia ou conceito.

Os vocábulos ou palavras são meios de transporte de ideias e conceitos construídos nas mentes das pessoas. Nas tecnologias de informação actuais estas palavras estão presentes em todo o lado, como em documentos.

Com o crescimento das tecnologias de informação, a quantidade de informação disponível cresceu exponencialmente. Para que esta seja de fácil e rápido acesso, é necessária que esta esteja bem organizada. É necessário o uso de técnicas específicas para aceder rapidamente à informação.

The World Wide Web is a place with a huge amount of information. Nowadays, any person can access to a search engine (like Google) to find information relevant to its intents. For example, if one needs to find information related to a bar, the place to have a drink, using just the term “bar”, that is not enough. The search results could be ambiguous with several meanings like “a place to have a drink”, or a “piece of wood or metal”, or even “a pressure measure unit”. The problem is that computer systems can only represent text in the form of a set of terms. If the searches could be by meanings and ideas, and not by terms the results would be much better. However, to have the best results to a search, how can several ambiguous meanings of words be represented? How could a word full of meaning or an idea present in one’s mind (also known as concept) be represented by a computer system?

Nowadays, computers systems can represent a set of terms or words, a vocabulary. However, vocabularies do not represent ideas or concepts, they just represent words. In order to address the representation of concepts and ideas, one approach can be considered. This approach is the use of mechanisms to represent more than pure terms, to represent concepts. These are designated as Controlled Vocabularies (CV). CVs provide mechanisms to represent concepts consistently in a domain. CVs also provide a standardization for the representation of concepts, providing a structure for them.

The main goal of CVs is to provide a clear and uniform meaning to the concepts. Specifically, CVs try to solve some issues for the representation of text in computer systems. Getting the previous search example above of the term search. The concepts are in this Issues that are not solved just the representation of terms 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.

There are several types of CVs. Amongst others, Ontologies is one of them. 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.