Ahmad Mehrbod

Master em Engenharia Computação

Semantic and Syntactic Matching of Heterogeneous e-Catalogues

Dissertação para obtenção do Grau de Doutor em
Engenharia Industrial

Orientador: António Grilo, Professor Auxiliar com Agregação, FCT- UNL

Júri:

Presidente: Fernando José Pires Santana

Arguentes: João Pedro Mendonça de Assunção da Silva
Carlos Eduardo Dias Coutinho

Vgais: Virgílio António Cruz Machado
Richardo Luís Rosa Jardim Gonçalves
António Carlos Bárbara Grilo
António Aguiar Costa

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Setembro, 2016

Semantic and Syntactic Matching of Heterogeneous e-Catalogues

To my Family…

Acknowledgements

I would like to express my gratitude to all of those that directly or indirectly contributed and supported my work.

Firstly, I would like to express my sincere gratitude to my supervisor Professor Dr. Antonio Grilo for the continuous support of my Ph.D. study and related research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D. study.

Besides my supervisor, I would like to thank the rest of my CAT committee: Prof. Ricardo Gonçalves and Prof. João Pedro Mendonça.

I would also like to thank Professor Virgilio Cruz-Machado for his brilliant comments and suggestions and all my professors in Industrial Engineering department of Universidade Nova de Lisboa.

I sincerely acknowledge the financial support from project VortalSocialApps, co-financed by VORTAL and IAPMEI and the European Funds QREN COMPETE, and also would like to thank Fundação da Ciência e Tecnologia for supporting the research center UNIDEMI through the grant Projeto Estratégico PEst-OE/EME/UI0667/2014.

Last but not the least, I would like to thank my family: my parents and to my brother and sisters for supporting me in all my life. I would also like to thank all of my friends Aneesh Zutshi, Izunildo Cabral, Pedro Cruz, Pedro Tavares, Raphaela Vidal and Tahere Nodehi from UNIDEMI and João Gameiro, Tiago Ferreira, Hugo Felício, Nuno Milagres and Rui Barreira from Vortal who supported me during this research work. At the end I would like express appreciation to my beloved wife Samira for her understanding, encouragement and support that made this burden possible and my lovely children Anahita and Arad who are my motivation to continue.

Abstract

In e-procurement, companies use e-catalogues to exchange product information with business partners. Matching e-catalogues with product requests helps the suppliers to identify the best business opportunities in B2B e-Marketplaces. But various ways to specify products and the large variety of e-catalogue formats used by different business actors makes it difficult.

This Ph.D. thesis aims to discover potential syntactic and semantic relationships among product data in procurement documents and exploit it to find similar e-catalogues. Using a Concept-based Vector Space Model, product data and its semantic interpretation is used to find the correlation of product data. In order to identify important terms in procurement documents, standard e-catalogues and e-tenders are used as a resource to train a Product Named Entity Recognizer to find B2B product mentions in e-catalogues.

The proposed approach makes it possible to use the benefits of all available semantic resources and schemas but not to be dependent on any specific assumption. The solution can serve as a B2B product search system in e-Procurement platforms and e-Marketplaces.

Keywords: Information Retrieval, e-Procurement, e-Catalogue, e-Tender, Semantic Search, Ontology, Vector Space Model, Product Classification System.

Resumo

Em contratação eletrónica, as empresas utilizam catálogos eletrónicos para a troca de informações sobre o produto com os parceiros de negócio. Correspondência de catálogos electrónicos com os produtos procurados ajuda os fornecedores a identificar as melhores oportunidades de negócio em mercados electrónicos B2B. Mas as várias formas de especificar os produtos e a grande variedade de formatos de catálogos eletrónicos utilizados por diferentes atores de negócios faz com que a correspondência seja difícil.

Esta tese de doutoramento tem como objetivo explorar potenciais relações sintáticas e semânticas entre os dados do produto em documentos de contratação e utililá-las para descobrir catálogos semelhantes. De forma a identificar termos importantes em documentos de contratação, catálogos standardizados e licitações eletrónicas são utilizados como um recurso para treinar um "Product Named Entity Recognizer" de forma a descobrir produtos referenciados em catálogos eletrónicos.

A abordagem proposta torna possível usar os benefícios de todos os esquemas e recursos semânticos disponíveis mas não deve ser dependente de nenhum pressuposto específico. A solução pode servir como um sistema de procura de produtos B2B em plataformas de contratação eletrónica e mercados eletrónicos.

Palavras-chave: Recuperação de informação, Contratação eletrónica, Catálogos eletrónicos, Licitações eletrónicas, Procura semântica, Ontologia, Vector Space Model, Sistema de classificação de produtos.

Definitions and Abbreviations

Term	Description
B2B	Business to Business
CPV	Common Procurement Vocabulary
CRF	Conditional Random Field
CSV	Comma Separated Values
cXML	commerce XML
E-Business	Electronic Business
E-Catalogue	Electronic Catalogue
E-Procurement	Electronic Procurement
IT	Information Technologies
NER	Named Entity Recognition
NLP	Natural Language Processing
OWL	Web Ontology Language
PNER	Product Named Entity Recognition
POS	Part of Speech
RDF	Resource Description Framework

TED	Tenders Electronic Daily
UBL	Universal Business Language
UNGM	United Nations Global Marketplace
UNSPSC	United Nations Standard Products and Services Code
VSM	Vector Space Model
xCBL	XML Common Business Library
XML	eXtensible Markup Language

Content

List of Figures

List of Tables

1

Introduction

1.1 Matching Heterogeneous e-Catalogues

With increasing competitive pressures, manufacturers must continually find ways to reduce costs, increase efficiency, and reduce lead time while at the same time seek greater access to markets in cost effective ways. E-procurement platforms help manufacturers reduce costs by having greater access to raw material suppliers while at the same time help them to sell across greater geographies and increase their market competitiveness (Ramkumar & Jenamani, 2012). This comes as no surprise, given one of the key competitive priorities for the 21st century is the maximization of Internet-based technologies such as e-procurement (Pearcy, Parker, & Giunipero, 2008).

E-catalogues play a critical role in e-procurement marketplaces. E-catalogues are procurement documents that explain the products subject of the procurement process. They can be used in both the tendering (pre-award) and the purchasing (post-award) processes. Companies use e-catalogues to exchange product information with business partners. Suppliers use e-catalogues to describe goods or services that they offer for sale. Meanwhile, buyers may use e-catalogues to specify the items that they want to buy (Ghimire, Jardim-Goncalves, & Grilo, 2013)(Ghimire et al., 2013).

Matching a product request from a buyer with products e-catalogues that have been provided by the suppliers, helps companies to reduce the efforts

needed to find partners in e-marketplaces (Kim, Kim, & Lee, 2002)(Lee et al., 2007).

The large variety of e-catalogue formats (Schmitz, Leukel, & Dorloff, 2005) which are used by various companies is one of the major challenges in the matching process. Since each business actor may use a different structure, classification and identification code for describing e-catalogues, it is not easy to match a product with the e-catalogue requested by another partner (Lee et al., 2007). This heterogeneity makes it difficult and time-consuming to integrate and query e-catalogues (Chen et al., 2010a).

While there are too many different standards for e-catalogues and product classifications in use, often companies do not follow standard formats and prefer to have their individual structures (Chen, Li, & Zhang, 2010). Hence, we often encounter a plethora of catalogue formats ranging from unstructured text to well-structured XML documents. This diversity results in the syntactic heterogeneity of e-catalogues.

While syntactic diversity comes from various schemas and formats in use, often the heterogeneity problem has a semantic dimension as well. Semantic heterogeneity of e-catalogues is due to various approaches to express and model the product concepts by different actors. Business partners may express the same concept using different keywords, classifications or taxonomies that cause to get diverse results in searching for the same product (Chen, Li, & Zhang, 2010).

The traditional approach to integrate e-catalogues is to transform different formats, schemas and taxonomies into a uniform catalogue model (Ghimire, Jardim-Goncalves, & Grilo, 2013)(Kim, Kim, & Lee, 2002)(Chen et al., 2010a). Usually, ontologies are used to define such uniform models and hand-coded rules are used to transform the catalogues to relative knowledge bases. In the homogeneous space that will be provided by the uniform model, the products can be defined or converted to well-structured objects that can be matched efficiently.

But because of variety of known or even unknown structures that are used by various companies in an e-marketplace, achieving a uniform model for e-

catalogues is usually not practical. Development of a uniform e-catalogue model requires a precise and detailed understanding of each of the various formats of catalogues (Benatallah et al., 2006). However, there is always a chance to encounter a new concept or schema which may cause difficulties in its interpretation. Furthermore, the transformation of the existing data to a uniform model can be costly, unscalable and tedious.

Since in the area of e-catalogues we often face up to plenty of models and developing a universal model is crucial, in this research work a practical, expandable and realistic mechanism to solve the problem is proposed. To reach this goal, this research work uses Vector Space Model that is common in the world of web search engines and customizes it to solve the matching problem of the e-catalogues.

Vector Space Model is an algebraic model for presenting text documents as vectors which is the base of many search techniques and document similarity methods. Using this presentation, the document can be compared and the search queries can be answered using simple mathematics. Vector Space Model (VSM) has several attractive properties and can be applied to both semantic (Mukerjee, Porter, & Gherman, 2011)(Widdows, 2008) and syntactic (Manning et al., 2008)(Carmel et al., 2002) aspects of the search problem. Although the basic functionality of Vector Space Model refers to keyword search in textual data, several efforts have been done to eliminate many of the problems associated with exact term matching and expand it to semantic matching in a wide range of search applications as well. Semantic search is a data searching technique in which a search query aims to not only find keywords, but to determine the intent and contextual meaning of the words used for search in order to improve search accuracy.

In semantic matching, traditional keyword-based VSM is adapted with vectors that are comprised of semantically defined entities, instead of keywords (Mehrbod, Zutshi, & Grilo, 2014b). Domain-specific semantic search typically involves recognizing entities in the query and search data and matching them up to entities that make sense in the particular domain. In order to reach this goal, the existing entities in the search domain should be extracted using Named Entity Recognition techniques. The product information usually is em-

bedded in text which imposes a barrier on collecting, comparing and analysing the product information. Here the problem is to match free-text, semi-structured or even structured product descriptions to their related entities in a semantic resource. In simple words, the problem is to detect a known product in a piece of data.

Named Entity Recognition (NER) extracts information automatically from a given set of documents, thus requiring lower human effort than other approaches to semantics, such as hand-coded knowledge bases and ontologies (Turney & Pantel, 2010).

The proposed method uses Vector Space Model and Named Entity Recognition to measure the syntactic and semantic similarity ratio of providers' e-catalogues with a buyer's e-catalogue or call for tender. Instead of developing semantic or syntactic models and combining them to universal models that try to cover all possible cases, the idea here is to use any available syntax and semantic information of e-catalogues to interpret product data without any model assumption. The matching process uses the syntactic and semantic metadata for interpreting each e-catalogue as much as the information is available for the system. But is not dependent on this information and uses the basic mechanism of VSM for tolerating unknown formats.

1.2 Research Questions

Matching products queries on buyer's side to the product data on supplier's side can help both parties to achieve business goals in digital marketplaces. Companies use e-procurement tools, processes and techniques to purchase their required goods and services. The companies usually share their product data in the form of e-catalogues in B2B e-marketplaces. While suppliers publish their offered products and services in the form of e-catalogues, buying organizations can also benefit from the usage of the buyer e-catalogues to announce their needs.

The shared data can be a valuable resource for search engines to find the best suitable results for the business actors in the marketplaces. The product da-

ta can lead the matching mechanism in finding and recommending the right suppliers to a buyer. In this sense, the search engine uses the product data from an e-catalogue in order to find similar and related e-catalogues that is referred as e-catalogue matching.

The main barrier on matching product data with the same or similar products mentioned in different e-catalogues is the heterogeneity of e-catalogues. Since various companies utilize different structures, schemas and standards to create the catalogues, it is not straightforward to identify similar products especially in multi-resource marketplaces.

Some procurement marketplaces especially in public sector try to force beneficiary companies to follow a specific standard. Though the standards do not cover all aspects of the problem and sometimes are difficult to integrate with in-house procurement systems, companies barely follow e-catalogue standards.

In order not to impose fixed structures to the companies, the e-catalogue heterogeneity problem is treated using data integration models. The idea behind it is to uniform all e-catalogues which come from different resources into a covering data model that makes the matching process easy. But achieving such model and transforming the catalogues to the model is a critical issue of this method. This leads to the first research question of this Thesis:

- **Research Question 1**

  How can buyers and suppliers match their e-catalogues in an efficient way, with no restrictions regarding data integration models?

  Since the integration process can be difficult, costly and not extendable, the question here is how to exploit the product data without restricting the e-catalogues in an integrated model. Hence, the problem is to develop a flexible method which is able to find out various e-catalogues regardless of the structure and the content model. How can this method figure out the content of the e-catalogues without transforming them into a reference model? How will it exploit the possible available structures, models and standards? And how will it tolerate missing such knowledge?

Moreover, several documents have been used during the procurement process. Among these documents, public tender notices and contract awards are published publicly in order to increase transparency in public procurement. These documents are not only valuable for transparency goals, but also contain worthwhile information for business actors about products and services that are being purchased. Similarly, in the private procurement sector also B2B e-marketplaces announce the tender calls and make them available to the suppliers who are looking for business opportunities.

In the other side, suppliers publish their product e-catalogues to make them available for buyers in B2B e-marketplaces. Regardless of public or private sector of the procurement process, the main search scenario in a procurement marketplace is to search in tender notices for finding business opportunities. This sets the ground for the second research question:

- **Research Question 2**

  How can suppliers improve their efficiency in finding business opportunities in e-procurement platforms using the content of their e-catalogues?

  Current opportunity search and tender notification systems in procurement platforms use only simple keyword-based and column-oriented search mechanisms. The question here is how to find the suitable business opportunities for a supplier based on his product and services. How can a supplier receive the tender calls that are similar to the supplier's products instead of having to check all tenders published in the business sector? How can a supplier have a list of all available opportunities form various procurement platforms and marketplaces ranked based on their similarity to the supplier's products and services?

1.3 Propositions

The basis of this research work is to develop a flexible and extendable method to search and match similar products in procurement documents that

have been published using various syntaxes and semantic in B2B marketplaces. This research work has been developed based on the following propositions:

- **Proposition 1**

  Information retrieval techniques can be used to make a flexible model of the existing concepts and data in e-catalogues to search similar products in B2B e-marketplaces.

  This thesis aims to exploit an information retrieval solution based on Vector Space Model to develop an e-catalogue matching mechanism. The flexibility of this method makes it possible to apply it to heterogeneous and diverse data environments. This helps us to use all available product information and structures in defining the searchable elements and indexing the product data without requiring mandatory schemas.

  Modelling the product data in a vector-based space makes it possible to calculate the similarity of e-catalogues and search queries. The definition of the searchable elements in the model is the key factor that specifies the similarity measure and the matching mechanism.

  A multi-layer matching mechanism will be used to exploit available information and tolerate missing information. The multilayer mechanism starts with measuring the similarity based on the syntactic and structural features of the e-catalogues. In the next layer, this will be extended to discover potential semantic relationships among product data to find semantically similar e-catalogues. If suitable input data for a layer cannot be found, the matching mechanism still can use the other layers or basic functionalities of VSM to make a matching. Therefore, any missing structure or data definition doesn't affect the whole matching process.

  Every semantic search mechanism needs to identify the mentions of the desired items from the search context. The searchable elements that will be indexed by a search engine determine the search functionality. Therefore, extracting the elements from search corpus is a critical task in developing a search mechanism. Consequently, search mechanisms usually exploit

Named Entity Recognisers to extract the mentioned items and the determine the definition of the items from a knowledgebase or an ontology. The semantic extension of the matching mechanism measures the similarity ratio of e-catalogues using semantic relationships of data attributes defined in domain ontologies. The ontologies that are built based on procurement product classification systems are used in an iterative process to extract the semantic relationships among product data and enrich the search indexes with synonyms, similar and semantically related elements.

Therefore, the semantic extension of the e-catalogue matching engine will be supported by a Named Entity Recogniser which identifies the meaningful searchable elements from the procurement documents. The Entity Recogniser has to be trained with several known samples in order to be able to figure out similar occurrences in procurement documents. In order to make required training set, an extensive resource of publicly published tender notices will be used in a stepwise method to make training samples.

Although Product NER is becoming more and more attractive in e-commerce information systems, there is no work in the area of B2B e-commerce. The develop B2B PNER process that can serve as the basis for other B2B information retrieval systems, e-Procurement platforms and e-Marketplaces. This will support the search mechanism to match various expressions of the same products.

- **Proposition 2**

  The existing product data in procurement documents can be exploited to support in a much more efficient way suppliers and buyers to find business partners and opportunities in B2B e-marketplaces.

  The information retrieval techniques are the basis of the search engines and full-text search methods in heterogeneous contexts. The main characteristic of such techniques that makes them successful is their capability to be redefined and customized for various kinds of search problems. By defining the underlying elements, this research work will customize information retriev-

al techniques in order to develop an e-catalogue matching engine to cope with syntactic and semantic heterogeneity in e-catalogues. The goal is to find and match similar products in different procurement documents.

Tenders and e-catalogues that are being published in various resources have different semantics, taxonomies and schemas. This heterogeneity can be managed using information retrieval techniques. The customization of information retrieval techniques makes it possible to take intended factors into account for calculating the similarity ratio between tenders and e-catalogues in the search domain. The similarity of the products will be calculated based on all available information including syntactic and structural features of the e-catalogues and at the same time, potential semantic relationships among product data to find semantically similar e-catalogues.

Using the matching mechanism, suppliers can search for similar tenders to their e-catalogues. The flexibility of the matching mechanism provides the opportunity to search tenders gathered from various procurement notification systems. The matching results will be a ranked list of the tenders based on the calculated similarity ratio to the supplier's e-catalogue. This will help the suppliers to find suitable business opportunities with less effort.

1.4 Research Methodology

The development of this thesis has followed methodologies for the various phases of the thesis.

1.4.1 Collection of Literature Review

The theoretical basis for this research work is based on three complementary Research Areas including State of the art on E-catalogue Matching, Concept-based Information Retrieval techniques, and Information Extraction as shown in Figure 1.1. Thus a literature review of all these three areas was performed to identify the characteristics of the proposed matching mechanism.

After a review on the benefits of matching e-catalogues, possible matching e-catalogue scenarios and the barriers on the matching process in procurement marketplaces, the State of Art in E-catalogue Matching discusses the previous solutions provided on the academic literature for solving the problem. Existing solutions are reviewed and categorized in five different groups. This helps us to be familiar with the available solutions, their pros and cons and available resources that can be reused in solution development.

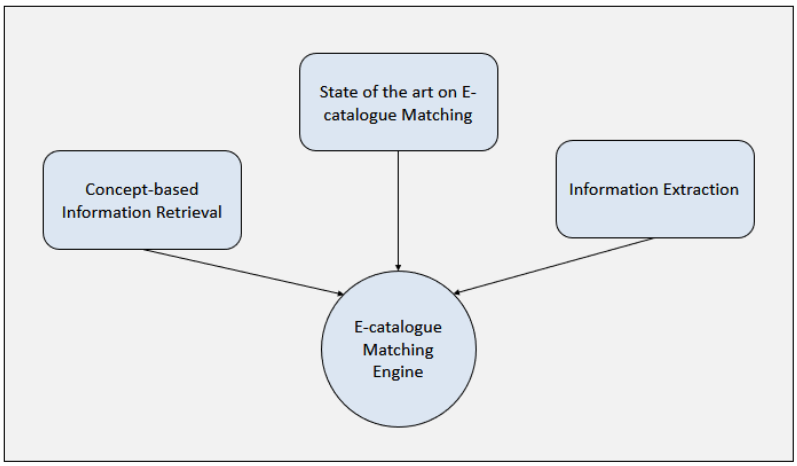

Figure 1.1 Research Areas that lead to the development of the matching mechanism

Information Retrieval and Information Extraction Researches explore the theoretical background needed to develop a flexible and extendable search engine for matching e-catalogues. This provides us with the base to build the matching mechanism. Not all the elements of a search engine are explored within this process, but only those that deal with extracting, modelling and matching the products mentioned in procurement documents.

The study of Information Retrieval and Information Extraction concepts provides us the attributes of the search mechanism that should be defined and the elements that should be extended in order to customize the search mechanism for e-catalogue matching in a procurement marketplace. It also helps us to find a way to reuse the existing resources and knowledgebase in developing the matching mechanism. When the Information Retrieval and Information Extraction techniques are merged with e-catalogue matching concepts, a more fo-

cussed approach to developing a matching mechanism for exploiting procurement documents in product search is explored.

1.4.2 Development of the matching engine

A layer-based approach was used to develop a search engine that is applicable to match a large variety of e-catalogues coming from different procurement platforms. The matching process aimed to make a simplified representation of underlying product data in order to use all available information from documents but at the same time not to be dependent on availability of any specific information.

The implementation of this framework has been based on an information retrieval technique known as Vector Space Model. Information retrieval is to fetch data form a resource in reply to a search query. VSM is an algebraic presentation model for documents which is used by many search methods for indexing the search data. The core of the e-catalogue matching engine is made based on VSM that is extended in different layers for exploiting available product information available in procurement documents. The syntactic extension layer helps the matching mechanism to model the value of the data in various levels of a procurement document. The value of each level is adjusted using suitable coefficients based on the syntactic similarity of the documents. The coefficients are customized using boosting masks for standard e-catalogues. The semantic extension layer provides the ability to detect same, similar and related products expressed using synonym words in different procurement documents.

The semantic layer uses Information Extraction techniques in order to detect the product mentions form the documents. These extracted product elements form the searchable units of the concept-based information retrieval used in the semantic layer of the matching engine. The combination of Information Retrieval and Information Extraction techniques enable the matching mechanism to use existing product information to match similar procurement documents originating from various resources.

1.4.3 Validation of the matching mechanism

The e-catalogue matching engine provides us with a flexible search mechanism that can be used in by business actors in various search scenarios in a procurement e-marketplace. Three test cases have been chosen to test the range of applicability of the Matching Engine in retrieving procurement documents. Furthermore, an extra test case is used to demonstrate the accuracy of the Information Extraction block of the matching mechanism. The selection of the test cases has been done to ensure that we have a variety of matchings based on different search scenarios and various features of the matching layers. Supplier Finder simulates the search for finding a suitable provider for a product by a buying organization. Opportunity finder simulates the search by a supplier in a procurement portal for finding business opportunities. Multi-Resource Matching tests the ability of the e-catalogue matching mechanism in using available semantic resources for matching procurement documents coming from different resources. B2BProduct NER Accuracy analyses the efficiency of the Information Extraction method which is used as a complimentary block of the matching engine. The four selected test cases have the following main objectives. A comparison is also given in Table 1.1.

Table 1.1 Test objectives for different test cases

Test Objectives	Supplier Finder	Oportunity Finder	Multi-Resource Matching	B2BProductNER Accuracy
Semantic Matching	✓	✓	✓	✓
Extracting the concepts		✓	✓	✓
Matching Different Structures	✓		✓	
Matching Related Products	✓	✓	✓	✓
Boosting Masks	✓			
Multi-Resource	✓		✓	✓

I. Supplier Finder

- To demonstrate the e-catalogue matching capabilities for finding a supplier by using a product e-catalogue.

- Using syntactic and semantic term extension to match different e-catalogues coming from various resources.
- Adjusting the effects of the data on the matching mechanism using semantic and syntactic features of the procurement data.
- Optimizing the effects adjustment for standard e-catalogues using boosting masks.

II. Opportunity Finder

- To demonstrate the e-catalogue matching capabilities for finding a business opportunity by using a product e-catalogue.
- Using semantic term and query extensions to match e-catalogues with potential procurement opportunities in various business sectors.
- Exploiting semantic data resources to interpret the data and find semantically related products and requests.
- Combination of the Information Extraction method in order to increase the performance of the matching mechanism.

III. Multi-Resource Matching

- To demonstrate the e-catalogue matching capabilities for finding business opportunities coming from deferent procurement portals by using a product e-catalogue.
- Using semantic term and query extensions to match e-catalogues with potential procurement opportunities with different classifications.
- Exploiting different syntactic and semantic data resources to interpret the procurement documents coming from various resources.
- Tolerating the lack of semantic resources for interpreting the product data using Information Retrieval methods.

IV. B2Bprodcut NER Accuracy

- To demonstrate the accuracy of the Information Extraction method used as a complimentary block of the semantic interpretation.

- Exploiting known data to learn to extract similar data from unknown data.
- Extracting product mentions from procurement documents that can be used as the searchable elements of the matching mechanism.

1.4.4 Data Collection

Data for all the test cases were collected from online procurement portals. The portals include online procurement resources that publish the business oportunities in public procurement publicly. For each test a set of randomly selected data has been downloaded and stored in local test repositories. In the case of supplier finder the test data has been stored in the development platform of Vortal company.

1.5 Structure of the Thesis

The rest of this research work has been structured as follow:

Chapter 2 starts with an introduction about e-procurement process and the position of e-catalogues in this process. It continues with a discussion about the importance of product search in procurement documents and discusses the problems ahead this goal. Finally, possible e-catalogue search and matching scenarios in procurement marketplaces will be discussed.

Chapter 3 describes the state of the art on e-catalogue matching. Through a literature review on various proposed solutions for e-catalogue matching, this chapter summarized previous works on matching e-catalogues and their pros and cons. The solutions are categorized in five different classes by a critical discussion.

Chapter 4 provides a background on two related domains (information retrieval and information extraction) that are used as the basic knowledge in developing the proposed e-catalogue matching mechanism. The chapter starts with the aspects of the information retrieval techniques and the reasons for selecting similarity-based approach for matching e-catalogues. Vector Space

Model as the basis of search mechanism is explained and its extension to concept-based VSM for semantic matching is described. The chapter ends with the explanation of information extraction and its application in finding product mentions form text documents.

Chapter 5 discusses the proposed e-catalogue matching mechanism based on VSM. In the first section of the chapter, the VSM is extended in order to develop a mechanism for matching e-catalogues coming from different resources. The method for exploiting the structure of a document in data modelling is explained. The second section describes the semantic layer of the matching mechanism. The section demonstrates the method of enriching the vector model to use available semantic resources in matching e-catalogues. Finally, the third section of the chapter, discuses an information extraction method which is proposed to extract the product mentions from procurement documents as the semantic search elements for the e-catalogue matching mechanism.

In Chapter 6, the evaluation results have been presented. Four test cases have been developed in order to validate the matching mechanism and information extraction methods. Each test started with an introduction on the test scenario that continues with the definition of the test. Then the data resource which is used in the test is demonstrated. Finally, the test results are reported and discussed.

Finally, Chapter 7 discusses the conclusions of the thesis. It briefly describes the motivation behind this thesis explaining the context of this research. It then discusses the contribution of this thesis and the considerations behind the proposed solution. Finally it highlights a roadmap for future research work based on this thesis.

2

E-catalogue matching

2.1 E-procurement Catalogues

Like any other kind of business, the relationship between suppliers and buyers in the procurement process is affected by information technology. The traditional ways that customers use to buy and suppliers use to interact with buyers is transformed to electronic procurement.

E-procurement (electronic procurement) is the purchase and sale of supplies, work, and services through the Internet or other electronic networks. It is considered to be a strategic tool for improving the competitiveness of organizations. E-Procurement helps to improve and simplify the way procurement operates allowing enterprises to identify opportunities and supply goods and services across markets (Ghimire et al., 2013).

An e-marketplace is a virtual space in an electronic network, an inter-organizational information system that allows buyers and sellers to participate trustworthy in the e-procurement process. These open electronic platforms facilitate activities related to transactions and interactions between multiple companies. An Internet-based electronic commerce platform matches multiple buyers and suppliers and enables transactions along with traditional project-based collaborative functions (Wang & Archer, 2007).

E-Procurement (Ramkumar & Jenamani, 2012) is considered to be a strategic tool for improving the competitiveness of organizations. B2B e-procurement chain, which is shown in Figure 2.1, consists of several necessary steps includ-

ing e-Sourcing, e-Noticing and e-Tendering, e-Awarding and e-Contract (also called e-Reverse Auctioning), e-Ordering, e-Invoicing and e-payment that can be summarized into two main phases: pre-award phase (tendering) and post-award phase (purchasing) (Kajan, Dorloff, & Bedini, 2012). Internet-based information systems and platforms are used in e-procurement to replace one, some or all stages of the traditional procurement process. In other words, e-procurement systems may provide an end-to-end solution that covers all procurement phases or dedicated solutions for some important aspects of the procurement process, such as search and selection (Roman, 2013).

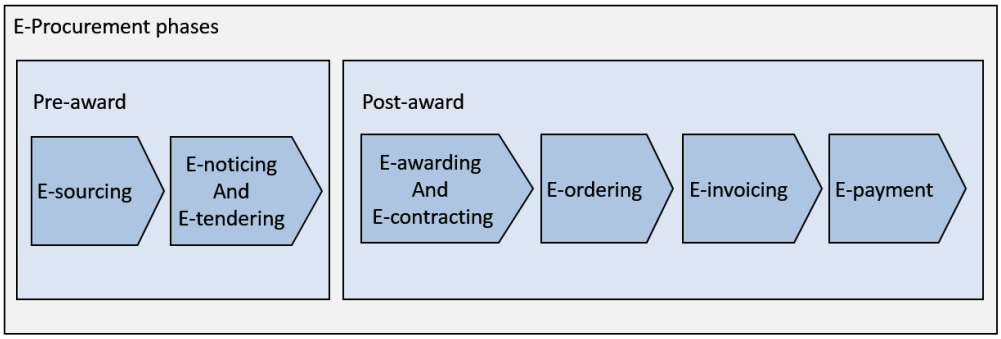

Figure 2.1 e-procurement phases

E-sourcing contains all preparatory activities conducted by the contracting authority (buying organization) to collect and reuse information for the preparation of a call. This process usually contains identifying the suitable suppliers that can be used in the awarding phase. E-sourcing gives the opportunity of marketing to the suppliers and its benefits for the contracting authority include facilitating the sourcing process, reducing prices by maximizing supplier competition and creating a repository for sourcing information (Interagency Procurement & Working Group (IAPWG), 2006) (Pedersen et al., 2012).

E-noticing is the advertisement of calls for tenders through the publication of appropriate contract notices in electronic format in a relevant official journal. An example of such journals is TED (Tenders Electronic Daily) which is the online version of the "Supplement to the Official Journal" of the EU, dedicated to European public procurement. E-noticing includes electronic access to tender

documents and specifications as well as additional related documents that are provided in a non-discriminatory way (Kajan, Dorloff, & Bedini, 2012) (Ordóñez de Pablos, 2012).

E-tendering supports the selection stage and acts as a communication platform between the buying organization and suppliers. This communication provides electronic access to tender documents and specifications for economic operators (suppliers) as well as support for preparation of an offer. Furthermore, it provides the possibility of submission of offers in electronic format to the contracting authority, which is able to receive, accept and process it in compliance with the legal requirements. Therefore, e-tendering covers the complete tendering process, usually including support for the analysis and assessment activities. It does not include closing the deal with a supplier but facilitates a large part of the tactical procurement process. It results in equal treatment of suppliers, transparent selection process, reduction in legal errors, clear audit trail, more efficiency in the tactical procurement process and improved time management of tendering procedures (Interagency Procurement & Working Group (IAPWG), 2006) (Kajan, Dorloff, & Bedini, 2012).

E-awarding is opening and evaluation of the electronic tenders received, and award of the contract to the best offer in terms of the lowest prices or economically most advantageous bid. E-contracting is the conclusion, enactment and monitoring of a contract or agreement through electronic means between the buying organization and the winning tenderer. It enables the closing of a deal with a supplier if parties agree on a price. They operate with an upward or downward price mechanism e.g. e-auctioning with upward price mechanism for the selling organization and e-reverse auctioning with a downward price mechanism for the buying organization (Ordóñez de Pablos, 2012) (Interagency Procurement & Working Group (IAPWG), 2006). In a reverse auction that is used in B2B procurement, the role of the buyer and seller is reversed, with the primary objective to compete for purchase prices downwards. In an ordinary auction, buyers compete to obtain a product or service by bidding a higher price while in a reverse auction, sellers compete to win the business by bidding a lower price.

E-ordering phase contains presentation and issuing of an electronic order by the contracting authority and its acceptance by the contractor. This process consists of creating and approving procurement requisitions, placing purchase orders, as well as receiving goods and services ordered, by using software systems based on the Internet. E-invoicing is the preparation and delivery of an invoice in electronic format and e-payment means electronic payment of the ordered goods, services or works. (Management Association, 2013) (Interagency Procurement & Working Group (IAPWG), 2006).

E-catalogues are electronic representations of information about the products and services of an enterprise (J. Z. Huang et al., 2005), considered as a key enabler in both phases of e-procurement process (Pedersen et al., 2012). In the pre-award process, e-catalogues are used by suppliers to submit offers about goods and services and in the post-award, they are used to exchange information about goods and services offered under the contract.

Companies use e-catalogues to exchange product information with business partners. While suppliers create catalogues to make their product and service content available to their customers, buying organizations create catalogues to specify the items that they want to buy and consolidate product content from diverse suppliers and make it available to their users.

A particular application of e-catalogues in e-procurement that is not studied enough in literature is to use them as input to provide a suitable call for tenders. In this sense, e-catalogues are not only usable in ordering and invoicing process, but also their contents can be reused by contracting authorities to describe goods or services in a call for tender (Icf - Ghk, 2014).

The use of e-catalogues in B2B procurement can significantly benefit both buyers and suppliers due to the automated processing that e-catalogue management tools can offer. E-catalogues can form tenders or parts of them. The use of these tools can simplify the processes followed by suppliers to create offers, while buyers can automate processes for reception, evaluation, purchasing and invoicing (Pedersen et al., 2012).

2.2 *Matching Problem*

An e-Marketplace is an inter-organizational information system that allows a trustworthy e-Procurement process (Zhang & Bhattacharyya, 2008) (Grilo & Jardim-Goncalves, 2013a). One of the main critical success factors in e-Marketplaces is to address technical issues in order to afford the proper coordination in a heterogeneous environment (Alvarez-Rodríguez, Labra-Gayo, & De Pablos, 2014). One of these technical issues is to provide more intelligent search engines that assist in making decisions in more time efficient and accurate way (Kaptein & Parvinen, 2015).

An e-catalogue matching service in an e-marketplace which matches a buyer e-catalogue with product e-catalogues that have been provided by the suppliers, helps suppliers to reduce the efforts needed to find customers in e-marketplaces (Lee et al., 2007). Buyer e-catalogues are catalogues created by the buying organisations. Normally, such catalogues are limited to the goods covered by pre-negotiated prices, specifications and terms (Lysons & Farrington, 2006).

But most of the available B2B e-Marketplaces only provide simple keyword-based and category-based search services to their users for finding contract and tender notices. For suppliers that want to find business opportunities especially in public procurements, keeping track of all potential procurement opportunities from various procurement portals is time consuming and expensive, and typically not a part of their core business. To find suitable opportunities, they have to monitor all of the hundreds of procurement data sources. (Graux, Kronenburg, & August, 2012).

Helping the suppliers to identify the best suitable opportunities with automated processes will not only decrease the time required for locating and responding to opportunities but will also benefit the buying entities in making a decision over the proposals, because matching between opportunities and supplier catalogues will indirectly help in the submission of more closely related proposals.

E-procurement documents such as Contract Notices, e-Tenders and e-Catalogues can play a key role in the search process and be utilized in finding

business opportunities in e-Marketplaces. These documents contain information about products and services. The information about the products in various procurement documents along with another information like the categorization or classification code is of great importance for the suppliers to identify the most relevant opportunities (Ghimire, Jardim-Goncalves, & Grilo, 2013). This product information can be used by a product search mechanism in order to find and recommend similar products and services.

The large variety of e-catalogue formats which are used by various companies is a major challenge in the matching process. Since each business actor may use a different structure, classification and identification code for describing e-catalogues, it is not easy to match a product with the e-catalogue requested by another partner (Lee et al., 2007). This heterogeneity makes it difficult and time-consuming to integrate and query e-catalogues (Chen et al., 2010a) (Grilo & Jardim-Goncalves, 2013a). The matching problem is shown in a schematic view in Figure 2.2.

The problem of e-catalogue integration is more visible in B2B e-marketplaces than the B2C e-commerce websites, since the data of the catalogue-creating enterprise has to be imported into an information system (target system) of the catalogue-receiving enterprise (Leukel, Schmitz, & Dorloff, 2002). However, other e-marketplaces can also be suffered from this problem.

While, there are too many different standards for e-catalogues and product classifications in use, often companies do not follow standard formats and prefer to have their individual structures (Chen, Li, & Zhang, 2010). Hence often a plenty of catalogue formats (Ghimire, Jardim-Goncalves, & Grilo, 2013) ranging from unstructured text to well-structured XML documents exist in e-marketplaces. This diversity results in the syntactic heterogeneity of e-catalogues.

Syntactic diversity is only one side of the heterogeneity problem of matching e-catalogues. The other and yet more complicated side of this problem is the semantic diversity of e-catalogues. There are many ways for a user to express a given concept using different words. The same product concept can be expressed in different keywords, taxonomies or expressions. Figure 2.3 shows

syntactic and semantic heterogeneity problem of e-catalogues (used in Figure 2.2) for presenting the same product.

Hence, different users may use different terms to express the same product, which makes the matching process get different results when facing a synonym query (Lee et al., 2007). For example, two different buyers may use "*road maintenance work*" and "*road-repair works*" terms for expressing similar concepts in call for tenders. Therefore, suppliers usually prefer to use category-based search or subscribe to category-based alert services in B2B e-Marketplaces and receive all the tenders that are being published in their desired product categories. Consequently, suppliers will receive long and unsorted lists of tenders that need to be checked manually in order to find proper opportunities. This process can be very time-consuming especially in large e-Marketplaces.

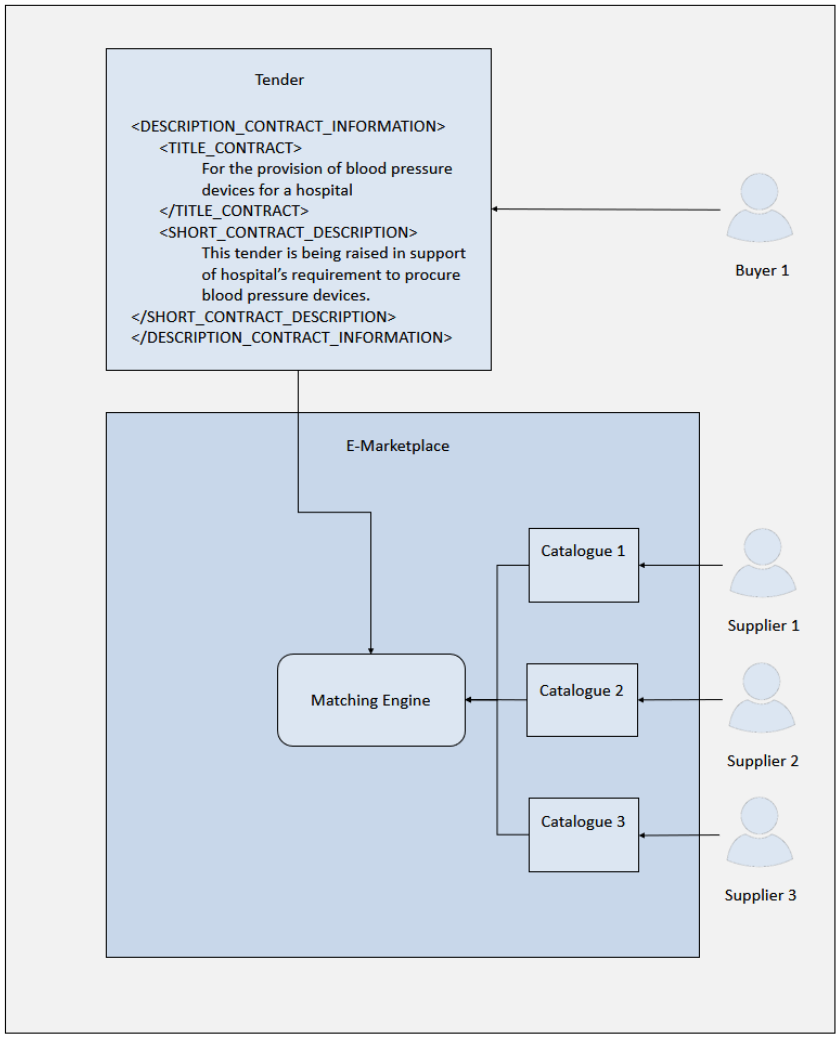

Figure 2.2 Matching Problem

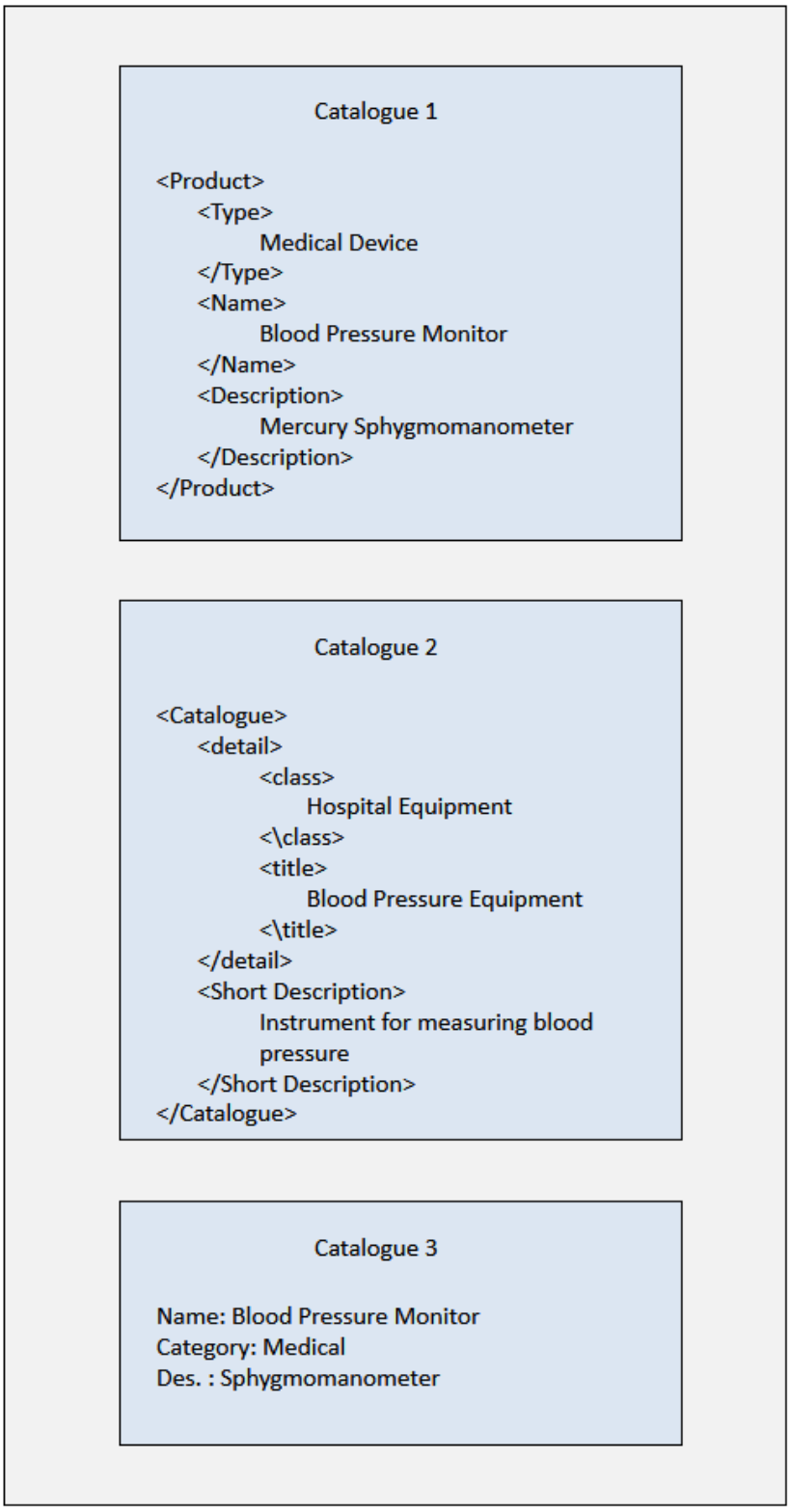

Figure 2.3 Heterogeneity of e-catalogues.

Furthermore, the classification categories can be too general and do not cover the details. The details are usually expressed in the description part of the tenders that can be used to improve the search precision. But as mentioned this valuable data is usually unstructured and heterogeneous.

These heterogeneities make it difficult and time-consuming to integrate and query e-catalogues. As a result, companies cannot use e-catalogues from their own e-procurement systems or other organisational information systems directly in an e-Marketplace or a tendering portal. They have to follow e-catalogue creation rules of each e-marketplace to provide acceptable document format for the e-marketplace.

2.3 Private and public procurement

One of the main factors that makes the procurement process and consequently the procurement marketplaces different from wholesale process and marketplaces such as Alibaba is the direction of the purchasing process. In wholesale websites the producers, suppliers and distributers start the process by offering the goods on the website; and the buyers search to find their needed goods, inquiry the price and features, negotiate the purchasing, payment and delivery process and then order (Guo & An, 2014). But in a procurement marketplace the purchasing process starts by the buying organization. This process is called Reverse Auction (in comparison with forward auction) and is used in B2B procurement in order to obtain the best price by encouraging competition between the suppliers (Jap, 2007). In pre-award phase, the buying organization makes a call for tender based on its requirements, publishes it publically or sends it to a prequalified list of suppliers based on the directives, regulations and its policies, and waits until a deadline for the suppliers to send their proposals. After evaluating the proposals and selecting the winner, the purchasing process will continue by making the contract, payments and delivery in post-award phase.

The procurement process can be categorized into public and private based on the sector of the buying organization. The private sector comprises private buying organizations, and the public sector comprises organizations owned by the national, state or local governments. While the process of acquiring goods and services in private procurement is customized to satisfy the needs of a particular private entity, public procurement is carried out within a specific legal framework to prevent corruption and provide equal opportunity.

In public procurement usually three types of procedures have been used by the contracting authority to procure goods, works and services. These methods are Request for Quotations (RFQ), Invitation to Tender (ITT) and Request for Tender (RFT - sometimes called Request for Proposal (RFP) especially in private procurement). The most common method especially for high-value purchases is RFT which suppliers are requested to deliver their proposed solutions, specifications and prices for solving a problem by a deadline. RFQ is commonly used for lower value purchases and when the product is well defined and the contracting authority wants to know about the prices. While the policy can vary between different public organizations, normally the project is awarded to the lowest price (Interagency Procurement & Working Group (IAPWG), 2006).

RFQ may be sent to only a few suppliers and their price confirmed by the procurement officer against past purchases. Finding suitable suppliers and providing a shortlist of potential suitable suppliers for the next steps of the procurement procedures is generally done by publishing an Expression of Interest (EOI). EOI is to inform tenderers of the context of the project, nature of proposed appointment and submission requirements (Urizar, 2013).

But for higher values a fair competition among the suppliers should be held using formal structured methods such as RFT and ITT. ITT that is also called Invitation to Bid (ITB) is used when the project is defined typically in major construction projects and similar to RFQ the competition is based on providing the lowest price. Request for Tender (RFT) procedure is used when the requirements are not fully definable at the time of solicitation. RFTs are the main sources for finding new business opportunities and provide new and innovative products and services for the suppliers.

Public and private sector institutions engage in procurement for similar goods and services. While there might be some differences in purchasing process of customized goods and services such as a new building or a customized software, the procurement process in the same in both sectors for standard products (Tadelis, 2012). But from the e-catalogue matching point of view in this research work, their potential differences in product search process are Intended.

Private buying organizations operate under institutional policies that are often customized for their business goals. They can source suppliers at will and even award direct contracts without a bidding process. For example, if a private company wants to purchase a good, it can contact various producers to inquire about product quality and pricing and negotiate a potential supply deal. If private organizations choose to invite vendors to submit bid proposals, they naturally focus on awarding contracts to suppliers with favourable terms and conditions. Therefore, even though all the suppliers don't have access to the call for tender and the search for finding opportunities might be disannulled, still the buying organization can benefit from supplier search scenario. The buying organization can use the product specification or a byer e-catalogue to search for similar supplier e-catalogues. The suggested suppliers can be invited to the tender or be inquired directly by the private sector buyer. Furthermore, the buying organization can exploit the similar supplier e-catalogues as the base for making the tender for sending to the selected suppliers.

But public buying organizations have to meet procedures and regulations and procure effectively by a fair bidding process. This provides the access to the open call for tenders for the suppliers and signify the opportunity search scenario as the most common search scenario in a public procurement marketplace. As discussed in the next section, in such search scenario suppliers search in published call for tenders in a marketplace or procurement journal in order to find suitable business opportunities.

Although private and public procurement may have some differences in product matching scenarios, they usually don't have significant technical differences in e-catalogue matching. The e-catalogue matching mechanism can be

applied to all search scenarios that an organization want to match similar products from procurement documents.

2.4 Matching Scenarios

In an e-procurement e-marketplace (Grilo, Jardim-Goncalves, & Ghimire, 2013) at least two main scenarios for finding similar e-catalogues are possible. First, a buyer who makes a call for tender needs to select some suppliers based on their product catalogues in order to invite them to bid for a tender. This process is the ITT procedure which is used in B2B procurement to obtain the best price in defined projects and standard goods (Jap, 2007). In this scenario, suitable suppliers can be selected based on the similarity of their offered products that published in e-catalogues to the products in the call for tender which is being made by the buyer.

The second scenario occurs when a supplier searches to find opportunities in an e-marketplace. In this process, the suppliers search in the tenders that are published publically using the RFP or RFT procedures. Since RFT is the formal method of invitation of suppliers for bidding for high-value projects, especially in the public sector, the published calls provide an important resource for searching business opportunities for companies. As mentioned this is the most common search scenario in a B2B procurement marketplace and a product matching mechanism can help the suppliers to find best suitable opportunities among several tender calls. A supplier may upload a product e-catalogue to the search interface as the search query in order to find the similar call for tenders as potential markets for his products or services. In other words, in the search scenario, a user has an e-catalogue and seeks similar e-catalogues or tender calls in the platform.

In order to prevent corruption and give equal opportunities to all competitors, public tenders have to be published openly in many countries. The tender notices that are published every day in procurement journals and e-marketplaces not only provide better value for money for governments, but also act as a valuable resource for several suppliers for finding business opportunities (Graux, Kronenburg, & August, 2012). Searching and selecting best suita-

ble opportunities among several tender calls especially from various tendering portals is a crucial yet time-consuming task for several business actors in e-procurement marketplaces. The product information provided by a supplier can be used by a product search mechanism in order to find and recommend similar product requests (Julashokri et al., 2011) and tender calls.

Even though these two scenarios are the most common applications of e-catalogues for searching by companies, e-catalogues recommended to be used in making call for tenders or purchase orders. Providing a tender based on e-catalogues not only eases the preparation process for the buyers by providing initial data, but also helps the suppliers to receive correct data.

In this matching scenario, an e-catalogue may be used for finding previous tenders and contract awards in order to reuse them in creating a similar call for tender. The content of an e-catalogue can be employed by the buyers as the input to submit a new call for tender. Therefore, e-catalogues are not only usable in ordering and invoicing process, but also can be reused by contracting authorities to describe goods or services in a call for tender (Icf - Ghk, 2014). This application of e-catalogues is considered by European commission that shows there is a need to extend the use of e-catalogues in pre-awarding phase.

In most European countries (European Dynamics SA, 2007)[1], suppliers use e-catalogues after they have been awarded, mainly for ordering activities and develop them according to buyer requirements (post awarding phase). However, the use of e-catalogues in e-procurement cycle is applicable and suggested in both pre-awarding and post-awarding phases. In post-awarding phase, an e-catalogue is usually considered as a management system for e-ordering and e-invoicing activities while the current focus is on the future and proper use of e-catalogues in pre-awarding phase as well (Icf - Ghk, 2014). In pre-awarding phase, e-catalogues can be exploited in forming a tender or a part of it. In this

[1] Report on *Electronic Catalogues in Electronic Public Procurement (2007) available at http://ec.europa.eu/internal_market/publicprocurement/docs/eprocurement/feasibility/ecat-vol-2_en.pdf*

case, the e-catalogue has the meaning of an electronic prospectus covering effectively and efficiently e-tendering purposes.

The EU legislative framework of public procurement Directives 2004/17/EC and 2004/18/EC, adopted in 2004, introduces for the first time a coherent and comprehensive framework for the use of electronic public procurement in the EU. Amongst its most innovative provisions, it authorises the use of e-catalogues as a tool for the electronic submission of tenders. In line with its Action Plan for e-procurement, adopted in 2004, the European Commission commissioned a study (European Dynamics SA, 2007) to analyse rules and current practices for the use of e-catalogues in both the public and the private sectors, with a view to formulating requirements and recommendations for their further development.

2.5 Summary

A product search service in an e-marketplace can help the suppliers to identify the best suitable opportunities and respond them in a shorter time (Guo & An, 2014). But most of the available B2B e-marketplaces only provide simple keyword-based and category-based search services to their users for finding contract and tender notices.

In e-procurement, companies use e-catalogues to exchange product information with business partners. Hence, an e-catalogue matching mechanism can be a solution to improve search capability of e-marketplaces and help the companies to find more opportunities. E-catalogue matching helps suppliers and buyers to find the suitable business partner in procurement marketplaces. Suppliers can use such services to find similar tenders and sell opportunities according to their products and the buyers can gain lower prices in a shorter time by finding appropriate suppliers.

The large variety of e-catalogue structures, expressions, and vocabularies which are used by various companies make it difficult to match a product request from a buyer (buyer e-catalogue) with products e-catalogues. While there are too many different standards for e-catalogues in use, often companies do

not follow standard formats. Hence, we often encounter a plethora of catalogue formats ranging from unstructured text to well-structured XML documents. This diversity makes it very expensive to solve the problem by achieving a general common structure.

3

Integration Models

Regarding the usage of e-catalogues in e-commerce, interoperability of e-catalogues (Catalogue integration) and personalization of e-catalogues are two main challenges which have been studied in the literature. Although these challenges are related and many researchers studied both together, the former is to match a search query with product e-catalogues and the latter is more focused on customizing e-catalogue selection based on user profile.

The heterogeneity of e-catalogues which come from various sources (Grilo, Ghimire, & Jardim-Goncalves, 2013) causes difficulty in finding same products from different e-catalogues. As mentioned, generally we encounter with two aspects of heterogeneity in e-catalogues which are semantic and syntactic diversity. Syntactic heterogeneity is the result of different document structures and catalogue formats while semantic heterogeneity is the issue of existing different words for expressing the same concept and different meanings of the words in various contents (Lee et al., 2007) (Leukel et al., 2002).

In order to deal with the integration problem of e-catalogues, several approaches and methods have been proposed. These previous works on matching e-catalogues can be classified into five categories as follow:

I. Standardization

II. Uniform Schema

III. Ontological Model

IV. Ontology Merging

V. Ontology Alignment

3.1 Standardization

Despite the widespread use of e-marketplaces with transactional and collaborative functions, there is today a plethora of electronic formats, product descriptions, and classification schemes, seeking to provide guidelines for the exchange of data between companies, and regarding e-procurement especially, the challenge of having common e-catalogues structures among buyers and suppliers (Grilo & Jardim-Goncalves, 2013a).

In order to avoid the product taxonomy diversity, classification systems such as CPV[2], UNSPSC[3] and eCl@ss[4] try to standardize the references that are used for describing goods and services which are the subject of e-procurement. Using a common classification system for products and services enables reliable and efficient exchanges of product data across organizations (Hepp, Leukel, & Schmitz, 2005).

CPV (The Common Procurement Vocabulary) (European Commission, 2007)(Council, 2002) has been developed by the European Union in order to facilitate the processing of call for tenders published in the Official Journal of the European Union. The aim is to use a single classification system to describe the subject matter of the public contracts.

UNSPSC (The United Nations Standard Products and Services Code) is a classification of products and services developed by the United Nations for use in e-procurement. UNSPSC is a horizontal branch spanning classification which

[2] ec.europa.eu/internal_market/publicprocurement/rules/cpv/index_en.htm

[3] www.unspsc.org

[4] www.eclass.de

is used in the US, the UK and Scandinavian countries. But it is also in use in many other countries and probably is the most used classification standard worldwide (Kajan, 2012).

eCl@ss (the cross-industry product data standard) is an ISO/IEC compliant industry standard for classification and clear description of products and services used in procurement, controlling and distribution.

Besides the classification standards, e-catalogue standards such as UBL[5], BMEcat[6] and cXML[7] allow the standardized exchange of product data as well as product classification system. In the other words, they recommend using the classification systems and furthermore propose common document schemas for unifying e-catalogue document structures usually for exchanging purposes.

BMEcat[8] is a standardized exchange format for e-catalogue data in the catalogue management. The BMEcat format allows the standardized exchange of catalogue data as well as product classification systems based on the XML technology. The BMEcat format is in widespread use in German speaking countries. The BMEcat format was initiated by the Federal Association of Materials Management, Purchasing and Logistics (BME), the leading German companies (including Bayer, BMW, German Telekom, SAP, and Siemens) jointly developed by the Fraunhofer Institute and the Duisburg-Essen University.

xCBL[9] (XML Common Business Library) is a XML component library for B2B e-commerce. This standard is created, maintained, and supported for use free of charge by anyone needing document definitions for e-commerce applications. xCBL is a set of XML business documents and their components. The last version of xCBL is xCBL 4.0 which is available as XML Schema and will be the

[5] www.oasis-open.org

[6] www.bmecat.org

[7] www.cxml.org

[8] www.bmecat.org

[9] xcbl.org

standard for future releases of it. xCBL 4.0 contains 44 documents in various namespaces. Each namespace represents a functional area such as order management, pre-order management, financial management, catalogue and etc. The catalogue namespace contains the xCBL documents that are associated with e-catalogue content creation, processing, and inquiries. The only document which has been published in this functional area is "ProductCatalog". This document covers the pricing and product descriptions for e-catalogue content and has a self-describing set of extensions for further characterizing products and services offered.

cXML[10] (commerce eXtensible Markup Language) is a protocol, created by Ariba, intended for communication of business documents between procurement applications, e-commerce hubs and suppliers. cXML is based on XML and provides formal XML schemas for standard business transactions, allowing programs to modify and validate documents without prior knowledge of their form. The current protocol includes documents for setup (company details and transaction profiles), catalogue content, application integration, original, change and delete purchase orders and responses to all of these requests, order confirmation and ship notice documents and new invoice documents.

UBL[11] (Universal Business Language) was developed by OASIS[12] (Organization for the Advancement of Structured Information Standards). Like xCBL, it is a library of standard electronic XML business documents. UBL 2.0 was released in 2006 and is endorsed at international level. In Denmark, UBL is mandated by law for all the invoices of the public sector. PEPPOL that is aiming at expanding market connectivity and interoperability between e-procurement communities uses UBL formats for content of electronic documents.

---

[10] cxml.org

[11] docs.oasis-open.org/ubl/os-UBL-2.1/UBL-2.1.html

[12] www.oasis-open.org

PEPPOL[13] (Pan-European Public Procurement Online) project initiated by the European Commission is an electronic transport infrastructure allowing governments and companies to connect their IT systems and reliably exchange data and business documents. It aims to develop tools and components that can be re-used by existing or new procurement systems at the national, regional or local level, in order to facilitate cross border participation in public procurements and to facilitate the cross border exchange of procurement data, e.g. through common implementations of standards and interfaces (Graux, Kronenburg, & August, 2012).

The e-PRIOR (open source e-procurement services) platform[14] is being developed in parallel with the PEPPOL e-procurement project to enable public administrations to connect to the PEPPOL network. The goal of PEPPOL is to enable public procurement across borders within the EU. This project used UBL 2.0 documents standards as a basis for developing e-catalogue and other procurement documents' schemas.

Although, the word 'e-catalogue' is often used interchangeably for product categorization systems, in the context of B2B procurement an e-catalogue is a document scheme for electronic exchange and transfer of product data between enterprises while classification standards are categorization dictionaries for these products. Therefore, in contrast to the classification systems such as eCl@ss and CPV, which describe how things can be characterized at an abstract level, e-catalogue standards such as BMEcat and UBL are about actual instances of classes which are described by distinct values in accordance to the dictionary. Some researchers such as (Yen & Kong, 2002) use the term catalogue, not only for standard classification systems but also for referring to categorizations dictionaries of e-commerce websites such as Amazon and Yahoo shopping.

The mandatory use of standards could reduce the effort of integrating heterogeneous e-catalogues, but it also increases difficulties of handling changes

[13] www.peppol.eu

[14] ec.europa.eu/dgs/informatics/supplier_portal/index_en.htm

and contexts in various e-catalogues (Guo, 2009). However, catalogue standards and classification systems are not sufficient to meet all the requirements of data exchange (Leukel et al., 2002). Many standards address vertical or even country-specific needs, thus, their relevance to global e-commerce is limited. Standardization processes are seldom transparent and open to new members. In addition, the participation of small and medium-sized companies in these processes is rather small (Schmitz, Leukel, & Dorloff, 2005). In (Lampathaki et al., 2009), authors analysed international standards that try to address data integration issues. They analysed a set of facts such as scope, completeness, openness, modularity, compatibility with other standards, ability to modify the schemas and maturity for some popular schema standards such as UBL, OAGIS, cXML and xCBL.

Consequently, often enterprises do not follow standard formats and prefer to have their individual structures (Chen, Li, & Zhang, 2010). Also, the variety of standards makes it impractical to reach the classification and schema unification goal. These standards differ in addressed markets, capabilities to represent product information, market acceptance, and standardization processes (Schmitz, Leukel, & Dorloff, 2005). This problem is more visible in multi-source e-marketplaces (Ghimire, Jardim-Goncalves, & Grilo, 2013) (Grilo & Jardim-Goncalves, 2013b). There are at least 25 standards relating to e-catalogue and product classification systems, and thousands of enterprise products databases and e-commerce sites(Kim, Choi, & Park, 2007)(Chen, Li, & Zhang, 2010)(Schmitz, Leukel, & Dorloff, 2005). Each standard tries to provide interoperability among various e-procurement systems. However, the large number of existing standards and their lack of effective integration (Chen et al., 2010b) is an obstacle to achieve this goal.

3.2 Uniform schema

One traditional approach to solve the integration problem is to transform different formats into a uniform catalogue model (Ghimire et al., 2013)(Ghimire, Jardim-Goncalves, & Grilo, 2013)(Chen et al., 2010a)(Liu et al., 2001) that serves as reference format. In order to achieve this general model, these approaches

formulate a formal model to represent various catalogues or extend an existing standard as the general model. Then mapping functions have been designed that can handle the transformation of different formats into the uniform model. For example, (Leukel et al., 2002) proposed a model to improve exchanging processes by extending XML e-catalogue standards. The authors argue existing industrial XML catalogue standards are not sufficient to meet all requirements of data exchange. Therefore, they extended the e-catalogue standards to support the coordination and the exchange more widely by proposing a process model and additional business messages.

(Liu et al., 2001) proposed a model as the definition of diverse product attributes collected from heterogeneous data resources and format translator to convert the product data to the proposed model. (Kim, Kim, & Lee, 2002) presented an e-catalogue model whose purpose is to provide a universal product catalogue repository in order to facilitate catalogue sharing and interoperability. In addition, proposed a model for product classification that allows flexible representation of product hierarchies. The model merges different category hierarchies in order to create one big category hierarchy that contains all the information from each of the category hierarchies while still maintaining the hierarchical information of the original hierarchies.

This approach, which is summarised in Figure 3.1, usually provides a service that acts as a central interoperable hub for collecting e-catalogues from various resources (Ghimire et al., 2013)(Grilo & Jardim-goncalves, 2013a)(Kim, Kim, & Lee, 2002).

The catalogues can be of different formats and can come from any marketplaces of platforms independent of each other. The e-catalogue service is responsible for receiving the catalogues and usually exposes a web service which is responsible for receiving catalogues, map to a default catalogue format, acknowledge the successful reception of catalogues and store them. Figure 3.1 shows the sequence of activities that place upon the reception of catalogues in this approach.

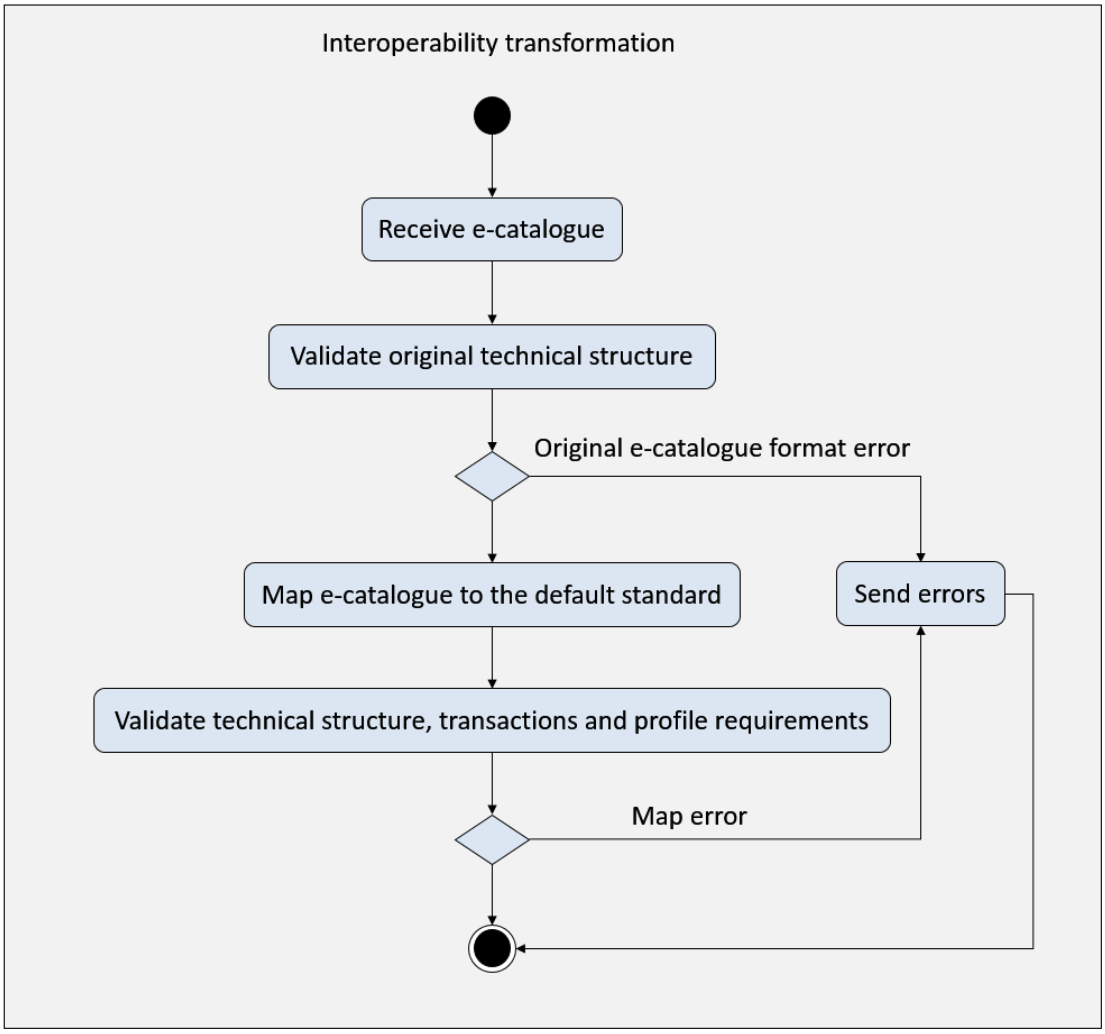

Figure 3.1 e-catalogues transformation to a uniform schema

But within this heterogeneous set of known or even unknown structures achieving a uniform structure for e-catalogues is usually not practical. Development of a uniform e-catalogue model requires a precise and detailed understanding of each of the various formats of catalogues (Benatallah et al., 2006). However, there is always a chance to encounter a new format which may cause difficulties in its interpretation. This problem is more crucial with enterprise specific formats that are used by companies. Furthermore, for transformation to a uniform model, e-catalogues must be completely validated and in conformance to the expected format with no tolerance to format deviations. Since usually each structure is transformed to the general model, it has to be completely compatible with the structure which the converter expects. Implementation of

such converters also can be a crucial and time-consuming task. After conversion of the e-catalogues, there is still a general need for a service to validate catalogue for syntax, completeness, values inserted, etc.

3.3 Ontological Model

Syntactic interoperability alone does not handle all the problems of integration. Even if the structured product information is available, it does not guarantee that the content can be precisely interpreted when e-catalogues use different taxonomies. So, ensuring semantic interoperability is inevitable in the interpretation of product information (Kim, Choi, & Park, 2007). Therefore, several efforts such as (J. Z. Huang et al., 2005), (Chen et al., 2010a) and (T. Lee et al., 2006) have encountered the integration problem by using ontologies to provide a universal semantic model for product data.

The purpose is to introduce generic attributes to design a universal ontology repository in order to facilitate e-catalogue sharing and interoperability (Chen, Li, & Zhang, 2010). The model is then used as a standard reference for e-catalogue transformation or development.

An ontology defines the terms used to describe and represent an area of knowledge. Ontologies are used by people, databases, and applications that need to share domain information. Ontologies include computer-usable definitions of basic concepts in the domain and the relationships among them. They encode knowledge in a domain and also the knowledge that spans domains. In this way, they make that knowledge reusable (Obrst, 2003). In the case of e-catalogue integration, these ontologies are a representation of products and services which include the definitions, properties, and relationships of the concepts that are fundamental to products and services (T. Lee et al., 2006).

Usually, these ontologies are constructed based on semantic concepts of either a product classification system or a product database. Many companies classify products according to generic or industry specific product classification standards, or by using proprietary category systems. Such classification systems often contain thousands of product classes that are updated over time. This im-

plies a large quantity of useful product category information for e-commerce applications. Thus, instead of building up product ontologies from scratch, which is costly, tedious, error-prone, and high-maintenance, it is generally easier to derive them from existing product classifications (Stolz et al., 2014) (T. Lee et al., 2006).

In order to encounter interoperability between catalogue systems of two e-commerce networks, (J. Z. Huang et al., 2005) proposed an ontological model for e-catalogues and converted the e-catalogues to the proposed integrated model. This model is composed of two parts including the structure of product classes and structure of product attributes.

Some projects (Alvarez et al., 2011a) have focused specifically on enriching the publically published tenders and contract awards' data with semantic tags or links to semantic definition references. One of the key issues for using this valuable procurement data sources in semantic matching is the lack of semantic definitions and heterogeneity of the data sources. While public procurement portals such as UNGM[15] (United Nations Global Marketplace) and TED[16] (Tenders Electronic Daily) publish call for tenders as a resource for finding procurement opportunities, they do not provide complimentary information that can be useful in semantic search of tenders.

The general approach of such projects is to integrate the data resources in a semantic schema such as RDF and to link the data to a semantic reference such as an ontology. Nečaský (Nečaský et al., 2014) developed a specialized vocabulary, called Public Contracts Ontology, for semantic definition of public procurement data especially contract awards and tender calls. Necessary extractors and transformers had been implemented for extracting public contract datasets from various formats (HTML, CSV and XML) and convert into RDF format corresponding to the developed vocabulary.

[15] ungm.org

[16] ted.europa.eu

The main target of developing the ontology and the semantic interlinking process was to implement an application for filing public contracts. The application helps contracting authorities to make contracts based on data about themselves and their contracts, suppliers to the contracts, products and services in the contract, and tenders proposed by the bidders. Furthermore, it provides a matchmaking service to help contracting authorities first, in finding similar contract awards to a call for tender that can help them in making contracts and second, in finding suitable suppliers to invite for bidding.

The goal of LOTED project (Valle et al., 2010) (Distinto, D'Aquin, & Motta, 2016) is to improve access to public tenders published on TED portal by republishing it according to a semantic model and providing links to data resources in order to allowing new applications to be built on top of the data. LOTED extracts structured information by checking RSS feeds provided by TED in a daily schedule and reformat it to RDF triples according to an ontology that has been explicitly designed to match TED tender structure. The outcome is an RDF triple dataset which is being updated daily with information extracted as linked data from the RSS feeds of the TED system, and exposed through a SPARQL endpoint. A custom made RDF extractor has been developed which parses the data of the tenders and transforms it into a structured RDF representation.

MOLDEAS (Alvarez et al., 2012) used semantic-web technologies and LOD (linked open data) to provide an integrated e-procurement platform to aggregate, publish and search tender notices. For this purpose, an ontological model has developed and unstructured information from the online version of European public procurement journal (TED) and some other national and regional public tender resources are extracted and transformed into this model. This data later was enriched with linking to product classification system and published via SPARQL endpoints. SPARQL queries can be used for search and retrieve the tenders. Since writing SPARQL queries can be difficult for a business user, a query expander was developed to convert the user query to an SPARQL query and enrich it with the semantic data links.

Product classification systems such as CPV and CPC had been transformed to RDF format that eases their application in semantic search and

matchmaking services. Then the data have been linked to these vocabularies and an RDF view for data has been made that is stored in a triple store and published using SPARQL endpoints (Alvarez et al., 2011b).

(Mynarz, Svátek, & Di Noia, 2015) developed a supplier finder service based on similarity of a call for tender to the history of successful previous contracts. The matchmaking is based on SPARQL queries on linked data of procurement data from TED and Czech public procurement journal. This project, reused the PC ontology to represent the contract's data and CPV vocabularies in RDF and then implemented a reasoning-based matchmaking service using SPARQL to find similar contracts' awards to a call for tender. The reasoning system uses the awarded public contracts as the learning cases or solved problems. The matching results is a ranked list of top-k matched suppliers for a call for tender as a query.

PPROC ontology (Muñoz-soro et al., 2016) is designed in order to contribute to the development of standards that may be used by administrators in publishing call for tenders and contract award notices. The purpose of ontology design is to facilitate access to publically published procurement data not only for contracting powers and tenderer companies, but also for general public. The goal is to give the citizen more information that will increase the transparency of public contracts.

These models have all the benefits of standard schemas and product classification systems and additionally improve the accuracy of the integration process using semantic relationships. But they also have the drawbacks of the standardization approach. It seems impossible to have a globally accepted reference model to create e-catalogues. Even though some publishers started to use ontological models to publish tenders, still we are far from convincing all organizations to follow one data model. For example, two local Spanish governments have started to publish procurement data as instances of PPROC ontology (Muñoz-soro et al., 2016), but it seems impossible to have a globally accepted reference model to create e-catalogues (Mehrbod et al., 2015).

Furthermore, ontology-based search requires the search corpora to be well annotated according to the ontology. A huge amount of information currently available worldwide in the form of unstructured text and transformation cost of

various structured documents into formal ontological knowledge is another issue for practical application of pure ontology-based search approaches. Transforming product specifications to such models is crucial and often relies on manual efforts by domain experts, usually leading to inadequate results.

3.4 Ontology Merging

Ontology merging approaches try to semantically unify e-catalogues by integrating related ontologies. Generally, this approach is similar to uniform model approach but targets the semantic integration instead of syntactic integration. In this approach instead of developing a universal ontological model to create e-catalogues or transform them to a common format, each e-catalogue is interpreted in its own ontological model. Developing various models for each type of e-catalogues is easier and more practical than creating a universal model for covering all kinds of e-catalogues.

Many companies classify products according to generic or industry specific product classification standards, or by using proprietary category systems. Such classification systems often contain thousands of product classes that are updated over time. This implies a large quantity of useful product category information for e-commerce applications. Thus, instead of building up product ontologies from scratch, which is costly, tedious, error-prone, and high-maintenance, it is generally easier to derive them from existing product classifications (Stolz et al., 2014) (T. Lee et al., 2006). Recently, (Stolz et al., 2014) developed a generic, semi-automated method and tool called PCS2OWL for deriving OWL[17] ontologies from product classification standards and proprietary category systems. The resulting product ontologies are compatible with the GoodRelations vocabulary for e-commerce (Hepp, 2008) that is used for annotating offerings and other aspects of e-commerce on the Web. GoodRelations is the on-

[17] www.w3.org/OWL

ly OWL ontology for e-commerce that is officially supported by both Google and Yahoo.

Since different e-catalogue ontologies being generated from different data sources are heterogeneous, the key to semantic integration of e-catalogues in this way is the mapping and integration of catalogue ontologies (Chen et al., 2010a)(Chen, Li, & Zhang, 2010). Ontology mapping is one of the techniques for semantic interoperability that combines two different ontologies into a new ontology that includes and reconciles all the information from the source ontologies according to semantic relations (Kim, Choi, & Park, 2007).

Therefore, different ontologies are combined into a new ontology that includes and reconciles all the information from the source ontologies according to semantic relations. For example (Kim, Choi, & Park, 2007) designed a product information mediation architecture by proposing an ontology mapping algorithm using both taxonomy and the attributes of underlying ontologies. The authors used ontology mapping to integrate different product category names, attribute names, data types and units of attributes into a single unified scheme from the customer's viewpoint. (Chen et al., 2010b) proposed a meta-model and a learning process to acquire the concepts, properties, relations and individuals of underlying ontologies in order to integrate some e-catalogue ontologies into an e-catalogue ontology.

The ontology merging approach can automate the process of solving semantic diversity between heterogeneous e-catalogue, but it could not correctly infer the meaningful information exchanges if there are no semantic mapping rules available (Guo, 2009). Furthermore, all the various matching cases in graphs or models must be predefined using such matching algorithms, which makes them difficult to define. (Kwon et al., 2008) tried to cover all the possible conditions in matching two structures. But within a heterogeneous set of structures, there is always a chance of encountering a new scenario that was not previously considered leading to a failure of the system.

3.5 *Ontology alignment*

Another approach to cope with the semantic heterogeneity problem of the reference ontologies is Ontology alignment. In this approach instead of mapping and integrating different ontologies into one ontology, correspondences between concepts of different ontologies are determined. Assuming that Ontology A and Ontology B are reference ontologies which have been used to annotate the content of Catalogue 1 and Catalogue 2 respectively, correspondence between concepts and relations in the two ontologies simplifies the semantic matching between two catalogues (Beneventano & Montanari, 2008).

In (Benatallah et al., 2006) the authors proposed an alignment method that doesn't require an understanding of each of the underlying e-catalogues. They used the concept of e-catalogue communities to facilitate the querying of a potentially large number of dynamic e-catalogues. A catalogue community is a group of e-catalogues from a domain i.e. catalogues offering products of a common domain such as the community of Laptops providers. It provides an ontological description of desired products (e.g., product categories, product attributes) without referring to any actual provider (e.g., Dell Computers). The ontology that is made from an e-catalogue community is used to interpret related user queries.

In order to achieve interoperability across similar or overlapping domains, instead of merging all ontologies to a global ontology, they used peer relationships among e-catalogue communities to allow sharing of e-catalogues information. A peer relationship is a link between a community and other communities. Once a link is formed, communities can forward queries to each other. The search algorithm first tries to answer each query in its own community. But if the query didn't have any answer, forwards it to the linked communities.

The terms used in community ontology can be different from one community to another. To help solve query mismatch problems, they used synonym-based matching approach. As part of the community ontology, each category (respectively, each attribute) is annotated with a list of synonyms in WordNet. WordNet also is used to assist community providers in defining the mapping between ontologies of two communities.

Making local ontologies for each community using its e-catalogues and proposed alignment process are more applicable and practical than making a global ontology by integrating all e-catalogues. But this approach depends on the efforts of community providers to define peer relationship links that are used in ontology mappings.

In order to solve this problem, (Mehrbod, Zutshi, & Grilo, 2014b) proposed to use a simple, automatic and applicable ontology alignment process based on modelling ontologies in a vector space. The e-catalogue matching process aims to find similar concepts in different e-catalogues by expanding semantic matching process using ontology alignment approach. Adding ontology alignment enabled the matching process to find semantically similar e-catalogues.

3.6 Summary

Based on two aspects of heterogeneity, syntactic integration and semantic integration of multi-source electronic catalogues have been approached to make e-catalogues interoperable. Though there is no one-to-one correspondence between type of heterogeneity and type of integration, some previous research, such as from (Ghimire, Jardim-Goncalves, & Grilo, 2013) and (Ghimire et al., 2013) considered more syntactic integration, others such as (Kim, Choi, & Park, 2007) were more focused on semantic integration and some such as (J. Z. Huang et al., 2005) studied both at the same time.

But regardless of the semantic or syntactic dimension of the problem, both solutions require integration of international product classification standards, enterprise product databases and product e-catalogue standards(Kim, Choi, & Park, 2007)(Chen, Li, & Zhang, 2010). The general solution in e-catalogue integration is to define a global model and transform e-catalogues to this uniform model. Though simple in theory, it is never used widely in industry. Therefore, these traditional solutions either for semantic integration or syntactic integration are dependent on universal formal models. But the variety of structures that are used by different companies makes it almost unachievable to have a uniform structure. Creating such general models has the following problems:

- Requires proper knowledge of the underlying catalogues' structures. But the structure of individual formats which are used by some companies are unknown for the platform and always there is a chance to encounter new formats.
- E-catalogues must be completely validated for conformance to their formats with no tolerance for format deviations. Since usually each structure is transformed to the general model, it has to be completely compatible with the structure which the convertor expects. Furthermore, development of such convertors is a crucial and time-consuming task.
- All the various matching cases in graphs or models must be predefined in matching algorithms. For example (Kim, Choi, & Park, 2007) tried to cover all the possible conditions in matching two structures. But within a heterogeneous set of structures always there is a chance to encounter a new unconsidered condition.
- Transforming product specifications to such models not only is crucial and relies on manual efforts of domain experts, usually led to inadequate results.
- The need to transformers or convertors reduces the scalability of the solution. The huge amount of historical data that has been published in procurement systems needs to be remodelled in each case.

Considering the above mentioned problems that makes it very difficult to define a reference model and keep it up-to-date and expenses of developing customized converters, the integration solutions are not used by e-procurement marketplace providers. For example, Vortal marketplace has 120000 suppliers that makes it impossible to study all e-catalogues to define the reference model and developing this huge number of converters.

4

Information Retrieval and Extraction

Interlinked open procurement data and ontological model solutions reformat data extracted from procurement resources and link it with data from other resources in order to make data suitable for using in new applications such as semantic matchmaking. For example, LOTED project republishes procurement data extracted from TED portal into a harmonized format which is suitable for linking with other datasets such as geographical data of the contracting authority (Graux, Kronenburg, & August, 2012).

As discussed in the previous chapter, the major barrier to integration approaches is to integrate and restructure the data based on the defined data models. The procurement documents coming from various and usually heterogeneous resources have to be transferred to the assumed data model by the solution in order to be accessible for the provided services. Whereas, many of open procurement datasets, as well as e-catalogues, have different formats and semantics that makes it difficult to explore, analyse and use them in an integrated manner (Nečaský et al., 2014) (Valle et al., 2010).

Although such projects are successful in enriching procurement resources with semantic references to the data which is very helpful in reusing procurement data and search purposes, but they are hardly extendable to other procurement data resources. Integrating the data and reformatting it according to the relevant data models requires cleansing and transforming steps. These steps usually contain manual efforts or need resource specific transformers and con-

verters that affect the extensibility of the solution. According to the wide number of procurement data resources, such solutions can be expensive for companies that want to keep track of all potential procurement opportunities.

Furthermore, in such approaches, the information retrieval (IR) problem is reduced to a data retrieval task usually using SPARQL queries. However, the Semantic search should be combined with conventional keyword-based retrieval to achieve tolerance to knowledge base incompleteness (Castells, Fernandez, & Vallet, 2007).

Since developing a universal model to unify various e-catalogues is not practical, this research work uses a flexible model to solve e-catalogue matching problem. Similarity-based matching mostly using Vector Space Model which is the base of many search techniques and document similarity methods can be applied to both semantic (Mukerjee, Porter, & Gherman, 2011) and syntactic (Manning et al., 2008) aspects of e-catalogue matching problem.

The similarity-based matching that is common in web search engines, over few last years has become very popular in B2C and C2C e-Commerce product search as well (Vandic & Milea, 2014). In this sense, the search algorithms are customized for finding products from various data resources.

Several semantic search approaches combined Vector Space Model as the traditional keyword-based search with semantic search methods to provide concept-based search (Wei, Barnaghi, & Bargiela, 2008). In such approaches, usually SPARQL queries are used as a part of the search process, but not an alternative to all the process. Although ontology-based search approaches show better performance than semantic VSM in a well-defined and structured environment that all the search corpora is well annotated, but in average the semantic VSM shows better performance (Castells, Fernandez, & Vallet, 2007).

Although the base concept of VSM has been published in 1975, because of its highly domain-dependent features, it is still being applied to many new domains of search problems. In the context of current research work, Vector Space Model will be used to measure the syntactic and semantic similarity ratio of providers' e-catalogues with a buyer's e-catalogue.

Instead of developing universal models that try to cover all underlying e-catalogues, the idea is to define rich searchable elements from underlying product data. This approach makes it possible to use available models for indexing and searching the data but not to be dependent on any predefined data model.

In this chapter, the basic concepts of the VSM will be introduced. These concepts will be extended and applied to the catalogue matching problem in the next chapter. The goal of the application will be to customize Vector Space Model for matching e-catalogues.

4.1 Similarity-based Matching

Compared to the solutions that are studied in the previous chapter which have transformation overheads and require knowledge about the target systems, some researchers such as (Lee et al., 2007) , (Kwon et al., 2008), (Kim, Choi, & Park, 2007) and (Vandic, van Dam, & Frasincar, 2012) have developed approaches which are not dependent on parsing and converting different e-catalogue schemas.

(Lee et al., 2007) provided an index structure, called Massive Catalogue Index in order to semi-automate the matching process of heterogeneous e-catalogues. The proposed approach eliminates the overhead of transformation and classification of e-catalogues in matching mechanism by providing index tables of common attributes. The main restriction to use the proposed approach is that all attributes of the system should be selected beforehand and this decreases the flexibility of the search mechanism.

A meta-search engine (Kwon et al., 2008) was developed that matches different product categories from various e-commerce websites. The aim of the search engine is to find the most similar supplier product category to buyer's desired product category among underlying categories and recommend the relevant supplier's e-commerce website to the buyer. Therefore, the main task is to match the categories, not matching the products or e-catalogues. In order to use the search engine, the user has to describe his intent using a subclass-superclass

relationship. Then a semantic extension of the query is made by expanding the class names using their synonyms from WordNet. The matching process is to calculate the relevancy ratio between the extended query and the suppliers' product categories. Results will be a list of suppliers that their product categories have higher similarity with the category hierarchy that the user is looking for.

In a more comprehensive approach (Kim, Choi, & Park, 2007), the product search process is divided into two main parts including context search for finding the category of the desired product; and attribute related information search for finding the desired product. In order to cope with semantic heterogeneity, the proposed category and attribute mapping methods regenerate a semantic query conforming to a specific shopping site's ontology with the original query written in a customer's ontology. But schema of the category and description of the attributes are pre-defined and fixed in the proposed model for product catalogues. The search engine makes the semantic translation of the user search query in SPARQL semantic query language and assumes that all underlying websites understand SPARQL standard queries in order to avoid syntactic heterogeneity.

In a similar approach (Vandic, van Dam, & Frasincar, 2012), a general product search engine called XploreProducts aggregates product information from different sources using standardized semantic web technologies and vocabularies. Semantic matching in the proposed platform is based on two main steps including product identification and category mapping from different web shops. Product identification is to identify the product names that represent the same product. The product name identification process consists of four procedures that work based on similarity of characters of the input strings. For example, the starting procedure of identification process accepts two product names as inputs and compare them using calculation of the cosine similarity between two product names. Even though this algorithm works well for detecting similar product names such as phone and telephone or names that are misspelled, but it is not able to find similarity between synonyms such as smartphone and mobile.

Since web-shops use different hierarchies and different names for describing product categories on the Web, the goal of the category mapping is to map a category from the source taxonomy to a category from the target taxonomy. To cope with this problem, the proposed solution uses an existing internal product category hierarchy and maps the new product categories to this internal taxonomy. The category mapping algorithm in (Vandic, van Dam, & Frasincar, 2012), computes a similarity score between the input category and each category in the target taxonomy and then chooses the target category that has the highest similarity. The algorithm assumes that the category taxonomy of the input category is available and maps this available taxonomy to its inner taxonomy. This approach works well for standard e-catalogues where the product classification is known, but usually, we encounter with unknown structures as well.

In an application of VSM in product search, (Vandic, van Dam, & Frasincar, 2012) have focused on adopting web search for matching products. Such works are usually have focused on product search in online shopping websites. Many product search engines that search on various shopping websites or marketplaces encounter similar problems in order to provide comparative shopping search services (Julashokri et al., 2011) (Ali et al., 2010). These services require collecting a given product information from various web pages of different websites in a template-independent manner (B. Wu et al., 2009). But this research work puts forward the application of VSM for matching B2B e-catalogues. An e-catalogue matching engine refers to a product search engine in the context of B2B e-commerce that matches a user search query with product e-catalogues.

In another approach, a product ontology is developed (L. Zhang, 2009) for annotating HTML documents and an ontology-based adaptation of the Vector Space Model is proposed for e-commerce product information retrieval. This approach tries to unify the structure of products in different websites that is the main shortcoming of this kind of systems and needs to develop data convertors or wrappers for each input. This approach works only for extraction and annotation scheme for specific known websites and any new website needs a new wrapper. In B2B e-marketplaces, it is almost impractical to make these wrappers for each company in a marketplace. Furthermore, the provided solution

works based on one ontology and cannot use multiple ontologies. As a consequence, there is no product name identification and no solution for the alignment of different categories of products.

4.2 Vector Space Model

Vector Space Model uses the occurrences of keywords or terms in documents to produce a table of vectors. Each document is represented as a vector of its constitutive terms. The result of vector construction process from n documents that consist of m terms is an n × m matrix (Figure 4.1) where documents are its rows and the terms from its columns. If a term exists in a document, its value or weight in the vector is non-zero and otherwise is zero. Depending on the application, different algorithms have been proposed to calculate the weights of the terms (Manning et al., 2008).

$$\begin{array}{c} \\ \mathbf{Document_1} \\ \mathbf{Document_2} \\ \vdots \\ \mathbf{Document_n} \end{array} \begin{array}{c} \begin{array}{cccc} \mathbf{T_1} & \mathbf{T_2} & \cdots & \mathbf{T_m} \end{array} \\ \begin{bmatrix} \mathbf{V_{11}} & \mathbf{V_{12}} & \cdots & \mathbf{V_{1m}} \\ \mathbf{V_{21}} & \mathbf{V_{22}} & \cdots & \mathbf{V_{2m}} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{V_{n1}} & \mathbf{V_{n2}} & \cdots & \mathbf{V_{nm}} \end{bmatrix} \end{array}$$

Figure 4.1: Matrix of term-vectors

Having a vector model of the documents, mathematical vector operations can be applied to determine the similarity of a document with another one or with a search query.

Documents that are similar to a given query can be calculated by comparing deviation of the angle between the vector of each document and that of the query (Figure 4.2). Closer weights indicate lower deviation angles and consequently more similar documents.

The simplest example is to use the deviation angle between vectors of frequent terms to calculate the relevance between text documents. The lower angle

between vectors of two documents shows that there are more equvalent weights in their vectors. Consequently, these documents have more common terms and are related together (Mehrbod, Zutshi, & Grilo, 2014a).

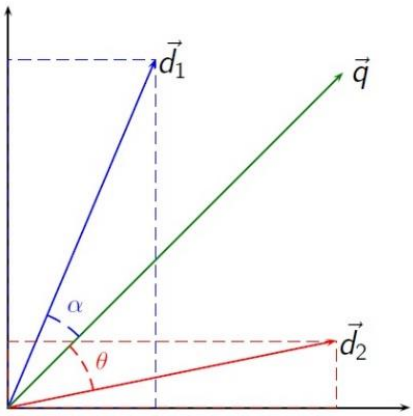

Figure 4.2 Deviation between angles in vector space

In practice, it is easier to calculate the cosine of the angle between the vectors, instead of the angle itself:

$$Cos\theta = \frac{d.q}{\|d\|\|q\|} \tag{4.1}$$

Where $d.q$ is the intersection of the document d_2 and the query q (i.e. the dot product of vector d_2 and vector q), $\|d\|$ is the norm of vector d, and $\|q\|$ is the norm of vector q. The norm of a vector is calculated as such:

$$\|q\| = \sqrt{\sum_{i=1}^{n} q_i^2} \tag{4.2}$$

As all vectors under consideration by this model are element-wise nonnegative, a cosine value of zero means that the query and document vector are orthogonal and have no match (i.e. the query term does not exist in the document being considered).

Depending on the application, several methods have been proposed to define the weights. Keywords are commonly weighted in order to reflect their relative importance in the query or document at hand. The underlying idea is that

terms that are of more importance in describing a given query or document are assigned a higher weight (Mukerjee, Porter, & Gherman, 2011)(Manning et al., 2008). For example, one well-known method of weighting the terms is TF-IDF that takes into consideration both document and collection statistics(Turney & Pantel, 2010).

TF-IDF, short for "term frequency–inverse document frequency", is a numerical statistic that is intended to reflect how important a word is to a document in a collection. The TF-IDF value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the collection, which helps to control for the fact that some words are generally more common than others.

Usually, natural language processing techniques are utilized to extract important terms automatically from the documents and queries. Among various types of processing that can be applied to text, usually Natural Language Processing analysers tokenize, lemmatize and remove stop words in the term extraction process. Tokenization is to decide what constitutes a term and how to extract terms from text.

Stop words are the most frequent and almost useless words. For example, some of the most common words such as *the*, *is*, *at* and so on that don't have a high value in the semantic intent of a text. Lemmatisation or stemming is to convert the different inflected forms of a word to the lemma form or stem so they can be analysed as a single item (Grilo, Ghimire, & Jardim-Goncalves, 2013). For example, all the words such as go, went, goes, gone and going have the same stem.

4.3 Concept-based VSM

The same concept can be expressed in different forms of language expression. As a result, the search engines get different results when facing a synonym query. The current state of product search engines cannot properly deal with semantic heterogeneity that can even affect the predicted growth of e-Commerce (Vandic & Milea, 2014). Therefore, researchers aim to use Semantic

Web technologies in information retrieval for efficient product discovery and presentation (Aanen, Vandic, & Frasincar, 2015). While Vector Space Model is used to deal with flat textual data, it is being extended since the last two decades to treat complex structured and semi-structured data (Tekli, Chbeir, & Yetongnon, 2009).

Keyword-based search engines extract common terms or keywords from search data and produce a table of vectors. Each row or vector of this table represents a searchable element that can be for example a web page in web search engines or a product data specification in product search engines. The advantage of this presentation is that simple vector operations can be used to determine the similarity ratio between the search query and search elements. In data searching phase, the same method is used to make a vector from the search query. The search query vector will be compared to the data vectors that are made in the indexing phase in order to find the similar items to the search query from the data repository.

While the keyword-based search mechanisms are used widely by search engines, they suffer from lacks of semantic interpretation of the search domain. In order to solve the drawbacks of the keyword-based search mechanisms, the semantic extensions to Vector Models represent the documents in the form of underlying concepts instead of the keywords (Turney & Pantel, 2010) that is called concept-based VSM (Widdows & Ferraro, 2008).

The similarity between terms can be found according to the semantic relationship between their corresponding concepts in an ontology and the similarity between documents can be calculated as the similarity of their concepts. An ontology is a description of the concepts and relationships that exist in a domain. The purpose of the description is to enable knowledge sharing and reuse (Elahi & Rostami, 2012).

After extracting of necessary information from underlying documents, the next step is to prepare them for matching mechanism. For example, in uniform model approaches such as (Ghimire, Jardim-Goncalves, & Grilo, 2013) this step consists of transforming the information into the uniform catalogue model. But in semantic search engines, this step is how to make vectors of keywords from documents. The main concern here is to expand traditional vectors of keyword-

based vector space model to vectors that consist of entity names instead of keywords (Mehrbod et al., 2015). Figure 4.3 shows the idea of developing concept vectors instead of term vectors.

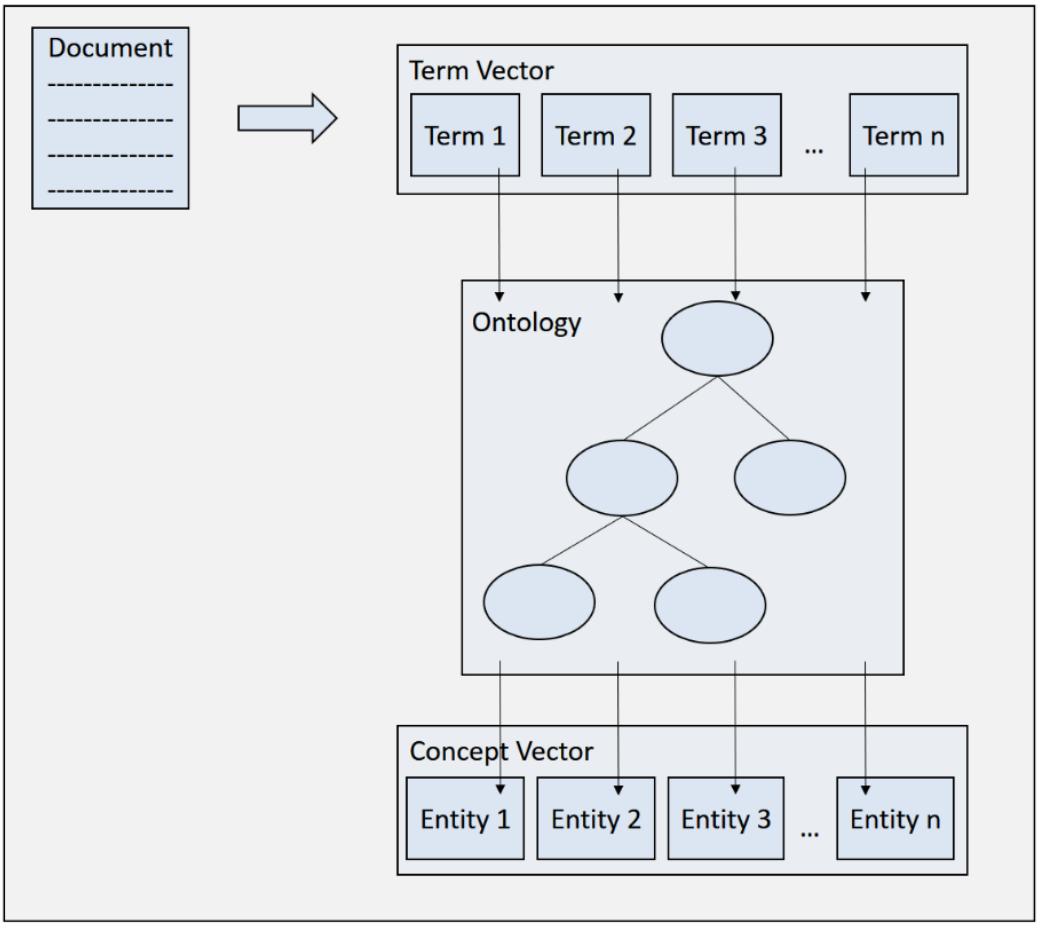

Figure 4.3 concept-based VSM

Vectorization represents the documents based on recognized name entities. This step defines the matching elements or keywords for the search mechanism. This definition determines the calculation method of semantic match degree between a query and the search results.

4.4 Information Extraction

Concept-based vectorization expands the traditional keyword-based vector space model to vectors that consist of entity names instead of keywords.

Vectorization process defines the matching elements or keywords for the search mechanism and represents the documents based on existing entities.

Finding the occurrences of entities such as movie names in a movie search system or product mentions in the e-catalogue matching engine that will be used as the searchable elements in concept-based VSM from the text is another challenge in developing e-catalogue matching mechanism.

The first challenge in searching heterogeneous resources is to identify the mentions that refer to the entities. But finding the occurrences of entities such as product mentions that will be used as the searchable elements from the text is a challenge. The product information usually is embedded in the text that imposes a barrier on collecting, comparing and analysing the product information. Furthermore, a product may have multiple different names while different products may have the same name. The problem is related to a research problem called Named Entity Recognition (NER) (Marrero et al., 2012), in which the goal is to find short and meaningful sequences of terms from data (e.g., a product name) (S. Wu, Fang, & Tang, 2012) and map them to the relevant entities in an entity collection, knowledge base or ontology (Lipczak, Koushkestani, & Milios, 2014).

NER serves as the basis for many other areas in Information Management such as semantic search, faceted browsing, recommender systems, and text categorization (Vandic & Milea, 2014). Since Named Entities can provide much richer semantic content than most vocabulary words, one of the fundamental building blocks of every semantic search engine should be a NER procedure that recognises the named entities from both queries and documents in the related context (Ahn et al., 2010).

NER task is divided into three not totally separated steps including (Eckhardt et al., 2014) mention identification, collecting entity candidates for each mention, and candidates' disambiguation. Actually, two first steps that are also called spotting (Lipczak, Koushkestani, & Milios, 2014) comprise the NER. Spotting scans the input text and looks for interesting sequences of terms and produces a set of possible mentions. Additionally, a list of candidate entities will be retrieved for every single mention. The candidate list will contain all possible senses that can be associated to a specific mention. The disambiguation

step that can be considered as a further step to NER is to link the entities to a knowledge source.

The mention identification selects relevant parts of the text. The relevant parts of the text are defined by the purpose of NER application. Mentions could be based on chosen token classes from Part-of-speech tagging (POS tagging) analysis, named entities or entity names from a knowledgebase, such as a product ontology. Candidates for each mention are collected by looking up entities in a knowledgebase or can be retrieved from a database created by clustering the same mention that occurs in different contexts. The candidates are then ranked or pruned.

Entity disambiguation can be done based on calculating the similarity between document and candidates. Usually, a mapping or similarity score is given to each annotation to show the disambiguation strength. This score can be calculated locally, modelling the mention-entity linkage compatibility; or globally, modelling the coherence among all the entities chosen to disambiguate all the mentions or a combination of both (Piccinno, Ferragina, & Informatica, 2014). The experiments demonstrated that the accuracy of extracted entities relies more on the successful recognition of correct entity mentions (Finding the mentions) rather than their disambiguation (Linking to the resource) (Lipczak, Koushkestani, & Milios, 2014).

Based on the application, various methods can be used to extract the named entities from data. While usually a combination of different techniques is used in entity extraction, the most effective approaches to NER are categorized as rule-based, dictionary-based and machine learning approaches (Melli & Romming, 2012) (Piccinno, Ferragina, & Informatica, 2014).

The rule-based or regex-based NER uses grammar and grammar-based techniques to matches incoming text against one or more predefined regular expressions. This approach typically uses hand-crafted linguistic rules (or patterns/templates) and seeks named entities by pattern-matching or template-filling. The entity candidates are constrained by a set of rule templates where each specific rule can be regarded as a relevant factor to identify the entities. For example, the Nokia Corporation has a series of cell phones named as 'N#' where '#' represents a number, for example 'N97' (S. Wu, Fang, & Tang, 2012).

Regex-based approaches need rich and usually expressive rules in order to achieve good results.

Dictionary-based NER, also known as gazetteer-based NER, matches text against a dictionary of (term, class) pairs to find occurrences of the dictionary's terms in the text. If the text contains the corresponding terms of a class, the mention is easily selected as the class type. This approach assumes the presence of a dictionary of names of target types and identifies the names in the text.

Machine learning approaches are usually sequence classifiers that try to classify a sequence of terms in the text to a class of entities. They use a training set of (term, class) pairs to train a model, and then use the model to predict the category of new (potentially previously unseen) terms. Learning approaches fare better across domains, and can perform predictive analysis on entities that are unknown in a text. Conditional Random Field (CRF) is the most used learning technique in PNER (Product NER) and NER (Bengfort, 2012)(Chen et al., 2014). Other mostly used techniques are Maximum Entropy Markov Model and Support Vector Machine (Putthividhya, Hu, & Ave, 2011).

CRF uses a classifier to predict a label for a sample that can be used in pattern recognition. The task of identifying product entities in the text can be represented as a sequence labelling task, in which each text token is labelled with a tag indicating whether the token begins, continues, or is outside of product entities. Therefore, in NER using CRF, the goal is to label a sentence (a sequence of words or tokens) with tags like *start of a product*, *inside a product* and *end of a product*.

4.5 Product Named Entity Recognition

NER is a subtask of information extraction that seeks to locate and classify words and phrases in text into predefined categories (S. Wu, Fang, & Tang, 2012)(Fang Luo et al., 2011). The most common task in NER is to extract names of persons, organizations and locations. But with the rapid development of e-Commerce, the need for Product NER (PNER) is increasing fast and several researchers have been focused on this task (Lipczak, Koushkestani, & Milios,

2014) (Procházka & Smrž, 2014) (S. Wu, Fang, & Tang, 2012) (Melli & Romming, 2012) (Piccinno, Ferragina, & Informatica, 2014). NER serves as the basis for many other areas in Information Management and this research work aims to develop a B2B PNER process that can serve as the basis for other B2B information systems.

A CRF model is used for PNER (Luo, Xiao, & Chang, 2011), which takes part of speech (POS) feature, contextual feature, ontology feature into account, in order to probe this domain in an effective way. Also, a method for identifying product named entity using CRF was developed (Fang Luo et al., 2011) which introduces the domain ontology features to the CRF model.

In a learning-based approach, naive Bayesian classifiers are used for extracting attributes and values from product specifications and annotating them regarding a predefined set of attributes (B. Wu et al., 2009). The proposed method extracts attribute name and value from web pages for a given object. In this process, the web page is parsed into a tree and then both the value and name extraction are considered as the task of classifying nodes in the tree.

A dictionary-based approach using data provided by Wikipedia (DBpedia) was used (Eckhardt et al., 2014) for mention detection. In this way, a set of alternative names is made for each entity by using DBpedia labels for the entity. Then the text is processed sequentially to find all the possible text mentions about entities. Finally, for each mention, a set of candidate entities will be available that may be its target.

Freebase[18] is one of the most frequently used datasets as the data dictionary for finding the product mentions in text (Toh et al., 2012) (Lipczak, Koushkestani, & Milios, 2014). Freebase contains tens of millions of topics, thousands of types, and tens of thousands of properties. By comparison, English Wikipedia has over 4 million articles. Each of the topics in Freebase is linked to other related topics and annotated with important properties like movie genres and people's dates of birth. There are over a billion such facts or

[18] https://developers.google.com/freebase/

relations that make up the graph and they're all available for free through the APIs or to download from our weekly data dumps. Freebase will be ported into WikiData[19] and be retired on June 30, 2015. Other publicly available datasets that can be used for PNER are AIDA, IITB, MSN, AQUAINT, and the one of the ERD Challenge (Piccinno, Ferragina, & Informatica, 2014).

Usually, a combination of the NER approaches is used by the PNER systems. For example, the winner of CPROD1 competition[20] proposed a combination of a dictionary-based matching model, a rule template model, and a conditional random field model, that combines the results using a blending model (S. Wu, Fang, & Tang, 2012). The dictionary-based matching model is used to identify the products using a list of products that is extracted from annotated product dataset. The Rule Templates model leverages the products naming rules and several semantic information. The Conditional Random Field model is trained to match the potential patterns which cannot be provided by simple statistical analysis. Since these models leverage different information, a hybrid approach is proposed which combines the results of these different models.

While dictionary-based matching can easily handle the correct answer, rule templates are good at dealing with semantic pattern and human naming regulation, and conditional random field can fully utilize the potential sequence information. After getting the mention symbols, an interactive mechanism is used to recognize the whole products name and retrieve the product items.

In another example, a combination of rule-based and dictionary-based approaches is used to extract the product names in CPROD1 competition (Toh et al., 2012). Regular Expression Patterns are used to capture possible model names such as Samsung Galaxy S3 model "I9300RWDX" where alphabets are

---

[19] http://www.wikidata.org/wiki/Wikidata:Main_Page

[20] The goal of PRODucts Contest #1(CPROD1) competition (Melli & Romming, 2012) was to automatically recognize product mentions in the textual content and disambiguate which products in the product catalogue are referenced by the mentions.

followed by a dash and then a mixture of characters and digits. Valid manufacturer names and product names are extracted using dictionary-based features. Two sources including Freebase[21] and BBY[22] Product Archives were used to generate a large name gazetteer. The matching process is to check if a string matches an entry in the gazetteer list or not. String matching composed of unigram and bigram matching. In unigram matching, only current word (both in lowercase and as it is in the text) is used in string matching. In bigram matching, a group of the current word and its previous word; and a group of the current word and its next word are considered in string matching process.

A further step to information extraction regarded to semantic data can be to recognize an entity of an ontology in the text-based data. Here the problem is to match free-text, semi-structured or even structured product descriptions to their related entities in an ontology or structured records in a classification system. For example, (Kannan et al., 2011) developed a system to relate textual product sales offers received from independent merchants with Bing product classification system (product catalogue) that is used by Bing Shopping search engine. The system uses a three stage semantic parsing in order to understand the semantics of the product descriptions. These steps consist of tagging the offer with attributes, identifying possible parses based on the tags, and finally obtaining an optimal parse. The process tries to extract all existing attributes of a product catalogue in a text string, then select a parse in which the maximum number of attributes agree in their values respect to a product. The product that has the highest similarity score with the input description will be selected as the result.

(L. Zhang, 2009) developed an ontology for annotating HTML documents and an extractor to extract product attribute information from product pages. The HTML pages are transferred into OWL formatted documents by taking advantage of the proposed ontology. The entity recognition process starts with

[21] http://wiki.freebase.com/wiki/Freebase_API

[22] https://remix.mashery.com/member/register

stop-words removing from documents, then all the classes from the ontology that matches terms of the document are extracted as desired concepts.

Once the reference ontologies or standard classification systems are available, they can be used to annotate a given catalogue of products and services. Usually NER process produces a semantically annotated text from the input text. Part of speech (POS) tagging is the most basic type of annotation. POS tagging, usually is done as a starting step of word disambiguation, detects the syntactic role of a word in the sentence (subject, object, …) and functional role of the word (noun, verb). The syntactic role of a word is more needed in applications such as sentence translation while the functional role is more useful in the detection of a word sense in a context.

4.6 Summary

This research work proposes to exploit Vector Space Model (Manning et al., 2008) to measure the similarity ratio of documents in order to match providers' e-catalogues with a buyer's e-catalogue. Vector Space Model is an algebraic representation of documents as vectors in a high-dimensional space that is widely used by web search engines and other information retrieval systems. VSM uses the occurrences of terms in documents to produce a table of vectors. Having a vector model of documents, mathematical vector operations can be applied to determine the similarity of a document with another one or with a search query. The simplest example is to use the deviation angle between vectors of frequent terms to calculate the relevance between text documents. While it is used to deal with flat textual data (i.e. classical free text documents), IR is being extended, since the last two decades, so as to treat complex structured and semi-structured data (Tekli, Chbeir, & Yetongnon, 2009).

Concept-based information retrieval systems use the semantic data concepts in making the term vectors and show documents in the vectors of underlying semantic concepts instead of frequent terms. This enables the search engines to find the documents not only based on the exact containing terms, but also considering the synonyms and related terms.

Hence, recognising the covered concept from the documents being indexed is important for concept-based information retrieval system. Therefore, a fundamental part of such search engines can be a Named Entity Recogniser that extracts the desired searchable elements from related documents.

The most effective approaches to NER can be categorized as rule-based, dictionary-based and machine learning approaches (Melli & Romming, 2012) (Piccinno, Ferragina, & Informatica, 2014). The rule-based approach typically uses hand-crafted linguistic rules and seeks named entities by pattern-matching or template-filling. The dictionary-based approach assumes the presence of a dictionary of names of target types and identifies the names in text. Machine learning approaches are usually sequence classifiers that try to classify a sequence of terms in the text to a class of entities. Learning approaches fare better across domains, and can perform predictive analysis on entities that are unknown in a text.

5

E-catalogue Matching Engine

The basic concepts of the VSM have been reviewed in Chapter 4. In this chapter, the concepts will be extended to solve the problem of matching heterogeneous e-catalogues. Vector Space Model can be used to determine the similarity ratio between documents. The most important factor in finding similarity using this model is the term definition method. Since the similarity ratio has been calculated based on the occurrence of common terms, the way that the terms are defined specifies the similarity measure. Therefore, this research work extends the term definition in a way that can represent both semantic and syntactic features of the e-catalogues. Then similar e-catalogues can be calculated by comparing deviation of the angle between their vectors.

The proposed extension can take advantage of all available ontologies and schemas that have been provided for e-catalogues in various researches. The idea is to interpret each e-catalogue syntactically in its schema and semantically in an ontology that is made based on its product classification system. Schemas and ontologies will be added to the matching process by adding the syntax of the structure and semantic of the ontology to the indexing and searching mechanisms of Vector Space Model. The matching process uses the available syntactic and semantic metadata for interpreting each e-catalogue. In the case of unknown formats, it tries to use the semantics of known format to discover the

possible concepts that may exist in an e-catalogue. Before explaining the details of the matching mechanism in the following sections, Figure 5.1 shows the overall overview of the vectorization process in a conceptual diagram.

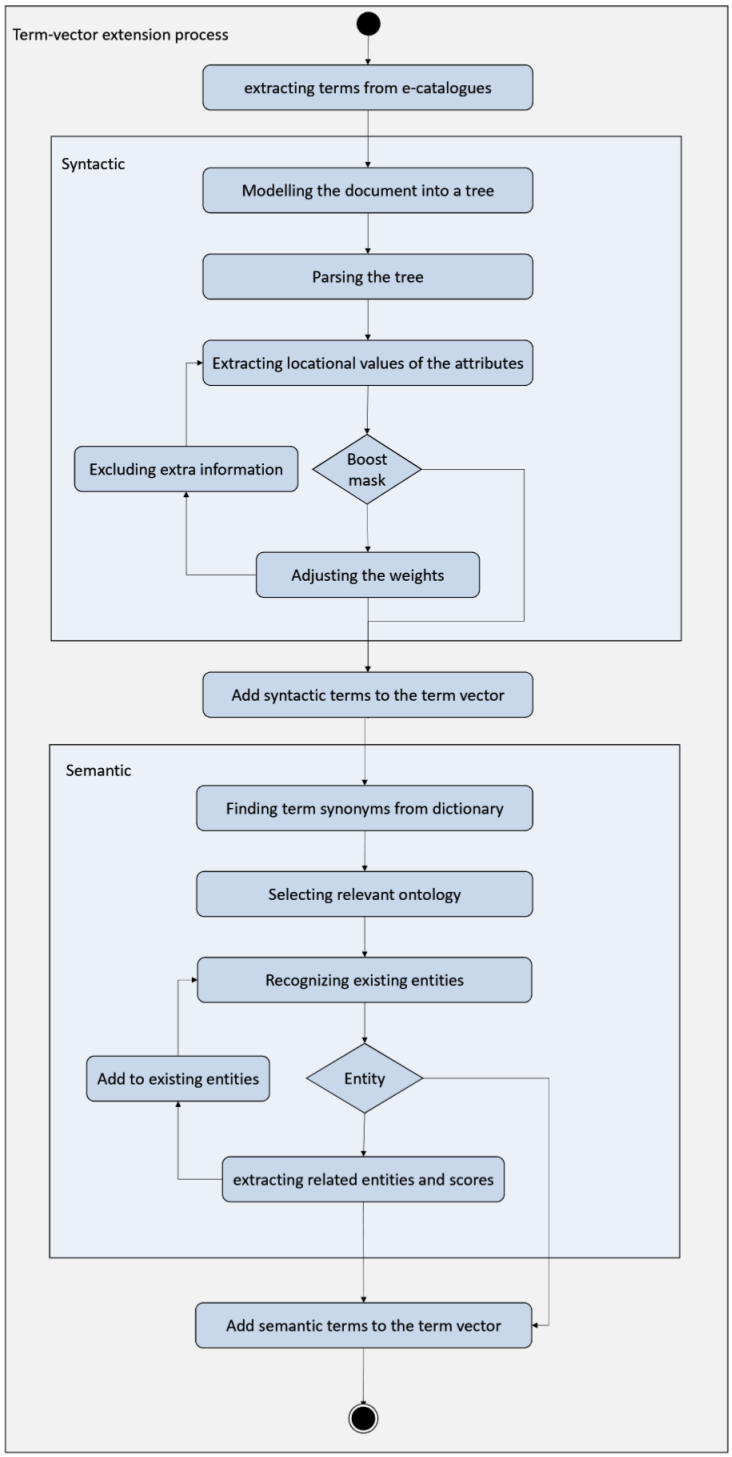

Figure 5.1 Term-vector extension process

Compared to traditional e-catalogue matching solutions that have transformation overheads and require knowledge about the target systems, the proposed solution is not dependent on parsing and converting different e-Catalogue formats.

The proposed approach for e-Catalogue matching uses combinations of values, names and paths of the attributes of structured e-Catalogue documents in term definition in order to find the syntactic correlation among e-Catalogues. This process is explained in the first section of this chapter.

In the second section, we will extend the matching mechanism by discovering and using potential semantic relationships of e-Catalogues in the term definition process. The best semantic resource that exists in an e-Catalogue is the product classification system that is used to define the products. Therefore, ontologies that are built from product classification systems such CPV, UNSPSC and eCl@ss (Stolz et al., 2014) will be used as the semantic reference to define the entities. These ontologies are rich sources of semantic information for interpreting product data of e-Catalogues (Vandic, Nederstigt, & Aanen, 2014). The semantic term definition process will determine existing entities and the semantic relationships among them in e-Catalogues. The idea is to enrich vector of each e-catalogue with semantic concepts that can be extracted from its terms. Adding extracted entities and their relations to the vectors will enable the matching process to match semantically related e-Catalogues.

In vector space based search systems, the definition and selection process of the terms is the major challenge. Actually, the terms are the searchable items and determine the matching mechanism and similarity measure. For example, if existing movie names in a text corpus are selected as the terms for a search engine, the system will be able to retrieve the documents about a movie when the name of the movie is being searched. As discussed in the fourth chapter, finding the occurrences of such entities like movie names or product mentions that will be used as the searchable elements from the text is a NER task.

In order to detect the existing meaningful product mentions in the e-catalogues, the third section of this chapter will discuss a NER process for recognising B2B Products from procurement documents. Term definition might represent the documents as vectors based on the recognized entities, properties

of the entities, available relations among entities, structure of the ontology and entity individuals. This step defines the matching elements or keywords for the search mechanism. This definition later determines the calculation method of semantic match degree between a query and search results.

5.1 Syntactic e-Catalogue Matching

In order to encounter the syntactic heterogeneity problem using VSM, three general cases have been considered for e-catalogues (Lee et al., 2007) (Mehrbod, Zutshi, & Grilo, 2014a).

- First, unstructured text such as PDF files which are common in online commerce.
- Second, structured or semi-structured e-catalogues which are unknown for the system such as enterprise-specific formats.
- Third, structured standard documents which are known for the system such as cXML and UBL e-catalogues.

XML is one of the most common formats for exchanging structured and semi-structured data and also standard e-catalogues in B2B e-commerce (Leukel et al., 2002). Among 25 e-catalogue standards, 16 of them are based on XML (Grilo & Jardim-Goncalves, 2013a). Hence technically, the syntactic matching process is to apply Vector Space Model to three groups of e-catalogues at the same time:

- Unstructured text documents
- XML documents
- Standard e-catalogues

5.1.1 Multilevel Term Definition

As the starting point, a Natural Language Processing tool is needed to extract the terms from e-catalogues. E-catalogue matching mechanism uses Lucene NLP analysers to tokenize, lemmatize and remove stop words from e-

catalogues. Since our application is to match e-catalogues and term matching is more valuable and important than phrase matching in this application, we filtered the stop words. Then, it makes a vector to represent the occurrence of terms in each e-catalogue document. Let D be a procurement document, after the tokenisation process, D can be presented as a set of its terms as:

$$D = \{t_1, t_2, t_3, ..., t_n\} \tag{5.1}$$

E-catalogues that are similar to a given search query can be calculated by comparing deviation of the angle between the vector of each e-catalogue and that of the query. If we present the term vector of document D as V(D), and let w_i be the weight of term t_i in the term vector, the similarity ratio of document D1 and D2 can be calculated as:

$$\text{Similarity (D1, D2)} = \frac{\text{V(D}_1\text{)} \,.\, \text{V(D}_2\text{)}}{\left|\text{V(D}_1\text{)}\right| . \left|\text{V(D}_2\text{)}\right|} \tag{5.2}$$

Where

$$\left|\text{V(D)}\right| = \sqrt{\sum_{i=1}^{n} w_i^2} \tag{5.3}$$

In order to associate the syntaxes in calculating similarity, levels of attributes in structured e-catalogues are also included in the term definition. XML documents are widely used to represent structured information. Any structured or semi-structured document can be shown using XML files. Hence, XML-based similarity becomes a central issue in the structured information retrieval. Since in conventional information retrieval, documents are unstructured data, Vector Space Model has been extended towards XML information retrieval (Leukel et al., 2002). Using these extensions structured and unstructured queries and documents can be presented in vector space model and the matching ratio of them can be calculated.

Hierarchical structures of XML documents are generally modelled as trees. In the traditional model, nodes of a tree represent XML elements and are labelled with corresponding element tag names. Since content is distributed at different levels of the tree, location of an attributes in the tree is effective on the value of the term (Tekli, Chbeir, & Yetongnon, 2009) and should be considered

in term extraction. Therefore, the name, value and location of attributes were exploited to define the terms.

Generally, locational paths of attributes in a structured document are considered as the terms. In this ways, values of attributes are disregarded in term extraction from structured documents. This approach is useful for structure-only comparing of documents (Lampathaki et al., 2009). But in the context of product features, the similarity measure is more sensitive to the values which have been saved in the e-catalogue structures. For example, we don't want to match two e-catalogues that have completely same structure but present different products. Therefore, in matching process of e-catalogues, values are crucial and are even more important than structures. Consequently, we used a structure-and-content tokenization process (Chen, Li, & Zhang, 2010) to define the terms.

```
<CatalogueLine>
  <Item>
    <Description>Paper</Description>
    <Name>Paper</Name>
    <StandardItemIdentification>
      <ID>12345678</ID>
    </StandardItemIdentification>
    <ManufacturerParty>
      <PartyName>
        <Name>Paper Manufacturer</Name>
      </PartyName>
    </ManufacturerParty>
  </Item>
</CatalogueLine>
```

Figure 5.2 A part of a structured e-catalogue (D1)

As an example, Figure 5.2 shows a portion of a UBL e-catalogue, D1, which is used by PEPPOL[23]. This e-catalogue can be presented using a tree as in Figure 5.3.

[23] Pan-European Public Procurement Online project, http://www.peppol.eu

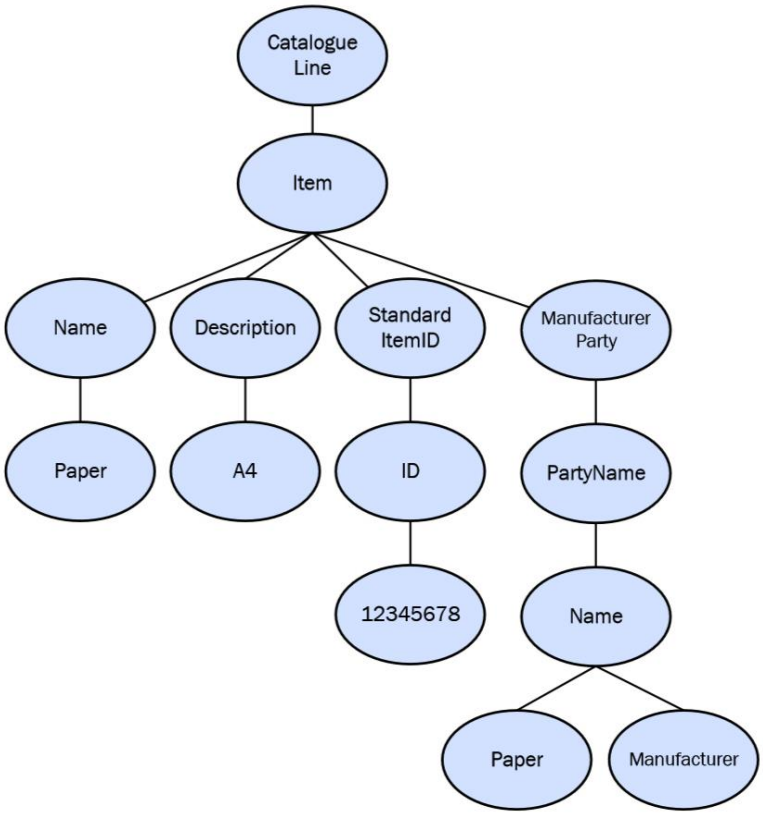

Figure 5.3 Tree model presentation of e-catalogue D1

In the matching process, this e-catalogue should be similar to e-catalogue D2 in Figure 5.4 that has the word *paper*. Moreover, it should have a higher similarity ratio with document D3 in Figure 5.4 that has the word *paper* in attribute *name* and even higher to document D4 in Figure 5.4, that has *paper* in the hierarchy of *name* and *item* and so on.

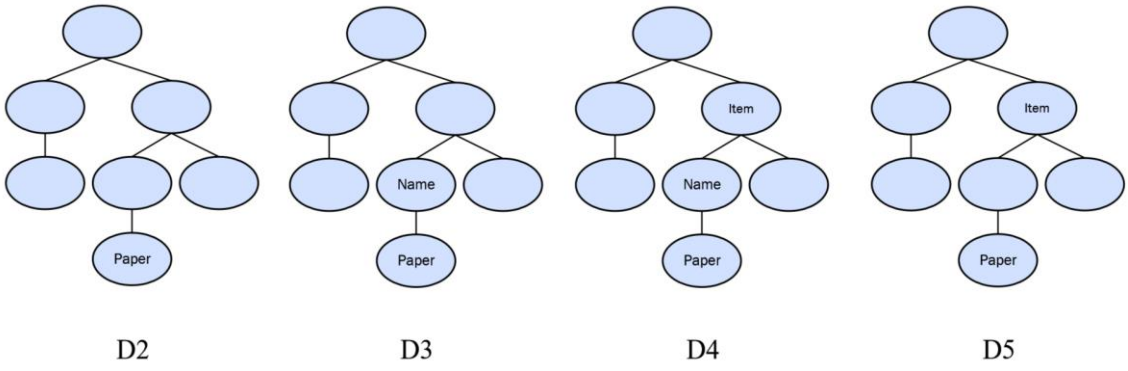

Figure 5.4 Similar e-catalogues to D1

One way of doing this is to define a term as a value together with its position within the XML tree. Figure 5.4 illustrates this representation. We use all the sub-trees of a document that contain at least one value as terms (Schmitz, Leukel, & Dorloff, 2005). In other words, we first take each value (*paper*) as a term. This help the matching process to match this document with unstructured documents or structured documents such as D2 that have same values but in a different structure. Next, we add values with the last level of their position (*name/paper*) to the terms. It helps matching process to increase similarity of D1 with structured documents such as D3. Then we continue adding levels of positions (*item/name/paper*) to the terms to the root of the tree. It keeps increasing similarity of D1 with structured documents such as D4. These position-adjusted terms constitute a set of leaf terms together with their paths to the root of the document tree as:

$$P = \left\{ //t.t \middle| t \text{ is a leaf} \right\} \tag{5.4}$$

Therefore, document D is not considered just as its basic term. It is seen as a union of its basic terms and its positional terms:

$$D \leftarrow D \cup P \tag{5.5}$$

And the term vector of the document will be:

$$V(D) = V_{Base}(D) \cup V_{Syn}(P) \tag{5.6}$$

Table 5.1 shows all possible terms for the tree of Figure 5.2. Note that having values attached to all the terms helps the search process to avoid matching documents with the same structure but different products. Therefore, D1 will have one common term with D2, two common terms with D3 and tree common terms with D4 which guarantees more ratio of similarity for documents with resembling structures. Note that having value attached to all the terms helps the search process to avoid matching documents with the same structure but different products.

Table 5.1 All possible terms for D1

Value	Terms
Paper Manufacturer	Manufacturer Name/Manufacturer PartyName/Name/Manufacturer ManufacturerPar-ty/PartyName/Name/Manufacturer Item/ManufacturerParty/PartyName/Name/Manufacturer CatalogueL-ine/Item/ManufacturerParty/PartyName/Name/Manufacturer Paper Name/Paper PartyName/Name/Paper ManufacturerPar-ty/PartyName/Name/Paper Item/ManufacturerParty/PartyName/Name/Paper CatalogueL-ine/Item/ManufacturerParty/PartyName/Name/Paper
12345678	12345678 ID/12345678 StandardItemIdentification/ID/12345678 Item/StandardItemIdentification/ID/12345678 CatalogueL-

	ine/Item/StandardItemIdentification/ID/12345678
A4	A4 Description/A4 Item/Description/A4 CatalogueLine/Item/Description/A4
Paper	Paper Name/Paper Item/Name/Paper CatalogueLine/Item/Name/Paper

Documents such as D5 should have a lower matching ratio with D1 as compared to its matching ratio with D3 and D4. Because the same value for an attribute (*item*) exists in both documents but not necessarily in the same path order. Therefore, the terms of Table 5.2 have also been added to the terms of D1 to cover this type of similarity. In order to give a lower similarity ratio to the e-catalogues that match D1 using the terms of Table 5.2 instead of the terms of Table 5.1 (don't have completely the same structure), the weight of a term is divided by twice the number of nodes between the value and the attribute. With this simple approach, we don't have to change the similarity formula as proposed in (Kim, Choi, & Park, 2007) and (Chen, Li, & Zhang, 2010).

Table 5.2 Additional terms for the last entry in Table 5.1

Value	**Terms**	**Weight Ratio**
Paper	Item/Paper	½
	CatalogueLine/Item/Paper	½
	CatalogueLine/Name/Paper	½
	CatalogueLine/Paper	¼

5.1.2 Boosting Masks

Standard e-catalogues are sources of diverse types of information. They usually include not only product data but also general document information and partners' data. This extra information can mislead the product search process. Furthermore, various attributes of product data can have different values in the matching process. For example, classification code of a product is more important than a description of the product in the matching process. Hence, tables of coefficients are used for known formats to adjust the impact of each attribute in the matching process. Let w_i^p be the weight of the positional term p_i in P, let C_p be the coefficient assigned to p_i in the boosting mask, where $C_p \in [0,1]$, the following equation calculates the discussed weight for each positional term:

$$w_i^p = \begin{cases} 1 & otherwise \\ C_p & \exists boostingmask, 0 \le C_p \le 1 \end{cases} \tag{5.7}$$

Now, the syntactic term vector of an e-catalogue can be shown as:

$$V_{Syn}(D) = \left\{ w_i^p \middle| p_i \in P \right\} \tag{5.8}$$

And

$$\left| \mathrm{V_{Syn}(D)} \right| = \sqrt{\sum_{i=1}^{m} {w_i^p}^2} \tag{5.9}$$

Figure 5.5 shows the coefficients for the sample e-catalogue of Figure 5.2. These coefficients are values between 0 and 1 which are multiplied by the weights of terms. Undesired information such as partners' data can be simply excluded from matching process by putting 0 coefficients. Using this simple mechanism, a new known structure can be easily added to the search system. Default values for all coefficients are 1 which reduces the status of an e-catalogue to an unknown structure for the matching process.

```
<CatalogueLine>
  <Item>
    <Description>.5</Description>
    <Name>.8</Name>
    <StandardItemIdentification>
      <ID>1</ID>
    </StandardItemIdentification>
    <ManufacturerParty>
      <PartyName>
        <Name>0</Name>
      </PartyName>
    </ManufacturerParty>
    <CommodityClassification>
      <ItemClassificationCode listID="CPV">1</ItemClassificationCode>
    </CommodityClassification>
  </Item>
</CatalogueLine>
```

Figure 5.5. Coefficients for the sample e-catalogue

5.2 *Semantic e-Catalogue Matching*

In order to solve the syntactic heterogeneity in the previous section, we applied Vector Space Model to the e-catalogues matching problem (Mehrbod, Zutshi, & Grilo, 2014a). In this section we want to expand the solution to cope with the semantic heterogeneity of e-catalogues as well (Mehrbod, Zutshi, & Grilo, 2014b). This section uses not only the attributes and their locations, but also the semantic interpretation of the terms in the matching process.

One of the most common approaches in semantic Vector Space Model is to expand the terms using their synonyms in a vocabulary set. Many researches exploit WordNet lexical database to enrich the vectors with the synonyms of the terms (Turney & Pantel, 2010). WordNet is a large lexical database of sets of synonyms that is freely and publicly available. Synonym sets are interlinked by means of conceptual-semantic and lexical relations. The structure of WordNet makes it a useful tool for computational linguistics and natural language processing. This approach is useful, simple and practical. Therefore, it is used as a part of the proposed term expansion process and will be explained in the next section, but the proposed semantic matching mechanism is more comprehensive than this. Furthermore, the proposed approach uses domain specific semantic resources for e-catalogue interpretation.

5.2.1 Ontology Deriving

Semantic applications require semantic references for interpreting data. The best semantic resource that exists in an e-catalogue is the product classification system that is used to define the products and services of the e-catalogue. As previously mentioned, many efforts have been made to build up product ontologies from existing classifications and standard classification systems. These ontologies are rich sources of semantic information for interpreting product data of e-catalogues (Schmitz, Leukel, & Dorloff, 2005) and can be used to enrich product descriptions with information from existing data sources.

As highlighted in chapter 3, (Stolz et al., 2014) developed a generic, semi-automated method and tool called PCS2OWL for deriving OWL[24] ontologies from product classification standards and proprietary category systems such as CPV, UNSPSC and eCl@ss. Making ontologies using classification systems is usually superior to building them up manually that is usually tedious, costly, and time-consuming in the domain of products and services. To date, ontologies for 13 product classification systems of different scopes, sizes, and structures such as CPV, UNSPSC and eCl@ss have been created using the PCS2OWL tool and are available online. Web Ontology Language (OWL) can be used to explicitly represent the meaning of terms in vocabularies and the relationships between those terms in an ontology. OWL has more facilities for expressing meaning and semantics than XML, RDF, and RDF-S, and thus OWL goes beyond these languages in its ability to represent machine interpretable content.

In addition to these ontologies that have been developed based on e-procurement classification systems for modelling procurement products, some ontologies have been developed for considering more details of the e-procurement process. Nečaský (Nečaský et al., 2014) argue that none of available product or e-procurement ontologies such as LOTED, MOLDEAS, KOCIS, Call for Anything, GoodRelations appears sufficient for matchmaking demand

[24] www.w3.org/OWL

and supply on the procurement market. Therefore, they designed Public Contract Ontology (PCO) that specifically targets public contracts.

PCO is comprised of two main classes, called *pc:Contract* and *pc:Tender* as the core concepts that are complemented using other ontologies' concepts. For example, it reuses *gr:offering* for modelling contract and tender items, *gr:PriceSpecification* for modelling pricing conditions and *gr:BusinessEntity* for modelling contracting authorities from GoodRelations Ontology[1]. Regardless of supplementary concepts such as locations, dates and so on that can be defined or reused to enrich such ontologies, one of the most concepts for a procurement ontology is the class used for modelling the classification of products and services. PCO reused Simple Knowledge Organization System (SKOS[2]) to express code lists and classifications based on CPV.

In a similar approach called MOLDEAS (Alvarez et al., 2012), an ontology for public procurement data has been designed. The main entity of this ontology is *moldeas-onto:Notice* class for modelling the tender notices, that is complemented by other classes such as *moldeas-onto:Country* and *moldeas-onto:Region* for modelling Geographical information. MOLDEAS reuses concepts of existing vocabularies and ontologies as well. For example, *gr:ProductOrServiceModel* class of GoodRelations is reused for describing contract type. Similar to PCO, SKOS concepts have been reused for modelling the product classification system but it reused GoodRelations entities to link product scheme classifications.

Public Procurement Ontology, PPROC, is another ontology developed based on PCO. The main difference between this ontology and other similar ontologies is that PPROC is designed not only for procurement data modelling and data sharing of public procurement contracts but also for management of whole procurement process. Furthermore, it focuses on transparency on data sharing in public procurement.

PPROC reused the core concepts of PCO such as *pc:tender*, *pc:supplier* and *pc:contractingAuthority*, complemented it with inherited or new concepts such as *pproc:contract* and detailed attributes such as *pproc:awardDate* necessary for the detailed description of buyer profile. In contract classification, PPROC has extended the *pc:kindScheme* by defining several class taxonomies for different types of contracts. For product or service classification, a combination of CPV

vocabulary as the primary and *gr:offering* class of GoodRelations as the extended taxonomy are exploited (Esteban, 2015; Muñoz-soro et al., 2016).

As mentioned, the classification of the products and services that constitute a B2B document is one of the most useful information available in such ontologies for developing a product matching service. These classifications, normally are defined based on the available standard classification systems such as CPV, UNSPSC and GPC. Even though the structure and features of such Knowledge Organization Systems are very heterogeneous, they enable users to annotate information providing an agile mechanism for performing tasks such as exploration, searching, automatic classification or reasoning (Alvarez et al., 2012).

Since the proposed e-catalogue matching mechanism focuses on utilizing the product data on the search process, the vocabulary used on the procurement ontologies appear to have a higher impact on the matching results than the other classes defined in the ontologies. In a tender, e-catalogue, contract award notice or generally a procurement document rather than the products or services, other main information sections such as contract restrictions, buyer and supplier profile and deadlines are also available. Even though this information is a valuable resource of data for matching a search query to available calls or e-catalogues, employing it in matching process is another search domain and beyond the scope of this research work.

Therefore, the proposed e-catalogue matching engine doesn't use such data and focus on exploiting product related information such as title, description and classification of product items in the matching process. The existing procurement ontologies not only focused on all aspects of procurement data at the same time that may not have value added in product matching, many concepts of such ontologies have designed to address legislation specifics and process management issues than the product search process.

5.2.2 Ontological Matching

As discussed, it is straightforward to have an ontology for the semantic presentation of each e-catalogue. In the case of standard formats, the relevant

ontologies have been published by making ontologies from their product classification systems. For enterprise specific e-catalogues, ontologies can be provided by their companies using available tools such as pcs2owl. Once the reference ontologies have been derived from the standard classification systems, they can be used to describe a given catalogue of products and services (Beneventano & Montanari, 2008).

Semantic expansion of the e-catalogue matching process aims to find semantically related terms between e-catalogues. The idea is to enrich vector of each e-catalogue with semantic concepts that can be extracted from its terms. In this way the term-vector of each e-catalogue is expanded by terms of all semantic concepts that exist in the e-catalogue. Each distinct concept or property represents one extra term in the vector space. Adding these terms to the vectors enables the matching process to find semantically similar e-catalogues:

$$V(D) = V_{Base}(D) \cup V_{Syn}(D) \cup V_{Sem}(D) \tag{5.10}$$

In other words, each e-catalogue is considered as a combination of semantic entities and its previous terms including the basic keywords and the syntactic elements.

$$D \leftarrow D \cup P \cup E \tag{5.11}$$

Where E is the set of existing semantic concept in document D, extracted by the semantic term expansion process and P is the positional terms that have been extracted by syntactic term extension process.

Term expansion process that is shown in Figure 5.6 using pseudocodes, tries to extend *V(D)*, the term-vector of an e-catalogue, by synonyms, similar and semantically related concepts to the terms of the e-catalogue. The process starts with determining existing entities in the e-catalogue.

```
expand_vector (e-catalogue)
{
    TermVector = extractTermVector (e-catalogue);
    foreach Term in TermVector
    {
        TermVector = TermVector ∪ Select_synonyms_term (Term);//select synonyms from WordNet
    }
    Ontologies = selectOntologies (e-catalogue);
    EntityVector = {};
    foreach Ontology in Ontologies
    {
        foreach Term in TermVector
        {
            Entity = selectEntity (term, Ontology); //check if the term is an entity in the ontology
            EntityWeightCoefficient Entity = 1; //equivalency coefficient of syntactic and semantic similarity
            if ((Entity ≠ Ø) and (Entity ∉ EntityVector))
            {
                RelatedEntitiesSet = {};
                RelatedEntitiesSet = selectRelatedEntitiesSet (Entity, Ontology, EntityWeightCoefficient Entity);
                EntityVector= EntityVector ∪ {Entity} ∪ RelatedEntitiesSet;
            }
        }
    }
    return (EntityVector U TermVector)
}
```

Figure 5.6. Term expansion process

```
selectRelatedEntitiesSet (Entity, Ontology, EntityWeightCoefficient Entity)
{
    PREFIX owl: <http://www.w3.org/2002/07/owl#>
    PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
    SELECT ?RelatedEntities
    WHERE
    {
        {
            ?RelatedEntities owl:equivalentClass | rdfs:subClassOf ?Entity.
        }
        UNION
        {
            ?Entity owl:equivalentClass | rdfs:subClassOf ?RelatedEntities.
        }
    }
    If (RelatedEntities ≠ Ø)
    {
        Results = RelatedEntities;
        foreach RelatedEntity in RelatedEntities
        {
            EntityWeightCoefficient RelatedEntity = ½ * EntityWeightCoefficient
      Entity;
            Results = Results U selectRelatedEntitiesSet (RelatedEntity, On-
      tology, EntityWeightCoefficient RelatedEntity);
        }
        return (Results);
    }
    else
        return (Ø);
}
```

Figure 5.7. Related entities extraction process

```
selectOntologies (e-catalogue)
{
    If ( isStandard (e-catalogue) )
        {
            return ontology that is made based on classification system
            of the e-catalogue;
        }
    else
        {
        return all ontologies from the ontology repository;
        }
}
```

Figure 5.8. Relevant ontology selection

In order to determine existing entities, first the relevant ontology for the e-catalogue is selected based on its classification system (Figure 5.8). If no ontology can be found in the repository for the e-catalogue that can occur especially in the case of unstructured and unknown formats, the algorithm uses all available ontologies in the system and tries to recognize any available entity in the e-catalogue. The ontology repository of the system can be enriched by all available ontologies from various accessible resources such as (Hepp, 2008), (Stolz et al., 2014) and (T. Lee et al., 2006).

Let O be the selected ontology for e-catalogue D, the process should determine E as a subset of O that consists of the semantic concepts of D:

$$E \subset O \tag{5.12}$$

For this purpose, each term is compared with the entities of the ontology to check if there is an entity for describing the term. The list of terms already enriched by the synonyms of terms. So far, E can be considered as:

$$E = \left\{ e_{t_1}, e_{t_2}, e_{t_3}, ..., e_{t_n} \right\} \subset O \tag{5.13}$$

Where e_{t_i} is an entity for describing term t_i of document D in ontology O. While e_t can be determined from t using exact string matching, a NER based method has been used for this function. The NER method will be explained in details in the next section.

Next, a list of all related entities to each e_t is extracted using an iterative process that is implemented using a recursive procedure shown in Figure 5.7. The procedure that is called *selectRelatedEntitiesSet,* accepts an entity and an ontology as input and recursively returns all related entities to the given entity from the given ontology. Therefore, for each e_t we will have a set of related entities that will be added to E in order to enrich $V_{Sem}(D)$:

$$E = \{e_{t_1}, e_{t_2}, e_{t_3}, ..., e_{t_n}\} \cup \{e_1, e_2, e_3, ..., e_o\} \subset O \tag{5.14}$$

And,

$$V_{Sem}(D) = \{w_i^e | e_i \in E\} \tag{5.15}$$

In order to extract related entities, *owl:equivalentClass* and *rdfs:subClassOf* predicators were used. The procedure can be extended easily with a more complete list of predicators. Note that the entity extraction process discards repetitive entities in order to avoid infinitive loops in the extraction of the related entities.

In order to justify the impact of the extracted entities on semantic similarity ratio, by the geodesic distance of each entity to the relevant term, a weighting coefficient has been considered for each entity. The default value for the weighting coefficient is 1 and in each level of relationship chain, it is divided by 2. This weighting coefficient will be multiplied by the weight of the entity in the term vector. The default value is set to 1 that shows an entity in the first level of relationship chain, shown as e_t, has the same value as the relevant term *t*. But the values of the entities decrease with the increase in their geodesic distance in ontology graph from the relevant terms:

$$\text{EntityWeightCoefficient} = \frac{1}{2 \times \text{geodesic\_distance}(e_t, e)} \tag{5.16}$$

Therefore, w_i^e that is the weight of semantic concept e_i in $V_{Sem}(D)$ can be calculated as:

$$w_i^e = \begin{cases} 1 & \forall t \in D \mid \exists e_t \in O \\ \dfrac{1}{2 \times \text{geodesic\_distance}(e_t, e)} & \forall e \in O \mid \exists (?e_t : \text{Relation}_{\text{Distance}}\ ?e) \\ 0 & \neg \exists e_t \in O \end{cases} \quad (5.17)$$

As mentioned the numerator of the Entity Weight Coefficient fraction is set to 1 by default that shows an entity (in the root of relationship graph) has the same value as a term in the term vector. To give a higher importance to the semantic similarity than the syntactic similarity, the default value can be increased to c_e. By assigning different coefficients to the weights, the impact of each type of similarity on the overall similarity ratio can be controlled:

$$|V(D)| = c_d . |V_{Base}(D)| + c_p . |V_{Syn}(D)| + c_e . |V_{Sem}(D)| \quad (5.18)$$

or,

$$|V(D)| = \sqrt{\sum_{i=1}^{n+m+o} c_i w_i^2} \quad (5.19)$$

Finally, the extracted entities are added to the term vectors that enables the matching process to find semantically related e-catalogues. Consequence term vectors have some terms related to the semantic concepts, in addition to terms relevant to syntactic structures. The similarity ratio that is calculated using such vectors is a combination of all these similarity measures.

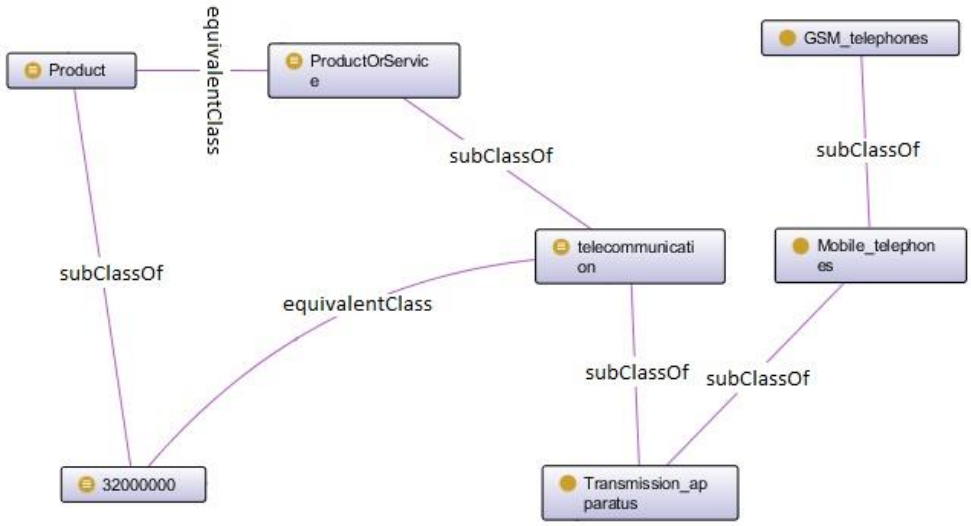

Figure 5.9 A sample ontology based on CPV

As an example, suppose the ontology of Figure 5.9 is selected for the e-catalogue of Figure 5.10. This sample ontology is made based on the CPV standard classification system and will be used to expand the terms of the sample e-catalogue.

```
<CatalogueLine>
  <Item>
    <Description>Smart phone</Description>
    <Name>Mobile</Name>
    <StandardItemIdentification>
      <ID>RM-917</ID>
    </StandardItemIdentification>
    <ManufacturerParty>
      <PartyName>
        <Name>Nokia</Name>
      </PartyName>
    </ManufacturerParty>
    <CommodityClassification>
      <ItemClassificationCode listID="CPV">32250000</ItemClassificationCode>
    </CommodityClassification>
  </Item>
</CatalogueLine>
```

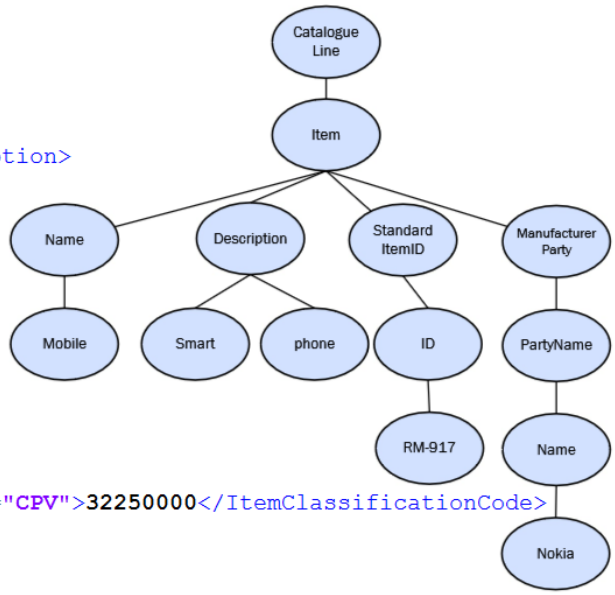

Figure 5.10 A part of a structured e-catalogue

Table 5.3 Syntactic Terms for *Mobile* in Figure 5.10.

Value	Terms	Ratio
Mobile	Mobile	1
	Name/Mobile	1
	Item/Name/Mobile	1
	CatalogueL-ine/Item/Name/Mobile	1
	Item/Mobile	½
	CatalogueLine/Item/Mobile	½
	CatalogueLine/Name/Mobile	½
	CatalogueLine/Mobile	¼

Table 5.4 Related entities to the term *Mobile*.

Term	Related Entity	Coefficient
Mobile	Mobile_telephones	1
	GSM_telephones	$^1/_2$
	Transmission_apparatus	$^1/_2$
	telecommunication	$^1/_4$
	ProductOrService	$^1/_8$
	Product	$^1/_{16}$
	32000000	$^1/_8$

Table 5.3 shows the syntactic terms, *P*, that were added to the vector of this e-catalogue for the value *Mobile* in the attribute *name* by syntactic term ex-

pansion process. For a complete sample of the syntactic term expansion please refer to the previous section. Semantic term expansion process extracts the related entities to each of these terms, *E*, from the selected ontology *O*. Table 5.4 shows the related entities to the term *Mobile* extracted from the sample ontology.

5.2.3 Synonym Matching

As discussed, product data is distributed in various levels in structured and semi-structured e-catalogues. Attribute values that are the leaves of the tree in a tree presentation of such documents, constitute the main product aware data. For interpretation of such data, we have to use domain specific semantic resources. Therefore, as explained in the previous part, we used product classification ontologies to expand the terms.

Since the terms in other levels of the tree are not necessarily domain-specific terms, in this section we want to use a more general semantic resource to interpret the data of upper levels. One of the most common approaches in semantic Vector Space Model is to expand the terms using their synonyms in a vocabulary set. Many researchers exploit WordNet lexical database to enrich the vectors with the synonyms of the terms (Turney & Pantel, 2010).

WordNet is a large lexical database of sets of synonyms that is freely and publicly available. Synonym sets are interlinked by means of conceptual-semantic and lexical relations. The structure of WordNet makes it a useful tool for computational linguistics and natural language processing. This approach is useful, simple and practical. Therefore, we have used it as a complimentary part of our semantic term expansion process.

As an example, consider term *CatalogueLine/Item/Name/Mobile* in Figure 5.10. According to Table 5.4, seven related entities are available for attribute value *Mobile* and this term will be expanded as Table 5.5 using the explained approach in the last part.

Table 5.5. Ontological term expansion.

Term	Related Entity	Expanded Terms
Mobile	Mobile_telephones GSM_telephones Transmis-sion_apparatus telecommunication ProductOrService Product 32000000	CatalogueLine/Item/Name/ Mobile_telephones CatalogueLine/Item/Name/ GSM_telephones CatalogueLine/Item/Name/ Transmission_apparatus CatalogueLine/Item/Name/ telecommunication CatalogueLine/Item/Name/ ProductOrService CatalogueL-ine/Item/Name/Product CatalogueL-ine/Item/Name/32000000

In order to complete term expansion process using synonyms from WordNet, we consider *element* and *component* as synonyms of *item;* and *label*, *brand* and *fullname* as synonyms of *name* in the dictionary. Table 5.6 shows the results of the synonymous expansion of the term *CatalogueL-ine/Item/Name/Mobile* in Figure 5.10 and one of its ontological expansions (term *CatalogueLine/Item/Name/Mobile_telephones*).

Table 5.6. Synonymously term expansion.

Term	Synonymously Expanded Terms
CatalogueL-ine/Item/Name/Mobile	CatalogueL-ine/Item/Name/Mobile CatalogueLine/element/Name/

	Mobile CatalogueLine/component/Name/Mobile CatalogueLine/Item/Label/Mobile CatalogueL-ine/Item/Brand/Mobile CatalogueLine/Item/Fullname/Mobile CatalogueLine/element/Label/Mobile CatalogueLine/element/Brand/Mobile CatalogueLine/element/Fullname/Mobile CatalogueLine/component/Label/Mobile CatalogueLine/component/Brand/Mobile CatalogueL-ine/component/Fullname/Mobile
CatalogueL-ine/Item/Name/Mobile_telephones	CatalogueL-ine/Item/Name/Mobile_telephones CatalogueLine/element/Name/Mobile_telephones CatalogueLine/component/Name/Mobile_telephones CatalogueL-ine/Item/Label/Mobile_telephones CatalogueL-

	ine/Item/Brand/Mobile_telephones CatalogueLine/Item/Fullname/Mobile_telephones CatalogueLine/element/Label/Mobile_telephones CatalogueLine/element/Brand/Mobile_telephones CatalogueLine/element/Fullname/Mobile_telephones CatalogueLine/component/Label/Mobile_telephones CatalogueLine/component/Brand/Mobile_telephones CatalogueL-ine/component/Fullname/Mobile_telephones

5.3 B2B-Product NER

As explained in the previous section, the proposed semantic term expansion compares each term of a vector with the entities of the relevant ontology to check if there is an entity for describing the term. The process adds the potential related entities to the vector in order to enrich the vector with the semantically related concepts to the existing terms. In order to find the entities from the e-catalogues the matching mechanism uses NER during indexing phase.

Hence, as a requirement of the developed e-catalogue matching engine, this section presents a B2B-Product NER process that can serve as the basis for other B2B information systems such as semantic search mechanisms and document classifier systems in B2B e-Commerce as well. Although several works have been done on developing semantic search, semantic modelling and se-

mantic matchmaking on B2B documents, a NER model for entity recognition in B2B context is the missing part such works. This will help the semantic search mechanisms to find product mentions from various B2B documents that can be a necessary block for e-procurement information retrieval systems such as the proposed e-catalogue matching engine, and a complementary work for e-procurement semantic data modelling systems such as (Nečaský et al., 2014) and (Esteban, 2015). B2B-Product NER can help the information retrieval systems in extracting important and meaningful keywords in both data indexing and query generating tasks. Furthermore, the recognised elements can comprise the dynamic classifications for faceted search services in information retrieval systems.

For semantic data modelling systems (Nečaský et al., 2014)(Esteban, 2015), the NER process can help them in extending the modelling to new data resources. These systems that rely on the concept of linked open data, develop an ontology to model and publish the linked data; and a search system to query call for tenders. Mapping the data to the ontology concepts is the most expensive step in this process (Esteban, 2015). Therefore, usually the mapping and data transformation is done for the main data fields from known structures. The proposed NER mechanism can extract elements that are covered in text descriptions and other unstructured attributes and also can reduce the manual steps of the mapping in annotating new data resources that have different schema. To achieve this purpose, the results of the NER process can be saved into RDF or similar structures as defined in the procurement semantic models that later can be queried using SPARQL.

Among various kinds of NER techniques, learning-based approaches show reasonable results while requiring less manual work. Unfortunately, supervised learning approaches require large amounts of annotated training data in order to be effective. Since making such annotated data needs manual efforts, this thesis proposes to use standard e-Catalogues such as UBL, BMEcat and cXML as the training set in order to recognize entities from other e-Catalogues. The schema of a standard e-Catalogue is known and can be used to find the place of product mentions in the document. Furthermore, products in a standard e-Catalogue are usually defined according to a standard product classifica-

tion system such as CPV. These relationships are represented using identifiers that refer to the relevant product classes in a given classification system.

The main objective is to recognize mentions of products in e-catalogues and map them to the relevant entities in the given products classification system. This will help the search mechanism to match various expressions of the same products.

As mentioned the learning-based NER mechanisms usually require massive training sets to be accurate. Since providing a large annotated product dataset that could be used as the training set for CRF model needs manual efforts and can be very expensive and time-consuming, a self-learning method (Teixeira, Sarmento, & Oliveira, 2011) (Vieira et al., 2015) (Putthividhya, Hu, & Ave, 2011) has been used to bootstrap the training set.

This section adopts a supervised named entity recognition method for product named entity extraction problem from tender notices. The proposed method uses e-tenders as the training set in order to learn to recognize B2B-Product entities. The idea here is to use these already known product mentions and their references as the training data to train the model and then use the trained model to recognize the product mentions from other B2B documents.

5.3.1 Bootstrapping

While gazetteers can be used to perform named entity recognition through lookup-based methods, ambiguity and incomplete dictionaries lead to a relatively low recall. A learning-based approach which uses more general features can achieve higher recall while maintaining reasonable precision, but typically requires expensive annotated training data. In order to provide such training data, bootstrapping methods have been proposed to train learning-based NER models.

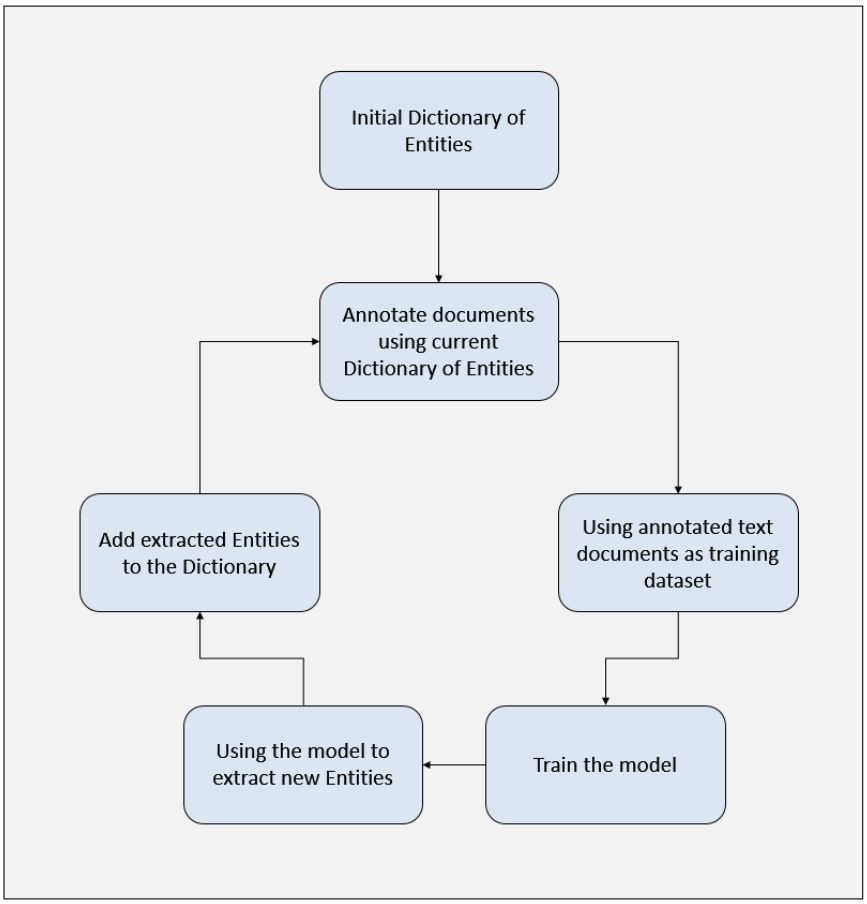

Figure 5.11 Bootstrapping process

While several bootstrapping approaches have been proposed for training NER classifiers, the general technique which is illustrated in Figure 5.11 is to start from a small set of labelled training examples and generate more training data from unlabelled data. Usually, a dictionary-based approach is used to annotate a text collection using an initial dictionary of already known entities. Then the annotated collection is used as the input for training the NER model. Finally, the trained model is used to extract new named entities from the text. Even though it is not crucial, this procedure can be repeated iteratively by adding new entities to the initial dictionary until the trained model satisfies the desired measures for the application.

For example in a bootstrapping approach, first, a labelled dataset is automatically generated by matching a manually prepared initial seed list of 6312 brand names to an unlabelled dataset (Putthividhya, Hu, & Ave, 2011). Then,

these auto-labelled training set is used to train a classifier to identify new entities from a separate set of unlabelled data. Thirdly, newly discovered entities are added back to the seed list. Thus, the original classifier for entity extraction can be improved through an expanded seed list i.e. by adding these newly discovered entities back to the seed dictionary, the supervised NER system can be retrained with the extended seed list.

An iterative bootstrapping method for extracting names of person was started by annotating person names on a dataset of 50,000 news items (Teixeira, Sarmento, & Oliveira, 2011). This was performed using a simple dictionary-based approach. Using the annotated news as the training set, a classification model was built based on Conditional Random Fields (CRF). Finally, the trained CRF model was used to extract the entities from the text. The extracted entities were again used for annotating the training set that was used to train a new CRF model. This cycle was repeated until the NER model stabilized. Reported results show that this bootstrapping approach stabilizes after 7 iterations, achieving high values of precision (83%) and recall (68%).

Also, a dictionary-based NER framework based on a Local Filters model was used to recognize Persian named entities (Khormuji, 2014). The method is first to detect named entity candidates using lookup dictionaries and second to filter false positives by filtering out noisy matches of the first stage. The integration is done by feeding the output of the dictionary-based system as the training set to a machine-learning classifiers.

A dictionary of known entities can also be used as a complimentary feature into a classifier in learning-based entity extraction[25]. But in bootstrapping, these seed values are used to either automatically generate labelled training data or to extend the initial dictionary itself. The method and the implementation details of the bootstrapping process depend on the application and can be different in various cases.

[25] For example this option is available in Stanford NER by adding RegexNER annotator into the NER pipeline

5.3.2 Learning-based B2B NER

We employ the idea of using a dictionary for providing the annotated training set needed for supervised learning from bootstrapping approaches. Although in bootstrapping approaches the training set is produced automatically, but the initial dictionary is still prepared manually by the experts (Putthividhya, Hu, & Ave, 2011). In this section, an automatic approach has been used to prepare the initial dictionary and no manual step exists on the proposed method.

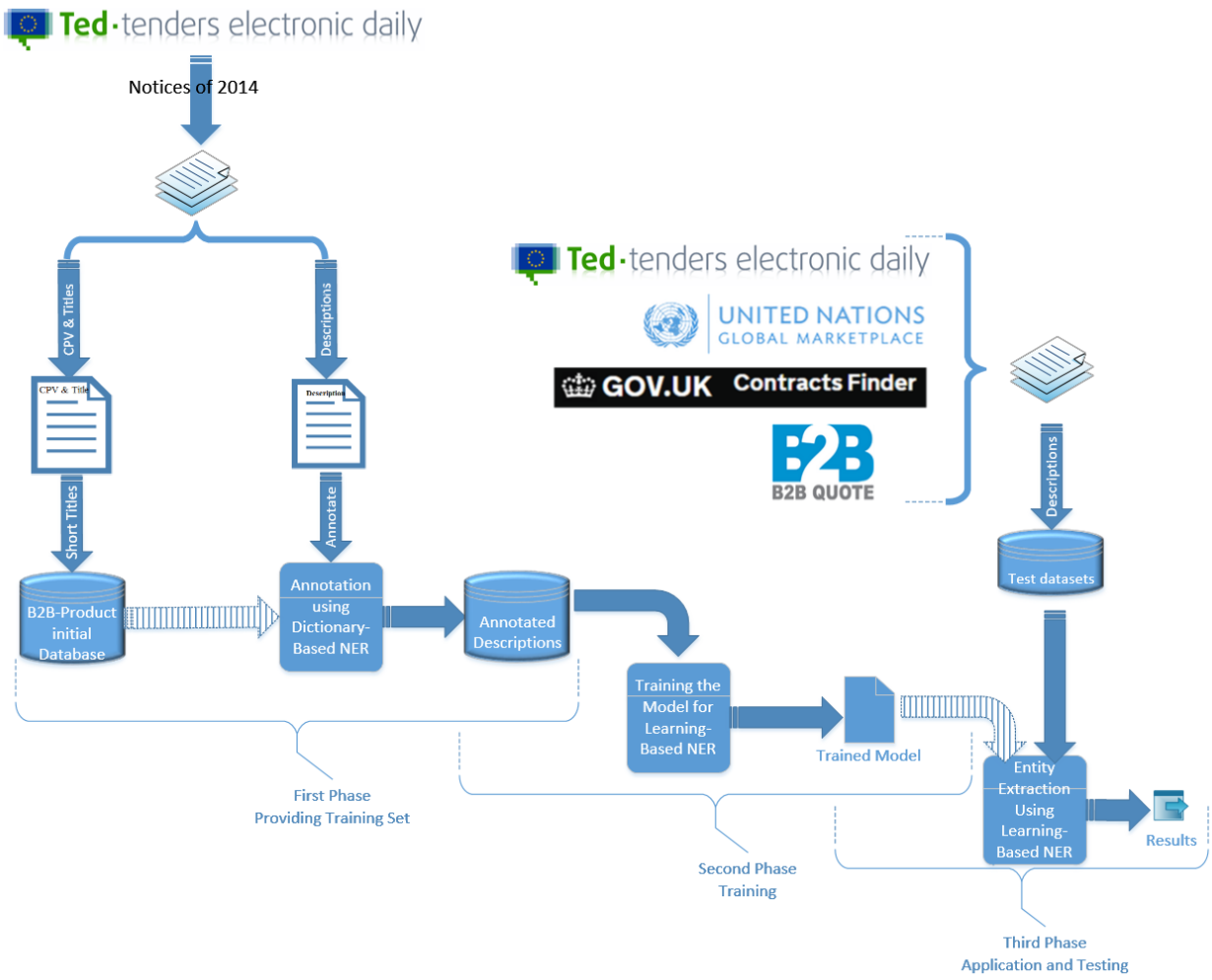

Figure 5.12 B2B-Product NER training and test Process

Figure 5.12 summarises all the process including making the training dataset, training the model and application of the model. As it can be on the Figure, in summary, the process starts by preparing a dictionary of B2B products from online available resources of B2B tenders. This dictionary is used in a gazetteer-based approach to prepare an annotated corpus. The annotated corpus has comprised the training set for training a learning-based model. Finally, the trained model is tested to extract B2B-Product named entities from four different datasets of B2B tenders. The following subsections explain the first and second phases and the next section will explain the third phase in details.

5.3.2.1 Making the dictionary

E-procurement documents such as Contract Notices, e-Tenders and e-Catalogues contain information about products and services. For example in Contract Notice commodities being procured are specified, Tender has the products being quoted as per the call and Catalogue has the products from the supplier's inventory (Ghimire, Jardim-Goncalves, & Grilo, 2013). Because of open tendering policy that tries to open up sufficient and fair competition between suppliers, e-Tenders are the most publically available resources of e-procurement documents. This huge resources of e-procurement documents can provide the required data for making the training set.

Tendering is a kind of reverse auction in which suppliers bid on the services or goods that buyers need (Du, 2009). In this procedure, different bidders generate competing offers on tenders and look to obtain an award of business activity in works, supply, or service contract (Dorn et al. 2009). In order to have a great improvement on the accessibility and transparency of tenders and provide equal opportunities to all suppliers, e-Tendering Marketplaces should be accessed anywhere globally. The publically noticed tenders in such marketplaces are used by various companies to find business opportunities.

Hence, we used e-Tenders of public Tendering Marketplaces as the data source. Tenders Electronic Daily (TED)[26] is the online version of the "Supple-

[26] ted.europe.eu

ment to the Official Journal" of the EU, dedicated to European public procurement. TED provides free access to business opportunities from the European Union, the European Economic Area and beyond. Every day, from Tuesday to Saturday, a further 1,700 public procurement notices are published on TED.

The main challenge here is to develop an automatic method to utilise such a huge and useful resource of procurement data in order to provide an annotated data resource. For this purpose, the titles and CPV references of tender notices from TED are used as entries of the dictionary of B2B products. The titles of tenders in TED refer to the products or services that are the subject of the tender. Therefore, the titles are used as the potential B2B-Product Entities for making the initial dictionary. Within these titles, those who consist of a few words are mostly the name of the desired product or service in the relevant tender. CPV codes are used in tenders in order to refer to the category of desired product or service in the Common Procurement Vocabulary classification system. These code references also contain several expressions related to the topic of the tender. These expressions are used to enrich the dictionary of short titles with longer product names. Figure 5.13 shows the title and Figure 5.14 shows the CPV references of a Tender Notice from TED.

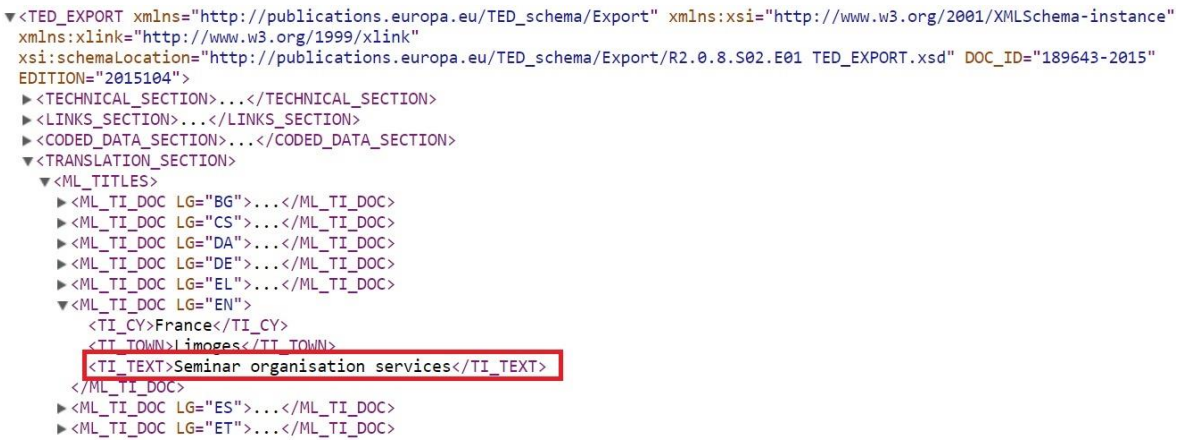

Figure 5.13 Title of a Tender Notice from TED

Therefore, the titles that consist of one word, two words or three words and all the CPV references are extracted from the tenders and considered as the primary dictionary of B2B-Product entities. Using this method after removing

duplicate entries, a primary dictionary including 776 one-word, 2507 two-word, 1396 three-word titles and 8691 CPV references has been made from 446419 tenders that had been published in TED in 2014.

```
<ORIGINAL_CPV CODE="30213000">Personal computers</ORIGINAL_CPV>
<ORIGINAL_CPV CODE="30213100">Portable computers</ORIGINAL_CPV>
<ORIGINAL_CPV CODE="30213200">Tablet computer</ORIGINAL_CPV>
<ORIGINAL_CPV CODE="30213300">Desktop computer</ORIGINAL_CPV>
<ORIGINAL_CPV CODE="30215000">Microcomputer hardware</ORIGINAL_CPV>
<ORIGINAL_CPV CODE="30230000">Computer-related equipment</ORIGINAL_CPV>
<ORIGINAL_CPV CODE="30231000">Computer screens and consoles</ORIGINAL_CPV>
<ORIGINAL_CPV CODE="30232000">Peripheral equipment</ORIGINAL_CPV>
<ORIGINAL_CPV CODE="30233000">Media storage and reader devices</ORIGINAL_CPV>
<ORIGINAL_CPV CODE="30234000">Storage media</ORIGINAL_CPV>
```

Figure 5.14 CPV references of a Tender Notice from TED

```
1  Roadworks   B2BProduct
2  Metalworking    B2BProduct
3  Parquet B2BProduct
4  Shelters    B2BProduct
5  Tents   B2BProduct
6  Wines   B2BProduct
7  Photocopiers    B2BProduct
8  Vaccines    B2BProduct
9  Wheat   B2BProduct
10 Lubricants  B2BProduct
11 Construction work   B2BProduct
12 Plasterboard works  B2BProduct
13 Road-repair works   B2BProduct
14 Lighting systems    B2BProduct
15 Photovoltaic cells  B2BProduct
16 X-ray devices   B2BProduct
17 Office supplies B2BProduct
18 Optical microscopes B2BProduct
19 Data-processing equipment   B2BProduct
20 Telecommunications services B2BProduct
21 Printing paper  B2BProduct
22 Site preparation work   B2BProduct
23 Wastewater pumping station  B2BProduct
24 Stadium construction work   B2BProduct
25 Medical software package    B2BProduct
```

Figure 5.15 B2B-Products Dictionary

Figure 5.15 illustrates a sample of the generated B2B-Products dictionary. The first column shows the product title and the second shows the type of the

named entity in the NER system. This file structure can be used as the dictionary for many dictionary-based NER tools (Llorens, Saquete, & Navarro-Colorado, 2013).

5.3.2.2 Annotating the data

In the next step, the primary dictionary was used as the seed list in a dictionary-based NER approach. The dictionary-based NER approach was applied to the tenders' descriptions in order to label product names in the tender descriptions using exact matching to the entries of the dictionary. Descriptions of the tenders convey more details about the desired products and services. From the Named Entity Extraction point of view, they have more words than the titles and comprised of sentences.

The Dictionary-based NER extracted the existing B2B-Product entities from the descriptions according to the primary knowledgebase gazette list and annotated them in the text. The output of this step is an annotated corpus which involves the B2B-Products annotated in product description sentences. Figure 5.16 shows a sample of the annotated corpus. In this file the Insurance service is marked as "B2BProduct" and the other parts of the sentence are tagged as "O".

```
1  Contract    O
2  to  O
3  provide O
4  Shropshire  O
5  Housing O
6  Limited O
7  's  O
8  Insurance   B-B2BProduct
9  services    I-B2BProduct
10 and O
11 related O
12 brokerage   O
13 requirements    O
14 for O
15 the O
16 next    O
17 3   O
18 years   O
19 .   O
```

Figure 5.16 Sample annotated Corpus

The result annotated description corpus is used as the training set for training a CRF model using Stanford NER tool[27]. The training set is provided in a standard format and can also be used in any other NER framework. Stanford NER is a Java implementation of a Named Entity Recognizer and provides a general implementation of CRF models. That is, by training a model, actually this code can be used to build sequence models for any task.

Since statistical NER systems such as CRF typically require a large amount of annotated training data (Khormuji, 2014), making the training set manually can be time-consuming, expensive and exhausting. Therefore, as mentioned, the annotated output of the dictionary-based NER system was fed to the CRF training process as the training set. In the same way, it can be fed to any other supervised machine-learning system that requires such a training set. This can provide an enormous training set for B2B product NER that is important in order to have a satisfying trained model without any manual efforts. The training set and its features for a NER system are the most important aspects of any NER system.

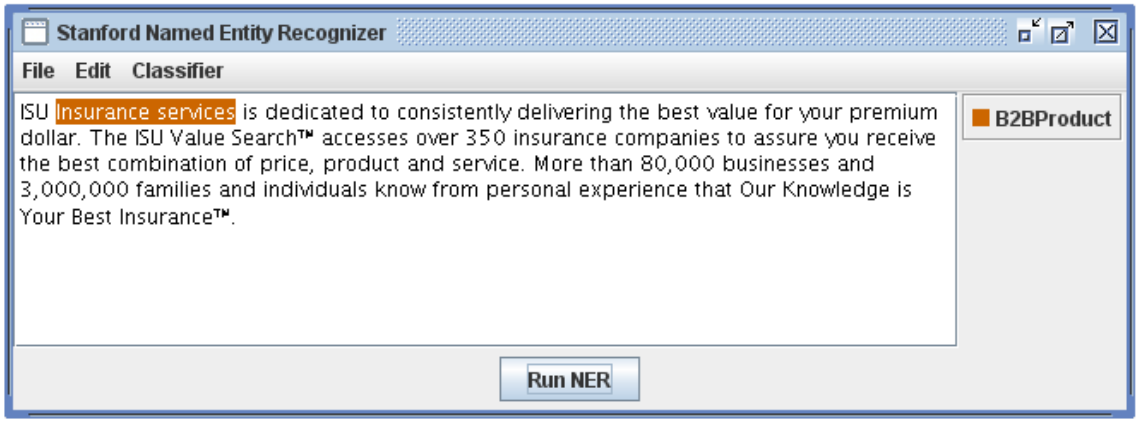

Figure 5.17 Named Entity Extraction from B2B context

[27] http://nlp.stanford.edu/software/CRF-NER.shtml

The last step is to use trained model for NER task in B2B context that is explained in details in next section. The trained model can be used in a NER tool (Figure 5.17) in order to extract the Named B2B-Product entities from e-Catalogues, e-Tenders and other B2B documents. The extracted entities not only include the products from the initial dictionary, but also newly discovered products as well as misspelled versions of the known products in the initial dictionary. Later, these entities can comprise the search elements of any semantic B2B product retrieval system.

5.4 Proposed Methodology Steps

The developed product matching method exploits the structures and semantics of product e-catalogues in term definition in order to use both syntactic and semantic concepts in calculating the similarity among products in B2B marketplaces. The e-catalogue matching engine that has been implemented using this method is capable of finding similar products from various types of e-catalogues.

The search engines utilize the semantic technologies to interpret the data elements and relationships among the data elements. It not only uses the structures, but also utilizes the semantic interpretation of the terms in data indexing and search processes.

The product classification systems and ontologies that are built based on (Stolz et al., 2014) are used in an iterative process to extract the semantic relationships among product data. Product classification systems are standard categorizations that are used for describing goods and services in e-procurement. These classifications can be used as semantic resources to describe the products mentioned in e-catalogues, tenders and other procurement documents as well.

The iterative process extends the term-vector of a product e-catalogue with synonyms, similar and semantically related terms of its terms. The process starts with determining existing entities in the e-catalogue. The procedure accepts an entity and an ontology as inputs and recursively returns all related entities to the given entity from the given ontology. Before that, the relevant on-

tology for the e-catalogue is selected based on its classification system. If no ontology can be found in the repository for the e-catalogue that can occur especially in the case of unstructured and unknown formats, the algorithm uses all available ontologies in the system and tries to recognize any available entity in the e-catalogue. The ontology repository of the system can be enriched by all available ontologies from various accessible resources such as (Hepp, 2008) and (Stolz et al. 2014).

In order to summarize the proposed approach for applying VSM to e-catalogue matching problem, all the steps are resumed as following:

1. Tokenization: extracting terms from e-catalogues using the NLP tool
2. Vectorization: making term-vectors
    a. Syntactical extension: Extending terms using name, value and location of attributes
        i. Representing the tree structure of the e-catalogues
        ii. Extracting locational values of the attributes
        iii. Excluding extra information and adjusting the importance of attribute
    b. Semantic extension: Extending terms using related concepts from a domain ontology
        i. Extending terms using their synonyms
        ii. Selecting relevant ontology
        iii. Recognizing existing entities
        iv. Iterative extraction of related entities
    c. Adding extended terms to the term-vectors
3. Calculating the similarity score

5.5 Summary

Based on many standards and data resource, e-catalogue search is affected by integration problems (Chen, Li, & Zhang, 2010). Since IR-based methods are applicable to wide range of structured and unstructured documents which we encounter in matching e-catalogues, this research work proposes a Vector Space Model approach to search e-catalogues. Furthermore, these methods target loosely structured data, thus useful and generally exploited for fast simple structured search and retrieval (Tekli, Chbeir, & Yetongnon, 2009).

In the proposed matching mechanism (Mehrbod, Zutshi, & Grilo, 2014a), VSM has been used to measure the similarity ratio of providers' e-catalogues with a buyer's e-catalogue. Combinations of values and attributes of structured e-catalogues have been used to find the correlation of documents based on the relationship of their common items. In order to associate the structures in calculating the similarity score, levels of attributes in XML documents are included in the search element definition. Then a simple table of coefficients has been used to specify the matching model for standard e-catalogues. This mechanism increases the search precision by removing unrelated information from the matching process and boosting weights of important tags.

The proposed solution exploits the structures as much as their details are known whilst it is independent of the structures. Structure independency provides the ability to match unstructured information as well as unknown structures. Although the structure can be discarded (Lee et al., 2007) in order to provide search indexes for matching e-catalogues regardless of the structures, the information existing in structures is valuable and can be helpful in matching process.

This syntactic matching approach has been extended using a semantic mechanism (Mehrbod et al., 2015) which is not dependent on the underlying ontologies and schemas. While the syntactic method uses a combination of values, names and location of attributes of structured information to find the syntactic correlation of e-catalogues, the semantic method uses domain ontologies to expand the matching mechanism with existing semantic relationships among data attributes. In this process vectors of each e-catalogue were enriched with

semantic concepts that exist in the e-catalogue. Adding semantic relationships to the terms of the vectors enables the matching process to find semantically similar e-catalogues. Combining semantic queries with information retrieval techniques makes it possible to use the benefits of all available ontologies and schemas but not to be dependent on them.

An entity recogniser has been trained to help the semantic matching mechanism in recognizing the products mentioned in e-catalogues. Such product mentions constitute the existing semantic concepts in the search domain. Learning approaches require large amounts of annotated training data in order to be effective. Since making such annotated data needs manual efforts and should be repeated manually for any new data resource, we develop an automatic self-training mechanism to train the required model for the learning-based NER approach.

The goal is to use these already known product mentions and their references as the training data to train the model and then use the trained model to recognize the product mentions from unknown e-catalogue structures.

6

Validation

In the previous chapter, a multilayer solution has been proposed to solve different aspects of the e-catalogue matching problem. In this chapter, the proposed mechanism will be tested in order to evaluate its capabilities in different possible search scenarios in a B2B e-marketplace. Four test cases as summarised in Table 6.1 will be applied to the proposed matching mechanism.

The first test scenario (Supplier finder) represent the capability of the proposed matching process in matching diverse structures and semantics of catalogues from various resources.

The second test scenario (Opportunity Finder) evaluated the ability of the e-catalogue matching engine for matching synonym queries in the tender search process. The tender search is one of the most common search scenarios in B2B e-marketplaces. The goal of the test is to improve search performance and to help the suppliers in finding more relevant opportunities.

The third test scenario also applied and tested the developed product matching engine in finding tenders from public procurement resources. But this test evaluated the capability of the matching mechanism to use available semantic data and tolerating absence of semantic information as well.

The B2B-Product NER test addresses the issue of extracting meaningful sequence of words from documents in e-procurement domain. Extracting

meaningful sequences of the words from text is an important task of information extraction called NER. In the case of e-Procurement domain, these meaningful sequences are the products, works or services that are mentioned in a tender or e-catalogue.

Table 6.1 Overview of the test cases

Test	Method	Target
Supplier Finder	Information Retrieval	Test the capability of the matching mechanism in solving semantic and syntactic heterogeneity
Opportunity finder	Information Retrieval	Test the capability of the matching mechanism in improving opportunity search and encountering synonym queries
Multi-Resource Matching	Information Retrieval	Test the capability of the matching mechanism in using available information and tolerating lack of information
B2B-Product NER	Information Extraction	Supporting the matching engine in recognizing the searchable items

6.1 Evaluation Measures

In information retrieval, precision and recall are the two most used metrics for performance measurement. The performance of search engines and matching mechanisms is shown and compared using these two factors and their combinations.

Recall is the ratio of the number of correct answers (relevant documents) that are retrieved, to the number of all correct answers (All relevant documents that exist in the search repository). Therefore, the Recall factor shows the ability of a search engine to retrieve more correct answers from the search area. For example, if there are 10 documents related to a desired query in the reposi-

tory and the search mechanism retrieves 7 of them, the recall measure will be 0.7.

$$R = \frac{\#\,\text{returned correct answers}}{\#\,\text{All correct answers}} \tag{6.1}$$

Precision is the ratio of correct answers (related documents) in all retrieved result set. For example, if the search method returns 10 documents (related and non-related) for a query, but only 7 of them are related to the query, precision is 0.7. Therefore, the Precision factor shows the accuracy of a search engine to retrieve less false answers from the search area. While recall shows the percentage of relevant results that the matching method is able to return, precision shows the accuracy of the method in returning more relevant results than irrelevant ones.

$$P = \frac{\#\,\text{returned correct answers}}{\#\,\text{All returned answers}} \tag{6.2}$$

In general, recall and precision vary inversely. Increasing the precision rise the risk of reducing the recall, because some relevant documents may be erroneously rejected. Conversely, increasing the number of recalled documents increases the risk of many non-relevant documents being returned. Finding a balance between these two metrics is dependent on the mission of the search engine.

Generally, the overall performance of a search engine is shown using Precision-Recall curves that represent these two inversely related metrics (Mehrbod et al., 2015). The performance curves can be used to compare different search methods.

While the common way of evaluating results of Information Retrieval systems is using Recall and Precision curves, for evaluating Machine Learning experiments usually a combination of these two parameters called F_1-score measure is used. F_1-score is a measure that considers both precision and recall. The measure can be interpreted as a weighted average of the precision and recall, where an F_1-score reaches its best value at 1 and worst at 0 (Powers, 2011)(Mehrbod et al., 2015)(Llorens, Saquete, & Navarro-Colorado, 2013).

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{6.3}$$

Precision, recall and F_1-score are defined based on the retrieved results set and correct results set. In other words, these measures are calculated on an un-ordered set of items that there is no difference between the set members. This feature makes them suitable for measuring Boolean search results where an item either belongs to the correct answer set or not.

In order to apply the measures to a ranked information retrieval technique, the calculations should be done for the sets of top *k* items in the retrieved results. In a ranked result set, the retrieved items are sorted by the similarity ratio to the search query and naturally the top list of the result set is considered as similar enough to the query. Usually a similarity threshold which specifies *k* or a direct approximation of *k* is considered based on the search domain and the magnitude of the search repository (Manning, Raghavan, & Schutze, 2008). For example for ranked search result sets of size about 30 items, the precion and recall values might be calculated for k_5, k_{10}, k_{15} and k_{20}.

Calculating the precisoin for Top k_x result items, gives us the precision of the search method for a specific recall point. For example, k_5 may show the precision of search engine se_1 for recal point r_1 and precision of search engine se_2 for recal point r_2 for the same query where r_1 is not necessarily equal to r_2. Therefore, prececions of se_1 and se_2 are not comparable on k=5. Furehtermoe, precision and recall values for top k_x for a query may not be defined if there are no correct answers among top k_x retrived answers.

Hence, in order to compare the performance of search engines the precesions should be calculated on standard recall levels (0 to 1 in increments of 0.1). The precision on each standard recall level can be calculated using interpolation. The interpolated precision value for recall level *r* is the highest measured precision value over all recalls greater than *r*:

$$p_{\text{int}}(r) = \max_{r' \geq r} p(r') \tag{6.4}$$

In order to make a fare compration, the interpolated precisions should be calculated for all queries available in the test collection. Hence, for each recall

level, the arithmetic mean of the interpolated precisions is considerd as the precision at that recall level.

The outcome will be a 11-point Interpolated Average Precision-Recall table that is usually shown as a curve. Each search mechanism in the comparation can be shown as a curve on the average interpolated precision-recall graph. In such graph that the curve to the top right shows better search performance, the search mechanisms can be easily compared.

Another measure that have become more common and accepted as the most standard measure is *Mean Average Precision* (MAP). MAP provides a single figure measure of quality across all recall levels. Among evaluation measures, MAP has been shown to have especially good discrimination and stability (Manning, Raghavan, & Schutze, 2008).

For a query, q_j, Average Precision is the average of the precision values obtained for the set of top *k* returned results after each relevant document is retrieved. That is, the precision is calculated for all *k* values that a new correct answer is retrieved in the results. The average of all these precessions comprises the Average Precision for query q_j.

The Average Precisions should be calculated for all queries $q_j \in Q$ in the test set. These values are then averaged for all queries to obtain Mean Average Precision as a single measure of the performance as:

$$AP(q_j) = \frac{\sum_{k=1}^{m_j} \text{Precision}(R_{jk})}{m_j} \tag{6.5}$$

$$MAP(Q) = \frac{\sum_{j=1}^{|Q|} AP(q_j)}{|Q|} \tag{6.6}$$

Where the set of the relevant documents (correct answers) for the query q_j is $\{d_1, \ldots, d_{m_j}\}$ and R_{jk} is the set of ranked retrieval results from the top results until you get to the document d_k. In other words, for each relevant document to

the query, k is the rank of the document in the retrieved results and $\text{Precision}(R_{jk})$ is the precision for top k retrieved documents.

When a relevant document is not retrieved at all, the precision value in the above equation is taken to be 0. For a single query, the average precision approximates the area under the precision-recall curve. Therefore, the MAP is roughly the average area under the precision-recall curve for a set of queries. Using MAP, fixed recall levels are not chosen, and there is no interpolation. The MAP value for a test collection is the arithmetic mean of average precision values for individual information needs (Manning, Raghavan, & Schutze, 2008).

6.2 Supplier Finder

6.2.1 Test Scenario

In order to test the capability of the e-catalogue matching mechanism in solving heterogeneity problem, a supplier finder service has been implemented in the e-procurement platform of Vortal. VortalNext is a large global e-procurement platform that allows buying organizations to purchase goods and services cheaper and more efficiently and gives suppliers access to a greater number of sales and revenue opportunities.

Vortal[28] is a leading Portuguese G2B2B (Government to Business to Business) e-sourcing an e-procurement operator and the third largest e-marketplace in Europe. Vortal currently has over 60,000 companies and 2000 contracting authorities connected to its platform, covering the markets of Public e-Tendering, AEC, Health, Energy & Utilities, and Industry & Office Supplies, mainly in Portugal, but now serving companies and contracting authorities in 3 continents. More 25 billion Euros have been awarded on its platforms.

[28] en.vortal.biz

The supplier finder service is developed as a part of an industrial research project called VortalSocialApps that provides a B2B social network for companies within the VortalNext e-procurement platform. The social network makes it possible for the companies to share their business information and e-catalogues directly from their local e-procurement systems in the platform. One of the major services that are provided by the platform to the users is the capability to search and find a company in the B2B social network. This service which is called "Supplier Finder" provides company search and e-catalogue matching facilities to the users and helps them to find a company using its business information. The e-catalogue matching mechanism that provides a framework for matching various e-catalogues originating from suppliers and buyers in the e-procurement platform helps companies to find partners and opportunities in the network. The service can be used to find companies using their profile data or/and their product e-catalogues.

Using this service, users are able to search within the e-catalogues that are uploaded by various suppliers to the B2B social network. Users can use an e-catalogue as the search query and simply upload their e-catalogues to find similar documents. This search mechanism allows users who prefer to specify the tag relations while searching (Tekli, Chbeir, & Yetongnon, 2009) to get rid of using content-and-structure queries (Carmel et al., 2002). Generally, two types of queries are possible to search within the structured data. The queries with structural constraints called content-and-structure, and those without constraints called content-only. In the other words, users can use e-catalogues as structured search queries but they don't have to use a specific structured query language to communicate with the search engine.

6.2.2 Data Gathering

In order to evaluate the matching performance, a set of product e-catalogues files in the following three general cases have been inserted into the e-catalogue repository.

- First, unstructured text such as PDF files which are common in online commerce.

- Second, structured or semi-structured e-catalogues which are unknown for the system such as enterprise-specific formats.
- Third, structured standard documents which are known for the system such as cXML and UBL e-catalogues.

The variety of document structures and schemas that have been used to make the search repository guarantees to model the diversity of e-catalogues used by various suppliers and buyers in the platform.

6.2.3 Test Definition

Next step is to calculate the semantic similarity between the search query and documents. The similarity between catalogues is measured based on the similarity between the products contained in each of them. In order to apply and evaluate the e-catalogue matching mechanism in supplier finder search scenario, the test that is shown in Figure 6.1, has been designed and implemented. Having made the term vectors for all e-catalogues in the repository, similar e-catalogues to a sample e-catalogue in the platform were searched. The matching ratio for the following states of the matching mechanism are shown in Figure 6.2:

- State 1: The basic functionality of Vector Space Model.
- State 2: The previously proposed syntactic e-catalogue matching mechanism.
- State 3: The proposed masking mechanism for standard e-catalogues.
- State 4: The proposed semantic extension for e-catalogue matching mechanism.

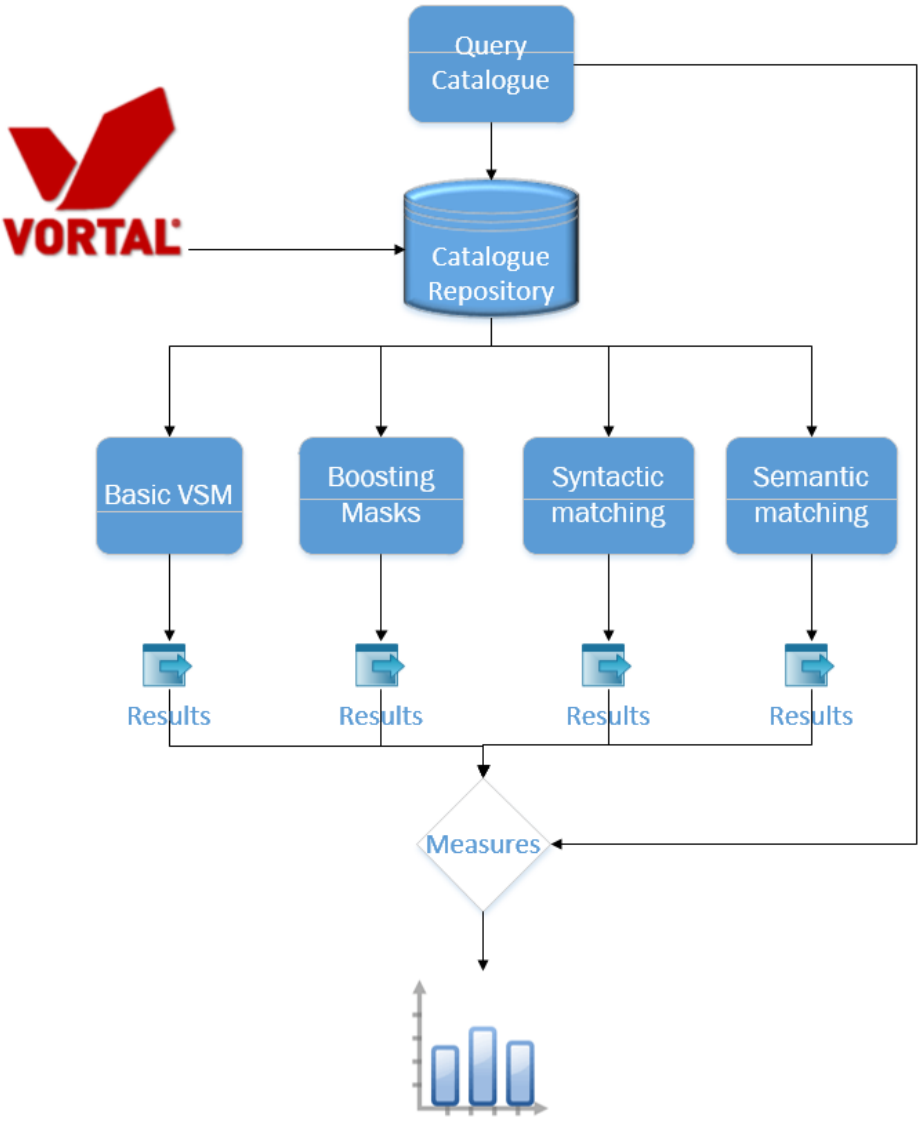

Figure 6.1 Supplier Finder Test

6.2.4 Test Results

Figure 6.2 shows the normalized score results for e-catalogue matching. The figure shows the result of using an e-catalogue as the search query and then searching for similar e-catalogues that have been provided to the platform by the suppliers. A higher score indicates a higher similarity ratio that boosts the position of an item amongst the search results. Thus, increased scores for related e-catalogues shows improvement in matching performance. The vertical axis shows the percentage of similarity ratio of each retrieved e-catalogue to the search query. Similarity ratios are calculated using four different mentioned

states. Although the number of retrieved documents are more, for the purpose of comparison, the horizontal axis only shows three retrieved e-catalogues as representative of the three mentioned groups of the e-catalogues in the repository. Document D1 is an unstructured text e-catalogue that contains associated information with the desired parts in the search query. Document D2 and D3 are respectively a structured XML e-catalogue and a standard UBL e-catalogue that contain the same information as D1. Therefore, the containing information of all documents (D1, D2 and D3) are same, but they are structured in different ways.

Using the basic functionality of VSM, the matching scores are lower for structured e-catalogues than unstructured documents, because structured documents have some extra data such as XML tags that represent their structures. But for the basic functionality, there is no difference between data and its structure. Therefore, this metadata that is considered as data by the basic method results in lower matching scores. Furthermore, in this state, there is no significant difference between standard structures and unknown structured formats for the system.

Using the proposed syntactic e-catalogue matching mechanism, this metadata results in better matching and therefore the similarity scores increase for the structured documents. The extra information in the structured documents has been used to make more terms for related vectors resulting in an increase in the matching scores. Calculated similarity score for the unstructured e-catalogue is equal to the calculated similarity score using the basic functionality of VSM. Obviously, there is no structure in the unstructured document to help the matching process. The increase in matching score is lower for the standard structure than the non-standard structured document. This anomaly is the result of extra information such as addresses and contacts in the standard e-catalogue that is not related to the product data.

The proposed masking mechanism for standard e-catalogues solves the matching score anomaly for standard formats by eliminating unrelated data from similarity calculation method. As it can be seen in Figure 6.2 this approach increases the matching score for the standard e-catalogue. Obviously, the mask-

ing mechanism doesn't have any effect on similarity ratio for unknown structures.

While the proposed syntactic matching mechanism only increases the similarity ratio for structured e-catalogues, the semantic matching mechanism improves the similarity ratio for all types of documents. Since semantically related terms can be detected in this approach, related e-catalogues to the search query have more matched terms with the query that lead to higher similarity scores.

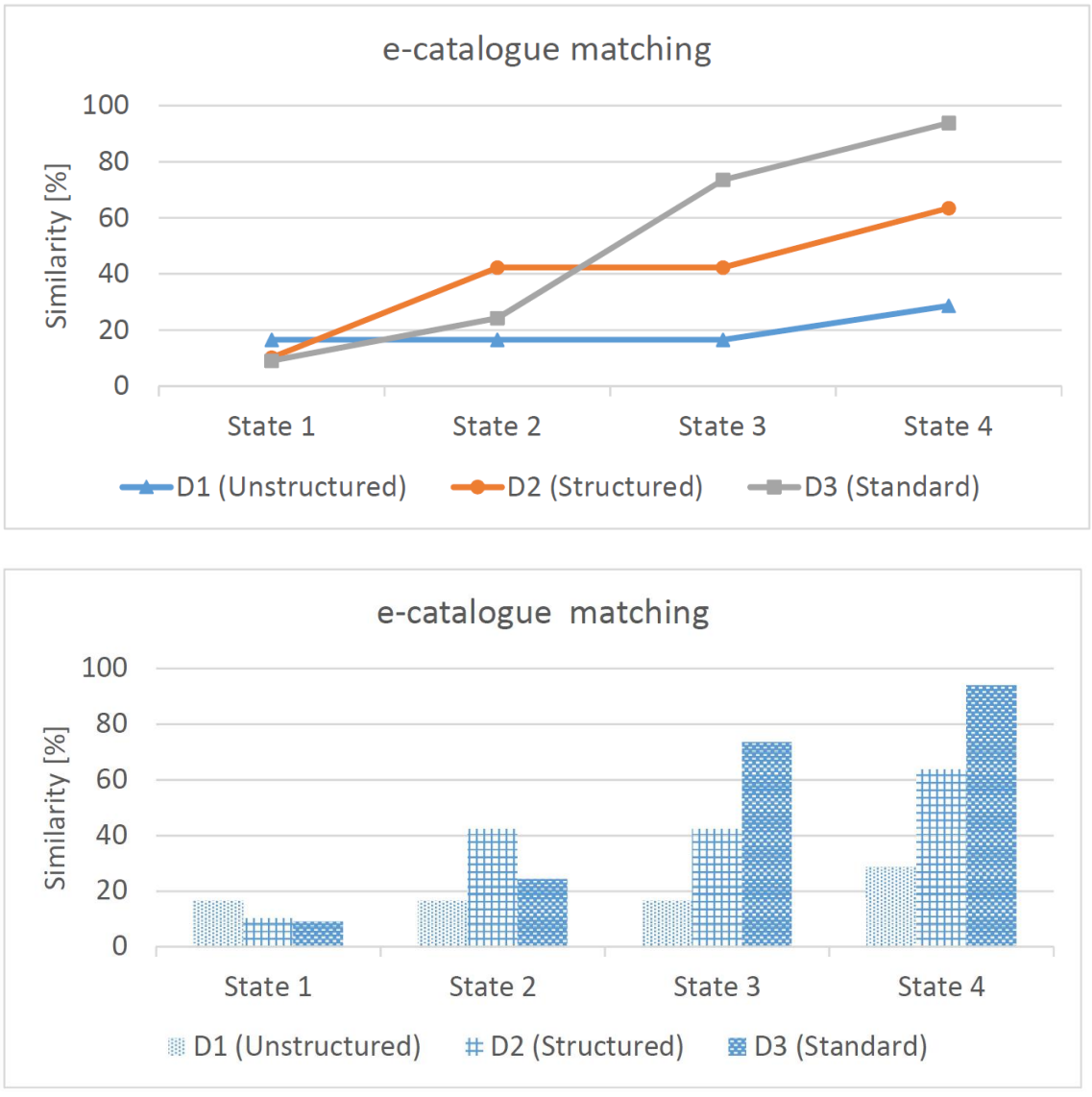

Figure 6.2 E-catalogue matching scores

The proposed matching mechanism not only improves the ranking result by increasing the similarity scores for related documents, it can also return more documents that is another important measure in search evaluation.

6.3 Opportunity Finder

6.3.1 Test Scenario

Public procurement, also called public or open tendering, is the purchase of goods, works or services by a public authority, such as a government agency. Open tendering opens up sufficient and fair competition between suppliers and ensures that public contracts are awarded fairly, transparently and without discrimination. This not only helps to achieve benefits such as increased efficiency and cost savings, but also can improve transparency in order to reduce corruption in public procurement services.

While transparency is one of the important factors in the efficiency of a public procurement system, the usage of e-procurement is another important efficiency factor (Molander, 2014)(Miyamoto, 2015). E-procurement digitalizes the important aspects of the procurement process, such as search, selection, communication, bidding or awarding of contracts; with a specific emphasis on efficiency, transparency and policy in the public sector (Roman, 2013).

Public e-procurement platforms allow reaching these objectives through a web-based open tendering e-marketplace. Since e-Tendering Marketplaces can be accessed anywhere globally, they can have a great improvement on the accessibility and transparency of tenders and provide equal opportunities to all suppliers (Grilo, Jardim-Goncalves & Ghimire, 2013).

In public procurement e-marketplaces, suppliers search published tender notices in order to find available business opportunities that match their products. Matching goods and services provided by a supplier with similar tender calls published by contracting authorities can be a sufficient way to find suitable business opportunities in B2B e-marketplaces. However, many tendering websites and marketplaces provide only simple keyword-based search. The

main drawback of such exact keyword-based search mechanisms in product search is their problem in detecting semantically similar products (S.-L. Huang & Lin, 2010). Applying semantic technologies to product search mechanisms of e-tendering e-marketplaces can help suppliers to find similar and semantically related tenders to their product e-catalogues.

In procurement industry, companies usually exchange their product information in the form of product e-catalogues. E-catalogues are used by suppliers to describe goods or services offered for sale and may be used by buyers to source goods or services, or to obtain product or pricing details. This product information can be used by a product search mechanism in order to find and recommend similar product requests (Julashokri et al., 2011). In this sense, a supplier delivers his product e-catalogue to the search engine and receives applicable tender calls sorted based on the similarity ratio to his productions.

This test evaluates the e-catalogue matching mechanism in public tender search. The previous test evaluated, the e-catalogue matching engine to find business partners by measuring the similarity ratio of providers' e-catalogues with buyer's e-catalogues. This new test applies and evaluates it in the tender search process using tenders published in a public tendering website, called Tenders Electronic Daily, in order to improve opportunity search service. TED[29] is the online version of the 'Supplement to the Official Journal' of the EU, dedicated to European public procurements. According to the EU rules on public procurements, information of public procurement contracts and notices published in the EU Member States, European Economic Area (EEA) can be accessed openly on TED.

6.3.2 Test definition

In existing tendering e-marketplaces, suppliers use keyword-based search engines to find products that match their conditions. The keyword-based search may have low precision especially when the users use synonyms for searching

[29] ted.europa.eu

for a product. Furthermore, theses search mechanisms cannot find potentially interesting products for the users that don't match their conditions exactly (S.-L. Huang & Lin, 2010).

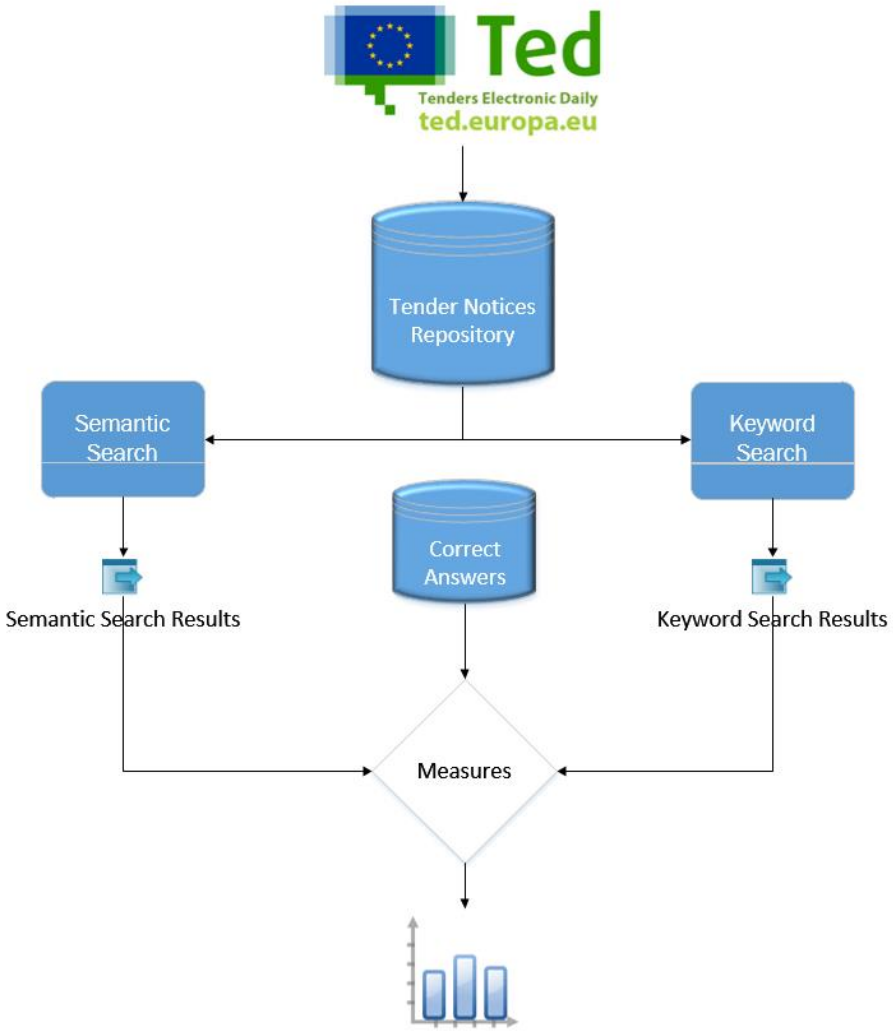

Figure 6.3 Tender search test

Semantic product search engines aim to encounter these problems by improving search capabilities using semantic web technologies in information retrieval process. Since the mentioned e-catalogue matching engine achieved good results in finding semantically similar products from e-catalogues, this test surveys the application of the engine for finding business opportunities in public tenders.

In order to apply and evaluate the e-catalogue matching mechanism to a tender search scenario, the following test has been designed and implemented. As it can be seen in Figure 6.3, the test evaluates the improvement of tender search performance using the e-catalogue semantic matching mechanics.

6.3.3 Data Gathering

The tender notices that have been published on the TED website are used as the tender repository in order to search for business opportunities. Tenders Electronic Daily (TED) is the online version of the "Supplement to the Official Journal" of the EU that is dedicated to European public procurement. TED provides free access to business opportunities from the European Union, the European Economic Area and beyond. Ted is updated five times a week, from Tuesday to Saturday, with more than 1,700 public procurement notices from the European Union, the European Economic Area and beyond. These procurement notices can be browsed, searched and sorted by country, region and business sector. All published tenders since January 2011 have been archived in TED that can be used as an enormous resource of half-million procurement notices per year for research purposes.

TED allows the users to search using a number of methods, including Business Opportunities, Business Sector (by CPV code), Place of delivery (by NUTS code) and Heading. Business Opportunities search is a structured search that the user can select interested country or countries for supplying and define the type of notice that is looking for. The notice types include contract (tender) notice, Design contest, prior information notice and Qualification system with call for competition. The search will return absolutely everything that's been published in the last issue and can be refined based on publication date, deadline and so on.

Business Sector search categorizes the notices using CPV codes that can help the user to narrow down the search results to the types of desired goods and services. Each code relates to an overarching category, with subcategories each having their own code. This allows the user to either search for a main category, e.g. 'construction', which would deliver many results and may be too

broad and time-consuming to trawl through, or search more specific underlying layers which sit under the main category, e.g. 'Sanitary fixture installation'.

Place of Delivery search uses NUTS codes (Nomenclature of Territorial Units for Statistics) in order to allow the user to filter search results based on country or region which is interested in tendering for. This is useful for suppliers that generally provide services locally and want to avoid results from other places. Heading provides the option to search by a type of authority, e.g. by European Economic Area, European Commission, Government Procurement Agreement, international institutions, Member State, agencies and so on.

Since the website provides the category-based search, the tenders from each business sector can be retrieved separately. In order to make a sample test repository, tenders from three different sectors including "Cable, wire and related products"[30], "Insurance"[31] and "Mobile telephones"[32] are collected. All tenders from the three mentioned categories that have been published in 2015 in the UK are collected and saved in the test repository. The test repository contains 28 tenders of "Mobile telephones" category, 107 tenders of "Cable, wire and related products" category and 550 tenders of "Insurance" category. Each category contains the tenders that are considered as the correct answers for the related search. Therefore, a search mechanism will be considered as 100% precise, if retrieves all the tenders from a category in response to a search query for the topic.

This small sample test set is used in a forthcoming demonstrative test in order to show the performance of the e-catalogue matching mechanism in a simple example. In the following, a large test set will be used to evaluate the performance of the matching engine in a fare test using more queries.

[30] CPV code: 44300000.

[31] CPV code: 66510000.

[32] CPV code: 32250000.

Table 6.2 TED Test repository based on main activities

Business section	Main Activity code	Number of Tenders
Housing and community amenities	A	2422
Social protection	B	321
Recreation, culture and religion	C	598
Defence	D	1073
Environment	E	574
Economic and financial affairs	F	496
Production, transport and distribution of gas and heat	G	345
Health	H	5172
Airport-related activities	I	240
Port-related activities / Maritime or inland waterways	K	172
Education	L	4334
Exploration and extraction of coal and other solid fuels	M	14
Electricity	N	571
Postal services	P	89
Railway services	R	380
Urban railway/light rail, metro, tramway, trolleybus or bus services	T	288
Public order and safety	U	732
Water	W	543
Not specified/Other	Z/8/9	7333

In order to make the comprehensive test set all tenders that have been published in English in 2015 on TED have been collected and categorized in 21 main business sectors. All procurement documents in TED can be retrived using the main activity criteria, such as education, health, housing and community amenities, etc. TED website allows the registered users to download XML packages including monthly records of all published notices. Having downloaded all notices published on 2015, we assorted the tenders based on their main activity into 21 groups as is shown in Table 6.2.

6.3.4 Test Results

Two different search mechanism including keyword search and semantic search have been tested and compared on the data repository. The keyword search is considered as the basic approach of searching for opportunities in open tendering websites. The e-catalogue matching mechanism that is explained in the previous section is used as the semantic tender search. The results of these two mechanisms are compared with the correct answers that are gathered using category-based search service of the TED in order to calculate search performance measures. In simple words, the results of the semantic matching mechanism are compared with the results of simple keyword search (as the basic search algorithm) for searching on the gathered tender repository.

As mentioned, Precision and Recall are two common metrics for measuring the performance of search engines. As an example, the collected repository contains 550 tenders in Insurance sector and if the search mechanism retrieves 400 of them in response to "insurance" search query, the recall measure will be 400/550. In the search repository, there are 135 tenders from two other sections that are not related to Insurance and considered as false answers. Hence, if the search engine returns 450 documents (400 correct and 50 wrong answers), precision is 400/450.

Table 6.3 shows the precision and recall calculated for both mentioned search mechanisms. The results are reported for three different synonym search queries including 'mobile', 'phone' and 'gsm telephones' on the repository for finding tenders from "Mobile telephones" section. All available tenders in this

section are considered as correct answers and the tenders from the other sections are considered as false answers.

Table 6.3 average interpolated precision-recall values

	Keyword-based search				E-catalogue matching engine			
Recall	Precision (phone)	Precision (gsm telephones)	Precision (mobile)	Average Precision	Precision (phone)	Precision (gsm telephones)	Precision (mobile)	Average Precision
.1	100	100	100	100	75	60	75	70
.2	0	0	100	33.33	71.42	35.71	83.33	63.48
.3			87.5	29.16	77.77	35.71	70	61.16
.4			90.9	30.3	55.55	38.46	76.92	56.97
.5			92.3	30.76	27.90	27.9	75	43.6
.6			87.5	29.16	22.58	22.58	77.77	40.97
.7			89.47	29.82	16.83	16.83	68	33.88
.8			0	0	0	16.37	18.44	17.40
.9				0		12.35	17.79	15.07
1.0				0		0	0	0

Since in practice we can calculate the precision in specific recall points, it is not easy to compare the results of different search mechanisms. In order to have comparable results, the precisions should report for standard recall levels. Thus, the averages of all three queries are used to plot the interpolated standard precision-recall curve in Figure 6.4.

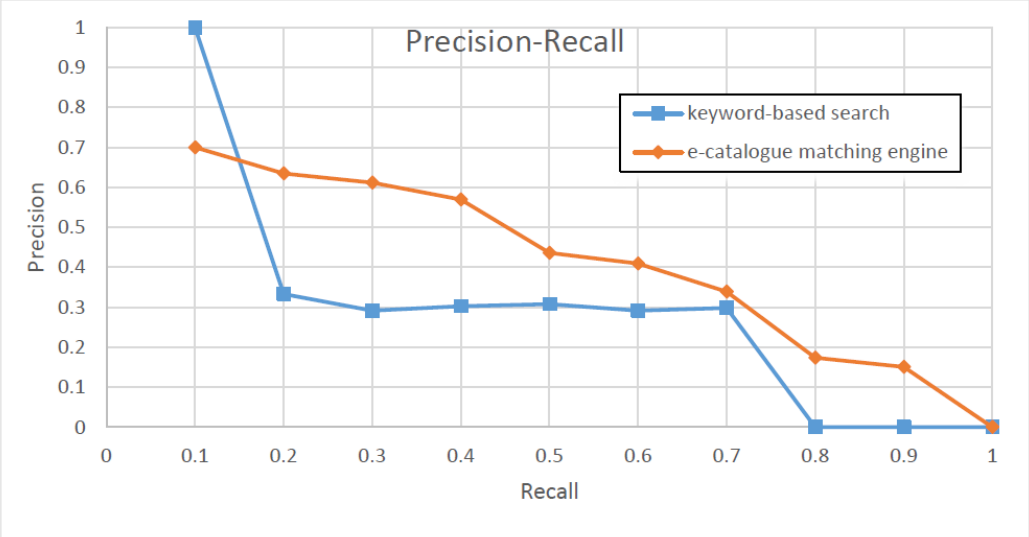

Figure 6.4 Precision-Recall curve for the sample test set

Each curve represents average interpolated precision-recall values for one search mechanism. In such curve, the one to the top right shows better search performance. As it can be seen in the figure, the semantic e-catalogue matching engine can improve tender search on a tendering website and consequently can help supplier in finding business opportunities. Even though the keyword-based search gets higher precision when using an exactly same keyword as exists in the data source, it has very low recall, especially when searching using synonym keywords.

As mentioned, after this demonstrative test, a comprehensive test also is done on the collected full test repository of 2015. In order to calculate the performances measures based on an extensive query set, the test has been repeated for all tenders available in the test repository as query for both semantic and keyword-based search mechanisms. I.e. catalogue files made using the data of each tender call used as the query to find similar calls from the relevant business section form the test repository and the test is redone for each query. The calculated Precision-Recall values for each query are interpolated to the standard Precision-Recall points in order to be comparable. The averages of all Recall and Precisions value are reported in Table 6.4 as the performance of the matching engine compared to the base matching mechanism in the relevant business

sector. Finally, the mean value for all the business sections is used as an indicator factor which is illustrated on Figure 6.5.

Table 6.4 Precision-Recall for comprehensive test set

Recall		.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
Keyword-based search	A	1	.690	.302	.302	.057	.026	.026	.010	0	0
	B	1	.787	.703	.387	.212	.091	.082	.037	.004	.004
	C	.870	.649	.525	.505	.429	.393	.217	.139	.044	.007
	D	.474	.337	.289	.258	.208	.174	.147	.108	.075	0
	E	.644	.288	.163	.155	.108	.109	.083	.068	.051	.021
	F	1	.836	.836	.533	.533	.062	.051	.051	.008	.008
	G	.461	.392	.350	.252	.215	.193	.116	.051	.034	.002
	H	.526	.323	.234	.052	.050	.014	.014	.005	0	0
	I	.518	.284	.144	.148	.116	.086	.063	.043	.021	.019
	K	.934	.636	.515	.413	.383	.352	.266	.235	.033	.033
	L	1	.625	.460	.358	.363	.195	.158	.047	.011	0
	M	1	1	.503	.503	.010	.010	0	0	0	0
	N	.761	.694	.616	.369	.248	.159	.103	.083	.060	.036
	P	.681	.478	.402	.317	.247	.198	.164	.125	.090	.069
	R	1	.703	.493	.414	.330	.287	.224	.154	.057	.028
	T	.789	.685	.598	.498	.359	.208	.125	.079	0.21	.005
	U	.727	.607	.543	.499	.296	.089	.078	.038	.035	.024
	W	.635	.574	.505	.422	.337	.293	.167	.122	.069	.039

Table 6.4 (continued) Precision-Recall for comprehensive test set

Recall		.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
E-catalogue matching engine	A	1	1	1	.711	.711	.059	.059	.059	.059	.059
	B	1	1	.765	.710	.617	.595	.526	.473	.449	.148
	C	.866	.886	.796	.796	.784	.786	.793	.798	.422	.128
	D	1	.776	.518	.499	.389	.358	.296	.136	.063	.039
	E	.739	.654	.565	.309	.163	.136	.093	.089	.065	.043
	F	1	.844	.844	.838	.838	.829	.774	.774	.662	.662
	G	.878	.80	.716	.552	.401	.198	.186	.133	.097	.042
	H	.530	.437	.384	.294	.231	.202	.149	.072	.049	.011
	I	.733	.701	.611	.521	.253	0.86	.081	.072	.033	.029
	K	1	.842	.725	.725	.198	.162	.162	.068	.009	.009
	L	1	1	.933	.933	.370	.370	.058	.058	.066	.066
	M	1	1	.823	.823	.483	.483	.474	.474	.066	.066
	N	1	1	.804	.804	.400	.400	.194	.194	.027	.027
	P	.855	.806	.692	.646	.529	.342	.147	.095	.036	.006
	R	.801	.741	.734	.730	.727	.684	.672	.619	.328	.078
	T	1	.665	.490	.504	.474	.393	.271	.072	.021	0
	U	.611	.522	.487	.449	.403	.335	.266	.170	.095	.004
	W	.893	.804	.735	.681	.634	.602	.506	.422	.347	.264

Table 6.4 (continued) Precision-Recall for comprehensive test set

Recall		.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
E-catalogue matching engine using NER	A	1	.952	.928	.928	.901	.904	.904	.890	.706	.706
	B	1	.899	.860	.786	.795	.754	.675	.451	.369	.080
	C	1	.962	.944	.911	.857	.756	.708	.638	.479	.479
	D	1	.796	.690	.380	.293	.150	.131	.087	.013	.013
	E	.752	.750	.732	.652	.549	.461	.280	.138	.082	.022
	F	1	.863	.863	.875	.875	.884	.865	.865	.748	.748
	G	.979	.744	.536	.415	.399	.376	.274	.230	.042	.042
	H	1	1	.502	.502	.502	.012	.012	0	0	0
	I	.879	.629	.540	.531	.450	.437	.250	.169	.051	.012
	K	1	.791	.728	.571	.571	.290	.194	.127	.100	.100
	L	1	1	1	1	.388	.388	.116	.116	.134	.134
	M	1	1	.933	.933	.870	.870	.893	.893	.517	.517
	N	1	.747	.638	.683	.646	.427	.427	.167	.003	.003
	P	.899	.755	.683	.655	.545	.464	.341	.271	.145	.049
	R	1	.873	.874	.874	.750	.756	.756	.561	.175	.175
	T	1	.758	.758	.535	.535	.114	.109	.109	.013	.013
	U	.757	.724	.626	.619	.547	.538	.513	.463	.208	.110
	W	1	.866	.823	.749	.548	.508	.460	.428	.363	.208

Knowing that in a Precision-Recall diagram the curve on the top-right shows the better performance, it can be seen in the Figure 6.5 that the e-catalogue matching mechanism provides better results compare to the key-word0based matching. But, in order to have a single-figure measure of quality for comparing the performance of the proposed e-catalogue matching method with the basic matching mechanism easily, MAP values are reported in Table

6.5. The MAP values are calculated using the average precisions of all queries across all recall levels on each business sector separately and then average of the MAPs for all business sectors together is reported at the end of the table.

Table 6.5 MAP Values

Business section	**MAP (Keyword-based search)**	**MAP (E-catalogue matching engine)**	**MAP (E-catalogue matching engine using NER)**
A	29.81	59.02	89.77
B	36.73	62.85	66.75
C	12.58	73.27	80.65
D	40.67	40.79	39.36
E	24.49	34.31	50.36
F	41.53	82.46	87.27
G	26.57	45.58	45.14
H	18.87	29.37	37.87
I	20.32	37.50	42.13
K	43.24	42.96	47.55
L	32.48	48.56	52.79
M	37.85	56.96	84.28
N	38.17	48.53	51.86
P	31.54	46.71	52.28
R	36.93	65.98	71.28
T	38.85	38.92	42.18
U	35.82	37.72	55.31
W	35.71	62.51	61.68
AVG	**32.99**	**50.78**	**58.81**

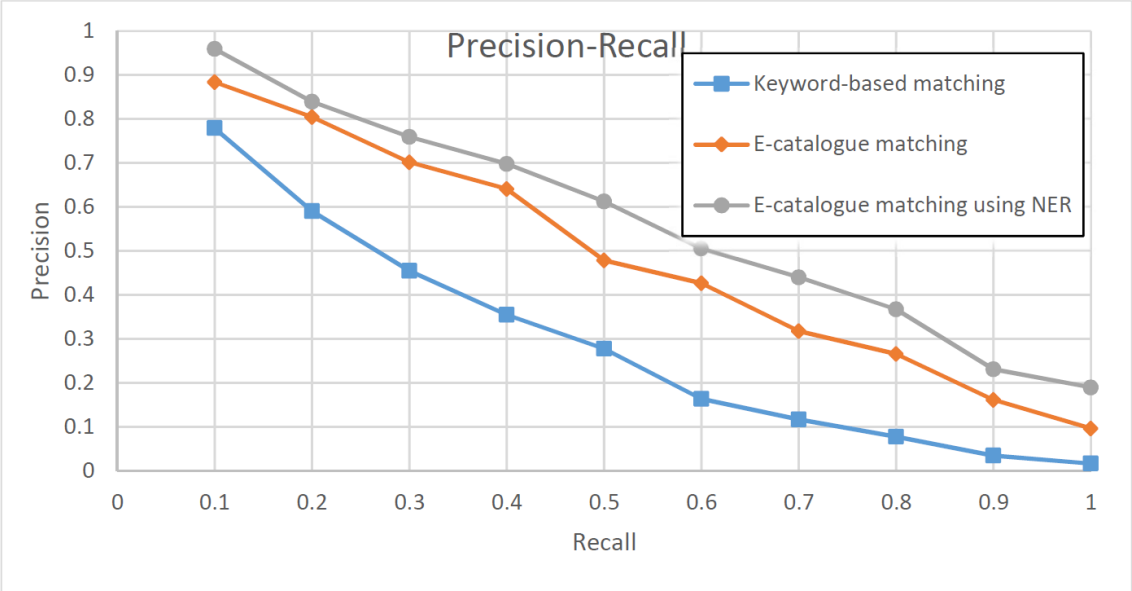

Figure 6.5 Precision-Recall curve for comprehensive test set

The evaluation results show the semantic e-catalogue matching engine can improve the search capabilities when searching using synonym terms in tender search. This helps the suppliers to achieve better results in finding relevant tenders that can facilitate finding business opportunities in public procurement marketplaces.

The matching mechanism was tested on different business sectors available in TED public procurement portal. The MAP table shows an overall improvement in matching performance regardless of the business sector. In order to test the independency of the matching mechanism to the dataset, an F-test has been used to show if the variances of the keyword-based result set and the e-catalogue matching result set are not significantly different.

As it can be seen in Table 6.6 and Table 6.7, F is lower than $F_{Critical}$ that shows two samples come from populations with almost equal variances. This means there is no significant difference in variances of the two results sets. Consequently, the matching algorithms shows almost similar improvement in all different business sectors.

Table 6.6 F-Test Two-Sample for Variances

	Keyword-based matching	e-catalogue matching
Mean	32.34222	50.77778
Observations	18	18
Df	17	17
F	0.354207	
P(F<=f) one-tail	0.019488	
F Critical one-tail	0.440162	

Table 6.7 F-Test Two-Sample for Variances

	Keyword-based matching	e-catalogue matching using NER
Mean	32.34222	58.80611
Observations	18	18
Df	17	17
F	0.251452	
P(F<=f) one-tail	0.003413	
F Critical one-tail	0.440162	

The second question that we should answer here is that if the e-catalogue matching mechanism shows a significant performance improvement compare to the keyword-based matching. This question can be answered by using a T-test. A T-test is a statistical test that can be used to determine if two sets of data

are significantly different from each other[33]. This test is most commonly applied when the test statistic would follow a normal distribution. The histogram in Figure 6.6 shows that both result sets follow almost a normal distribution.

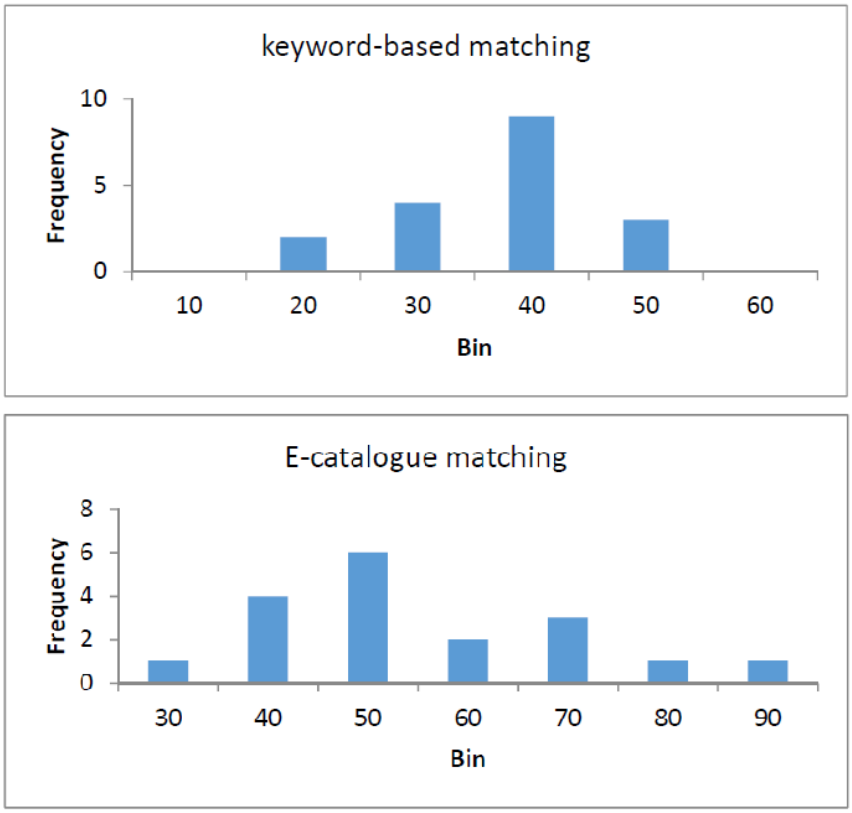

Figure 6.6 histogram of MAP result sets

The results of the T-test for comparing keyword-based result set with e-catalogue matching results are shown in Table 6.8 and the results of comparing with e-catalogue matching using NER results are shown in Table 6.9. In a T-test, absolute value of T_{stat} should be bigger than $T_{critical}$ in order to show a significant difference. As it can be seen in both cases the results show a significant difference.

[33] T-test is used since the population variance is unknown, otherwise we could use Z-test for the same target.

Table 6.8 t-Test: keyword-based Vs. E-catalogue matching

	Keyword-based matching	e-catalogue matching
Mean	32.34222	50.77778
Observations	18	18
Pooled Variance	164.2699	
Pearson Correlation	0.154858726460093	
Hypothesized Mean Difference	0	
df	17	
t Stat	-4.99601342536436	
P(T<=t) one-tail	0.0000552480907492111	
t Critical one-tail	1.73960672607507	
P(T<=t) two-tail	0.000110496181498422	
t Critical two-tail	2.10981557783332	

Table 6.9 t-Test: keyword-based Vs. E-catalogue matching using NER

	Keyword-based matching	E-catalogue matching using NER
Mean	32.3422222222222	58.8061111111111
Observations	18	18
Pearson Correlation	0.0716268688815951	
Hypothesized Mean Difference	0	

df	17
t Stat	-6.01755392511439
P(T<=t) one-tail	6.92619767697676E-06
t Critical one-tail	1.73960672607507
P(T<=t) two-tail	0.0000138523953539535
t Critical two-tail	2.10981557783332

6.4 Multi resource matching

6.4.1 Test Scenario

Searching and selecting the best suitable opportunities among several published tender calls especially from various tendering resources is a crucial and time-consuming task for several business actors in e-procurement marketplaces. Most of the tendering portals provide keyword-based search and category-based notifications to the subscribers for Tenders and Contract Awards. The idea is that the tender notification systems deliver tender opportunities to the suppliers, dramatically reducing the amount of time spent looking for these tenders. But according to the potential wide range of products in a business sector, a supplier may receive an extensive list of notifications which makes it difficult to find the best-matched opportunities with the supplier's product portfolio.

This problem can be solved using a search mechanism that is able to rank and sort the tender calls coming from various resources based on their similarity to a supplier's products or services. Consequently, suppliers can save time in searching and work on preparing proposals for the most similar calls to their products that definitely have more chance to win the competition. To reach this goal, the product information provided by a supplier can be used by a product search mechanism in order to find and recommend similar product requests (Julashokri et al., 2011) and tender calls (Mehrbod et al., 2017).

Since different e-procurement platforms follow their own standard formats for modelling contracts, tender notices, and catalogues, they will not be able to share this information with each other. The biggest disadvantage of this is that the users have to subscribe in all the platforms to be able to get access to various opportunities originating from different platforms and apply for the opportunities. While using standard formats helps to use the same information in different documents in the process flow and decrease the efforts needed to correct errors and fix problems by automating e-procurement process, standards do not provide sufficient coverage for all steps of procurement (Mehrbod, Zutshi, & Grilo, 2014a).

In order to improve interoperability of such systems, the e-catalogue search approach is applied to the problem of finding tender notices from open tendering e-marketplaces in order to find opportunities from heterogeneous tender resources. A test mechanism has been used to evaluate the performance of the search approach in finding related calls from different tendering websites. Tender notices from two major tender resources including United Nations Global Marketplace (UNGM) and Tenders Electronic Daily (TED) have been used to test the search mechanism.

One of the main features of the semantic mechanism used by the e-catalogue matching engine is that the search process is not dependent on any specific ontology and relevant annotated data. The search engine is able to use any ontology and tries to find the best ontology for interpreting the data among all available ontologies and in the absence of a suitable ontology uses basic keyword search methods. This approach makes it possible to use the benefits of all available ontologies and schemas but not to be dependent on them. I.e. the search service exploits the data and the structures of B2B documents in the matching process but is not independent of any pre-specified structure. For more details about the e-catalogue matching mechanism and used algorithms please refer to Chapter 5.

A test mechanism has been used to evaluate the performance of the search approach in finding related calls from different tendering websites. Tender notices from two major tender resources including United Nations Global Mar-

ketplace (UNGM) and Tenders Electronic Daily (TED) have been used to test the search mechanism.

6.4.2 Test definition

This test applies the mentioned e-catalogue matching engine to the public tenders' search and evaluates it on finding tenders from two different procurement resources. This will open the opportunity to find tenders from various data resources and marketplaces without converting data to a uniform model that can be time-consuming and costly.

Based on the application and the target data resource, each e-procurement ontology used a clarification vocabulary. As mentioned, this clarification vocabulary is more important than the ontology schema for searching product data especially when searching data from unstructured or semi-structured resources.

In this section, we want to study the effect of using different classification vocabularies on the tender search using the e-catalogue matching mechanism. The proposed e-catalogue matching engine allows using various ontologies for data indexing and query process. The effects of using two different vocabularies including CPV and UNSPCS on search performance have been compared in order to evaluate the search mechanism in tender notice search. This test shows the capability of the matching mechanism in using relevant vocabularies for interpreting tenders from different resources and tolerance of the search mechanism to encounter unknown data.

Figure 6.7 shows the overall view of the evaluation mechanism. Tender notices from UNGM and TED have been collected in the tenders' repository. Search results in three different states using CPV vocabulary, UNSPSC vocabulary and both at the same time have been illustrated and compared in Figure 6.8.

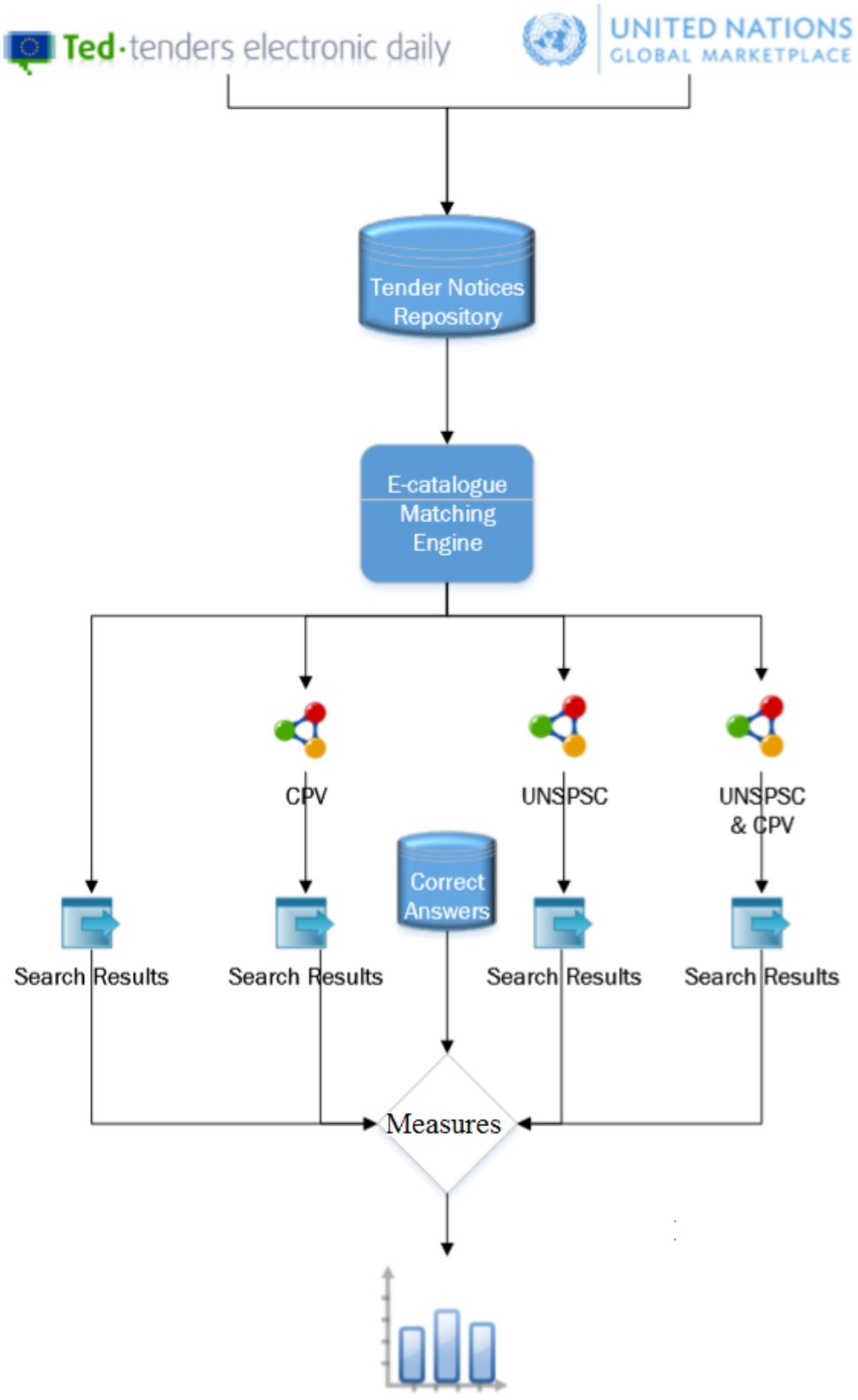

Figure 6.7 Classification vocabularies test

6.4.3 Data Gathering

The tender notices that have been published on TED and UNGM websites are used to provide a tender repository in order to search for business opportunities. These two portals are selected as two major resources of public tenders published using different structures and classification systems. The goal is to provide a heterogeneous repository of public tenders to test the business opportunity search scenario.

TED is the online version of the "Supplement to the Official Journal" of the EU that is dedicated to European public procurement. TED provides free access to business opportunities from the European Union, the European Economic Area and beyond. Every day, more than 1,700 public procurement notices from various European countries are published on TED.

UNGM is the common procurement portal of the United Nations. The United Nations represents a global market of over USD 15 billion annually for all types of products and services. The UNGM acts as a single window, through which potential suppliers may register in the UN vendor database. These organizations account for over 99% of the total UN procurement spent. The UNGM enables vendors to be aware of upcoming tender notices.

Tender notices in both websites are categorised according to the relevant procurement vocabularies and can be searched by the business category. TED uses CPV that has been developed by the European Union to classify products and services in procurement contracts and is mandatory in the European Union since 1 February 2006. UNGM uses UNSPSC which is the product coding developed and used by UN to describe the product and services that their agencies need.

The test repository constructed from tenders of these two different resources in order to evaluate the capability of the matching mechanism in dealing with tenders from heterogeneous resources. Tenders that have been published on 2015 collected to make the tender repository. In the TED websites, the tender archives can be downloaded for registered user in the form of XML files. For UNGM a JSOUP based extractor has used to gather the data. Table 6.10 shows a summary of the collected tenders.

Table 6.10 test repository

TED (CPV)			UNGM (UNSPSC)		
Code	**Title**	**Count**	**Code**	**Title**	**Count**
32250000	Mobile phones	37	43191501	Mobile telephones	23
44111000 44110000	Building materials Construction materials	46	11111600 22101600 30130000 30131700 40141700	Stone Paving equipment Structural building products Tiles and flagstones Hardware and fittings	20
66110000	Banking services	96	84120000	Banking and investment	36
79341000	Advertising services	53	82100000	Advertising	66
90700000 38000000 71630000	Environmental services Laboratory, optical and precision equipments (excl. glasses) Technical inspection and testing services	581	77000000	Environmental Services	254

64110000	Postal services	5	78102201	National postal delivery services	0
65100000	Water distribution and related services	11	83101501	Supply of water	1
45232430	Water-treatment work	16	70171501 83101506	Water quality assessment services Water treatment services	2
37520000	Toys	1	60141000	Toys	0
15000000	Food, beverages, tobacco and related products	150	50000000	Food Beverage and Tobacco Products	37
38300000	Measuring instruments	31	41110000	Measuring and observing and testing instruments	167
35113000	Safety equipment	6	46160000	Public safety and control	18
03111000	Seeds	2	10150000	Seeds and bulbs and seedlings and cuttings	13
72222300	Information technology services	47	81110000	Computer services	217

6.4.4 Test Results

Since for the measuring we have to define the correct answer set, tender notices from similar categories (listed in Table 6.10) of different portals combined to make correct answer set for each category. For example, tender notices from "Mobile telephones" (CPV code 32250000) and "Mobile phones" (UNSPSC code 43191501) are considered as the correct answers while searching for tenders about mobile phones. The test repository contains 23 tenders of "Mobile telephones" category and 37 tenders of "Mobile phones" category that constitute the correct answer set and more than 1100 tenders from other business sections that are considered as false answers for this test. Therefore, a search mechanism will be considered as 100% precise, if retrieves all 60 tenders from the mobile category in response to a relevant search query. If a search mechanism retrieves 20 tenders from the correct answer set and 10 from the rest of the repository, its recall is 20/60 and precision is 20/30. Table 6.11, Table 6.12 and Table 6.13 show the test results in three different states consequently. First, when the search mechanism uses the UNSPSC vocabulary for data indexing and search. Second, when it uses CPV vocabulary for data indexing and search. Third, when both vocabularies are used by search mechanism at the same time and when the search engine doesn't use any vocabulary.

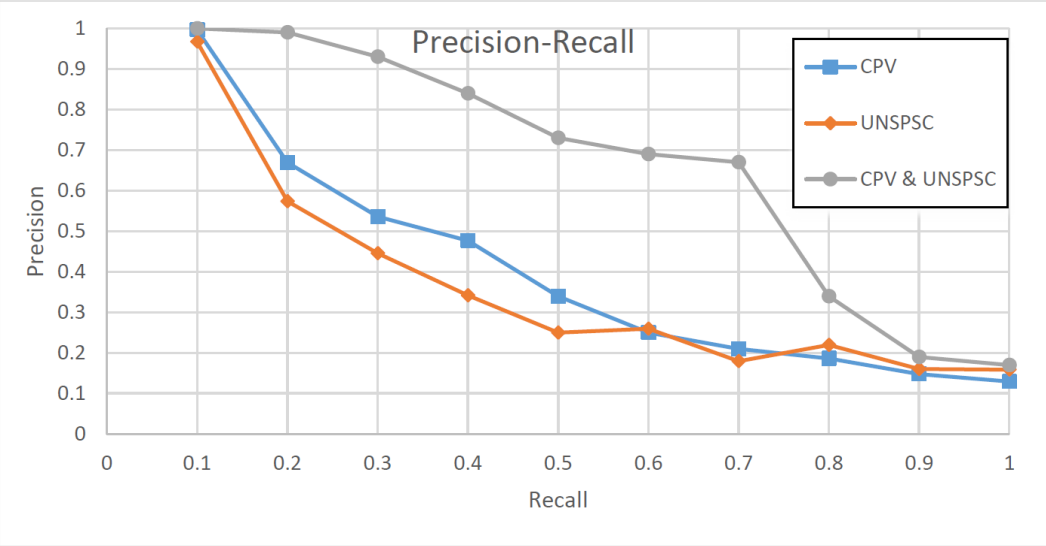

Figure 6.8 Test results in three different states

The results of these test cases are compared with the correct answers that are gathered using category-based search service of the TED and UNGM in or-

der to calculate the search performance measures. In order to compare the results easily, Figure 6.8 shows a summary of all tables. Each curve shows the average of all search results for finding related tender calls from all sections (last entry of each table) in the repository.

Table 6.11 test results using UNSPSC vocabulary

UNSPSC	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	MAP
10150000	1	0.49	0.51	0.51	0.24	0.18	0.18	0.07	0.07	0.07	36.87
11111600	1	0.79	0.61	0.34	0.26	0.13	0.10	0.11	0.11	0.11	36.09
41110000	1	0.72	0.59	0.41	0.09	0.09	0.09	0.10	0.11	0.10	33.33
43191501	0.95	0.25	0.24	0.13	0.11	0.11	0.09	0.09	0.09	0.09	21.78
46160000	0.93	0.44	0.28	0.25	0.15	0.11	0.11	0.11	0.12	0.11	26.54
50000000	1	0.40	0.38	0.29	0.13	0.9	0.10	0.9	0.03	0.03	28.83
60141000	1	1	1	1	1	1	1	1	1	1	1
70171501	0.9	0.61	0.32	0.32	0.24	0.16	0.16	0.08	0.09	0.09	34.85
77000000	0.85	0.53	0.35	0.17	0.15	0.11	0.10	0.09	0.09	0.09	25.83
78102201	1	1	0.53	0.53	0.43	0.43	0.07	0.07	0.05	0.05	41.92
81110000	1	0.49	0.40	0.19	0.17	0.14	0.13	0.13	0.11	0.10	28.95
82100000	1	0.59	0.43	0.20	0.14	0.12	0.12	0.11	0.10	0.10	29.44
83101501	1	0.11	0.11	0.14	0.14	0.14	0.12	0.12	0.07	0.07	26.78
84120000	0.90	0.56	0.44	0.27	0.26	0.14	0.14	0.14	0.14	0.14	31.69
AVG	0.96	0.57	0.44	0.34	0.25	0.26	0.18	0.22	0.16	0.15	35.92

Table 6.12 test result using CPV vocabulary

UNSPSC	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	MAP
10150000	1	0.63	0.51	0.51	0.25	0.20	0.20	0.08	0.04	0.04	36.84
11111600	1	0.65	0.63	0.61	0.32	0.20	0.18	0.16	0.14	0.08	36.11
41110000	1	0.64	0.46	0.44	0.10	0.10	0.04	0.02	0	0	33.33
43191501	1	0.60	0.53	0.53	0.30	0.18	0.18	0.18	0.05	0.01	21.79
46160000	1	0.48	0.18	0.17	0.15	0.13	0.13	0.12	0.09	0.07	26.56
50000000	1	0.58	0.54	0.46	0.37	0.22	0.19	0.05	0.02	0.02	28.83
60141000	1	1	1	1	1	1	1	1	1	1	100
70171501	1	0.79	0.54	0.54	0.47	0.24	0.24	0.15	0.17	0.17	34.86
77000000	1	0.66	0.57	0.33	0.16	0.13	0.14	0.13	0.06	0.02	37.23
78102201	1	1	0.56	0.56	0.50	0.50	0.14	0.14	0.01	0.01	43.9
81110000	1	0.54	0.26	0.20	0.15	0.12	0.13	0.11	0.10	0.03	26.8
82100000	1	0.51	0.48	0.47	0.45	0.14	0.13	0.11	0.07	0.04	29.66
83101501	1	0.55	0.55	0.26	0.26	0.23	0.16	0.16	0.13	0.13	26.78
84120000	0.95	0.69	0.62	0.52	0.36	0.16	0.13	0.12	0.12	0.10	31.29
Average	0.99	0.66	0.53	0.47	0.34	0.25	0.21	0.18	0.14	0.12	36.71

Table 6.13 test results using both UNSPSC and CPV vocaburalries

UNSPSC	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	MAP
10150000	1	0.99	0.92	0.78	0.64	0.64	0.64	0.62	0.08	0.08	39.01
11111600	1	1	0.98	0.93	0.76	0.75	0.71	0.17	0.14	0.13	40.30
41110000	1	1	0.92	0.75	0.67	0.67	0.60	0.19	0.06	0.03	28.29
43191501	1	1	0.97	0.88	0.77	0.76	0.82	0.70	0.078	0.01	36.15
46160000	1	1	1	0.97	0.81	0.65	0.60	0.14	0.13	0.12	25.69
50000000	1	1	0.99	0.95	0.89	0.88	0.86	0.33	0.32	0.32	38.60
60141000	1	1	1	1	1	1	1	1	1	1	1
70171501	1	1	0.98	0.82	0.69	0.61	0.50	0.25	0.16	0.16	48.50
77000000	1	0.98	0.83	0.65	0.6	0.57	0.68	0.17	0.13	0.11	32.55
78102201	1	1	0.91	0.86	0.72	0.52	0.51	0.52	0	0	44.64
81110000	1	0.95	0.86	0.7	0.59	0.53	0.52	0.13	0.09	0.05	27.05
82100000	1	1	0.99	0.96	0.85	0.84	0.76	0.16	0.15	0.15	34.57
83101501	1	0.94	0.8	0.68	0.52	0.53	0.52	0.21	0.14	0.14	39.41
84120000	1	1	0.94	0.91	0.76	0.72	0.71	0.16	0.17	0.16	38.01
Average	1	0.99	0.93	0.84	0.73	0.69	0.67	0.34	0.19	0.17	40.91

In order to compare the search results, the precisions are calculated for standard recall points that are shown on the interpolated standard precision-recall curve in Figure 6.8. Generally, the performance of a search engine is shown using Precision-Recall curves that represent these two inversely related metrics (Mehrbod et al., 2015). Finding a balance between these two metrics is dependent on the mission of the search engine. In such curve, the one to the top right shows better search performance. Each curve represents average interpolated precision-recall values for the matching mechanism in one of the mentioned states.

As it can be seen in the figure, the semantic product matching engine can use various vocabularies to improve tender search performance while searching tenders from various resources. The test tender repository includes tenders developed based on CPV and UNSPSC classification systems. The test results show the capability of the search mechanism to use relevant vocabularies for interpreting the data.

The CPV and UNSPSC curves show the search performance while using the CPV and UNSPSC vocabularies respectively. In each case, the search mechanism uses one vocabulary, consequently some part of underlying data is known for the system and some part is not. As it can be seen in Figure 6.8, the system has almost similar performance in both cases. While the search performance shows the system didn't fail to return the results, the performance is lower than the third state which both ontologies have been used to index and search the data. In CPV & UNSPSC state, the search mechanism uses both vocabularies at the same time to interpret the data that causes to obtain higher performance.

The test results show the matching mechanism not only is able to use available vocabularies to interpret the tenders in order to improve the search performance, it also can tolerate lack of existence of relevant vocabularies. By using different vocabularies, the search engine can find tenders from different resources of tendering websites and consequently can help suppliers in finding more business opportunities in procurement marketplaces.

6.5 *B2BProduct NER Accuracy Test*

6.5.1 Test Scenario

Published tenders in e-Procurement marketplaces are the main resources for finding business opportunities for suppliers. Besides the tender search, e-sourcing that acts as the starting point of the e-procurement process contains searching for suitable suppliers. Accordingly, e-tendering repositories such as UNGM and TED have an essential role in connecting contracting authorities and suppliers. However, e-Tendering marketplaces rerely provide supplier search and usually provide simple keyword-based search and category-based notifications to the subscribers for Tenders and Contract Awards. Semantic search can improve search capabilities of such marketplaces.

The matching mechanism has to find the product mentioned from B2B documents in all search scenarios. For example, for finding business opportunities for a supplier, the matching mechanism has to find product mentions from published tenders to match them with products in the supplier e-catalogue. Therefore, providing a NER system that can recognise the products mentioned in such tenders is a fundamental block in the process of searching for business opportunities. For this purpose, as presented in the previous chapter, a B2B-Product recogniser was developed and used as a part of the matching mechanism. In order to test the accuracy of the B2B-Product NER model, the trained model has been tested in two different test datasets, consist of an automatically annotated dataset and a manually annotated dataset.

Most of the available NER applications are developed in order to extract names of persons, organizations and locations from text. But the rapid development of e-Commerce increased the demand for Product NER (PNER). All previous researches in the area of Product NER, are focused on B2C e-commerce, but this research work applied NER issue to B2B e-commerce.

6.5.2 Test Definition

This section evaluates the performance of the trained model on two new corpuses of B2B product data. The objective of this evaluation is to measure and analyse the efficiency of the discussed training approach in training the learning-based NER mechanism and the accuracy of the trained model for NER task in B2B e-Marketplaces.

In order to evaluate the trained B2B-Product NER model, test datasets have been created automatically by applying the same method that has been used to prepare the training corpus to different data sources from four different tendering websites. Furthermore, small samples of each corpus are used to create manual test dataset in order to compare the performance of the model with the manual results that experts can achieve. Figure 6.9 shows the overview of the test.

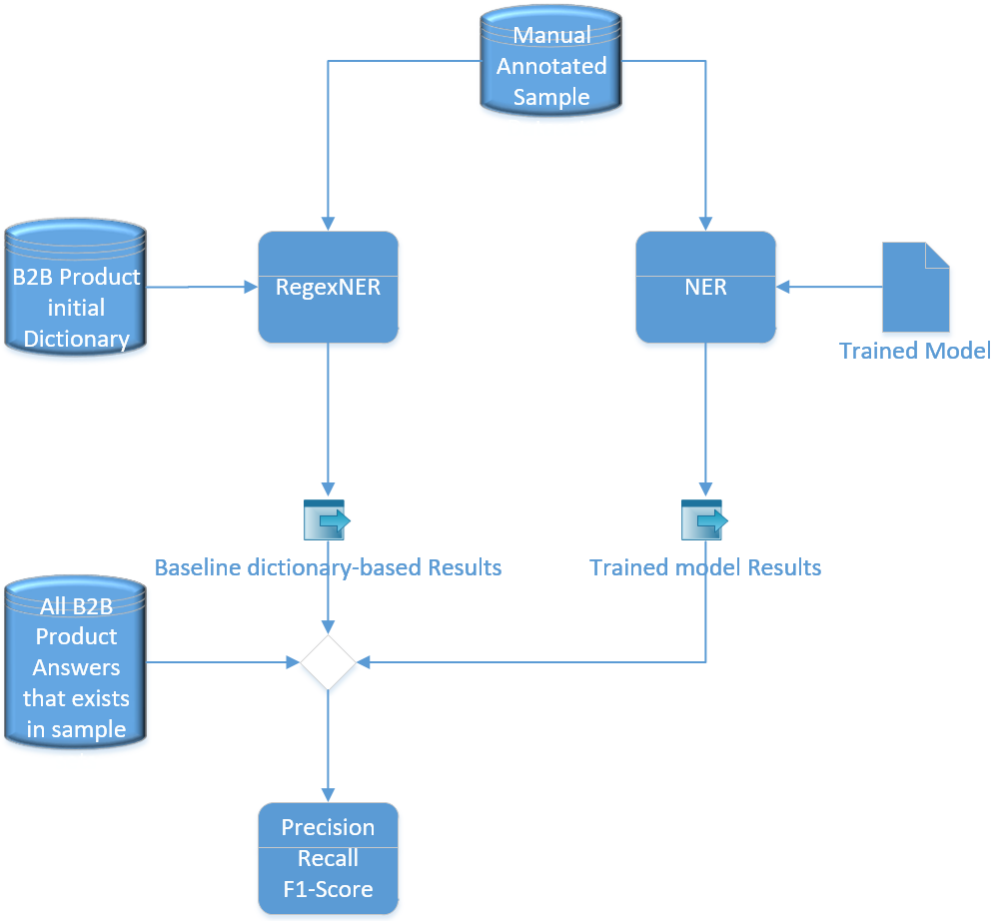

Figure 6.9 B2Bproduct NER test

6.5.3 Data Gathering

The TED 2013 test set includes 443,079 tenders that had been published in TED on 2013. The TED 2015 test set includes 270,467 tenders that have been published in TED on 2015 until October. These tenders have been used to prepare two different annotated test corpuses. Note that the tenders of 2014 have been used to train the model and two groups of tenders, one that had been published before and another that have been published after are used to evaluate the model.

Furthermore, the model has been evaluated on other test datasets from different tendering websites including The United Nations Global Marketplace, B2B Quote and UK Government's contracts finder.

The United Nations Global Marketplace (UNGM[34]) is the common procurement portal of the United Nations system of organizations. It brings together UN procurement staff and the vendor community. The United Nations represents a global market of over USD 15 billion annually for all types of products and services.

The UNGM acts as a single window, through which potential suppliers may register in the UN vendor database. These organizations account for over 99% of the total UN procurement spent. The UNGM enables vendors to keep abreast of upcoming tender notices. By subscribing to the Tender Alert Service, vendors can receive relevant business opportunities emailed directly. The UNGM also acts as an important procurement tool to shortlist suppliers for competitive bidding.

B2B Quote[35] was established in 2006 as a low cost, high quality and easy to use 'business to business' Tendering website. The website is focused on two ranges of tenders including low value public and private sector tenders and high-value tenders in the UK. It focuses on a specific range of Industry Sectors

[34] www.ungm.org

[35] www.b2bquote.co.uk

in order to provide a much more personalized service to customers, therefore, it doesn't follow any standard product classification system.

Contracts Finder[36] is the UK Government's marketplace for suppliers to find new procurement opportunities totally free of charge. It allows users to view and search the UK Government's pipelines of potential procurement activity and awarded contracts. It is a critical tool for addressing the Government's transparency commitments. The Public Contracts Regulations require most public sector bodies to advertise their new opportunities and contract award information here, so that all suppliers have better, more direct access to Public Sector work.

Contracts Finder provides search for information about contracts worth over £10,000 with the UK government and its agencies. It can be used to search for contract opportunities in different sectors, to find out what's coming up in the future and to look up details of previous tenders and contracts.

6.5.4 Test Results on automatic annotated test datasets

In order to evaluate the performance of the model, the precision and recall of the model have been calculated in the mentioned test datasets. Table 6.14 shows the results obtained from the application of trained model in a learning-based NER task on TED 2013 dataset, TED 2015 dataset, Contracts Finder dataset, B2B Quote dataset and UNGM dataset.

The evaluation results show the trained model is able to recognise the product mentions from various resources of B2B tenders. The discovered products not only include the known products from the initial dictionary but also include new product mentions as well as misspelled versions of the products, which is the advantage of learning-based NER approaches.

[36] www.gov.uk/contracts-finder

Table 6.14 B2B-Prodcut NER Evaluation results

Test set	Recall	Precision	F1-score
TED 2013	83%	94%	0.88
TED 2015	80%	94%	0.86
Contracts Finder	71%	98%	0.82
B2B Quote	67%	81%	0.73
UNGM	63%	74%	0.68

The trained model obtained a high ratio of precision in different test datasets. This means the NER process is able to correctly recognise the B2B-Products that are mentioned in the test tenders. Obviously, this will help any search mechanism which uses the NER model to correctly extract more important keywords for product search in B2B documents. The high ratio of recall shows the ability of the model to retrieve high ratio of the available B2B-Products from the search corpus. Therefore, the search mechanism will miss fewer keywords that should be included in the search process.

As it can be seen in Table 6.14, the model has obtained better results on TED and Contracts Finder than UNGM and B2B Quote. Actually, the difference comes from the different classification systems that are used by these e-marketplaces to categorise products. TED and Contracts Finder use Common Procurement Vocabulary (CPV) classification system that is common in Europe while UNGM uses United Nations Standard Products and Services Code (UNSPSC) and B2B quote uses its own categorization. CPV is developed by the European Union to facilitate the processing of invitations to tender published in the Official Journal of the European Union (OJEU) by means of a single classification system to describe the subject matter of public contracts.

UNSPSC is an open, global, multi-sector standard for efficient, accurate classification of products and services managed by UN Development Programme. Since tenders from TED has been used to train the model, the recogniser is more precise in detecting terms that are used to describe goods

and services in CPV based tenders. The accuracy of the model can be improved by using tenders from other resources in preparing the initial dictionary and the training process that can be future work of this research work.

6.5.5 Test Results on manually annotated test datasets

In order to provide test datasets for testing the performance of the trained model in extracting the products that are not mentioned in the titles and CPV references of the tenders, small samples from each data set are selected and the mentioned B2B products are annotated manually. The importance of this test is that some products may be mentioned in the descriptions of the tenders but don't be mentioned in the titles. Therefore, this test provides the performance evaluation in a new situation. Furthermore, since the test datasets are annotated manually the test results can be compared with the optimal results that can be achieved manually by experts.

Each test dataset includes a few randomly selected tenders and contains about sixty B2B product mentions. All available B2B product mentions in the sample dataset are annotated manually in order to compare with the B2B-Produtcs that are extracted by the model from these datasets. Table 6.15 shows the performance results of the model on manually annotated sample test data sets.

In order to have a comprehensible view of the obtained results, Table 6.15 also shows the results obtained by a dictionary-based NER mechanism using the initial dictionary on the same test datasets. The initial dictionary that had been made from the titles and CPV references of the tenders and had been used in providing the training set, is used as the dictionary for a dictionary-based NER approach. The obtained results can provide an overview of comparing the results of the trained model with the results of a non-machine learning mechanism on the same test data.

While a dictionary-based NER approach can provide very high Precision value, the problem of such approaches is the very low value of the obtained Recall that results in low F_1-Score. In other words, the dictionary-based NER looks for the occurrences of its dictionary seeds in the descriptions of the tenders and

retrieves them as the results. Since all the returned B2B-Products exist in the dictionary and therefore are correct answers, the precision is 100%. But the problem is this technique fails to extract the B2B-Products that don't exist in the dictionary which leads to low Recall value. As in can be seen in Table 6.15, the trained model improves the Recall factor by detecting more B2B-Producs that are don't exist in the initial dictionary.

Table 6.15 Evaluation Results on manual test datasets

Test set	Recall (trained model)	Precision (trained model)	F_1-Score (trained model)	Recall (initial Dictionary)	Precision (initial Dictionary)	F_1-Score (initial Dictionary)
TED 2013	42.30	95.65	58.66	27.86	100	43.58
TED 2015	53.08	84.31	65.15	22.58	100	36.84
Contracts Finder	45.45	83.33	58.82	18.18	100	30.76
B2B Quote	35.29	72.0	47.36	22.72	100	37.03
UNGM	32.0	72.72	44.44	18.64	100	31.42

6.6 Summary

In order to validate the proposed e-catalogue matching mechanism, we developed a supplier finder service (Mehrbod et al., 2015) in a G2B2B procurement platform called Vortal. The supplier finder service helps the buyers to find suitable suppliers on the platform based on the similarity between suppliers product e-catalogues to the buyers e-catalogue. The heterogeneity of the underlying product catalogues was the main reason to use the

concept-based VSM approach for developing a practical product matching engine. The developed product search mechanism is able to match similar product from various resources without the transformation and integration overhead.

The supplier finder has been tested in four different states in order to evaluate the different blocks of the matching mechanism. The syntactic matching mechanism is tested using a set of e-catalogues in various structured and unstructured formats. The experimental results show the matching process is able to match diverse formats of catalogues from various sources.

The semantic matching mechanism is tested using a set of e-catalogues from various classification systems and semantically heterogeneous resources. The matching process is capable of matching various types of catalogues that come from different sources. The experimental results show that the proposed approach improves the similarity ratio between similar e-catalogues compared with the basic approach of Vector Space Model and syntactic matching mechanism.

The semantic search mechanism tries to understand the contextual meaning of the words within the search domain. Accordingly, a B2B-Product recogniser has been developed that can extract "B2B Product" mentions from tenders and other B2B documents. The recogniser can be used as a fundamental search element extractor in semantic search process in B2B e-marketplaces. A self-learning approach has been adopted in order to train the required model for extracting B2B-Product mentions. The proposed approach uses already known product mentions in tenders as the training data to train the model and then use the trained model to recognize the product mentions from other documents.

The model has been tested using tenders that have been published in public procurement e-marketplaces including The United Nations Global Marketplace (UNGM), Tenders Electronic Daily (TED) which is dedicated to European public procurement, B2B Quote which is focused on tenders in UK and UK Government's contracts finder. The results show that the proposed approach achieved high values of precision and recall in different test datasets.

7

Conclusions

7.1 The problem and the motivation

In a procurement marketplace, many companies and organizations come together to purchase the products that they need or to sell their products. Tenders that are published by the buyers provide an extensive and valuable resource for suppliers to find business opportunities. E-catalogues that are used by companies to explain the products have an effective impact in this process. The product data that is published in the e-catalogues can be used as search queries to find appropriate tenders for a company. Matching an e-catalogue with the similar e-catalogues in the system helps the buyers in selecting proper suppliers. It also helps the suppliers to find new markets for their products in B2B marketplaces. This turns the search service to one of the major technical factors in the success of a procurement e-marketplace. But procurement e-marketplaces usually provide simple keyword search services for finding business opportunities.

While e-catalogues are widely used by suppliers and buying organizations to share the product and services information, diversity of structures and terms in creating e-catalogues is a barrier on their search ability. Since organizations may use different schemas, classifications and expressions for describing products in e-catalogues, it is challenging to match a product with the e-catalogue requested by another partner. This heterogeneity makes it difficult and time-consuming to integrate and query e-catalogues.

Enriching product data with semantic concepts can improve the searchability of the tenders and helps companies in finding suitable business opportunities. However, the heterogeneity of e-catalogue structures and tender resources is the challenge for finding related documents to a request. Integrating and publishing tenders, catalogues and other B2B documents in uniform semantic data models is suggested for avoiding the matching problem. But the integration process not only is difficult and expensive because of variety of structures that are used by different companies, it cannot tolerate lack of semantics and violations from the model assumptions. The documents have to be published or republished according to the semantic resources expected by the model which is not always feasible. These steps usually contain manual efforts that affect the extensibility of the solution. According to the wide number of procurement data resources, such solutions can be expensive for companies that want to keep track of all potential procurement opportunities.

Therefore, the main question covered by this research work was how can buyers and suppliers match their e-catalogues in an efficient way, with no restrictions regarding data integration models? In other words, how can the procurement documents be matched without forcing a data integration model to the companies, publishers or the matching solution providers?

7.2 Contribution of this Thesis

This research improves product search in e-procurement platforms by providing a semantic and syntactic matching mechanism for e-catalogues. Various procurement datasets, as well as e-catalogues, have different syntaxes and semantics which makes it hard to search and use them in an integrated manner. In consideration of cost, limitations and extensibility issues of the data integration models for matching heterogeneous e-catalogues, this research work proposed a flexible data indexing method to solve e-catalogue matching problem. Therefore, an information retrieval technique is used to encounter this problem and match various types of e-catalogues. This technique that is called Vector Space Model, is basically designed for text searches and is used by many search

engines. But because of its flexibility and simplicity, it has extended to apply to a wide range of search problems.

In the first step, Vector Space Model had been applied to find syntactic similarity ratio between e-catalogues. In order to implement the proposed matching process, an open source full-text search tool and a natural language analyser were used to extract terms from flat text files. Then the search tool was extended to consider the locational values of words in term extraction process when such information is available. The matching process has used combinations of values, names and location of attributes of structured documents to find the syntactic correlation of e-catalogues.

A table of coefficients is proposed to specify the matching process for standard e-catalogues as a boosting mask. This mechanism increases the search precision by removing unrelated information from the matching process and boosting the weights of important tags. Since an e-catalogues contains various information with different importance for matching process, this adjustment helps the process to benefit from customizing search mechanism for known structures.

In the next step, the matching process has been expanded to exploit both syntactic and semantic aspects in the calculation of the similarity ratio. Procurement ontologies were used to expand the matching mechanism with semantic relationships of the product data attributes. In this process vectors of each e-catalogue were enriched with semantic concepts that exist in the e-catalogue. Adding semantic relationships to the terms of the vectors enables the matching process to find semantically similar e-catalogues. The proposed approach makes it possible to use the benefits of all available ontologies and schemas but not to be dependent on them. I.e. the search service exploits the data and the structures of e-catalogues in the matching process but is independent of any pre-specified structure.

Coefficients corresponding to the geodesic distances of semantic concepts are used to adjust the effects of available semantic relationships on the overall similarity. The default values for the weight coefficients and their growth rates can influence the similarity measure. This helps the semantic matching mecha-

nism to give higher effect on the matching results to the terms that are semantically closer to the search query than the less related words.

The semantic e-catalogue matching process includes finding the potential entities to enrich the vector model with the semantically related concepts in the search domain. In order to detect semantic concepts from a procurement document, an Entity Recogniser has been developed. A supervised machine learning approach has been used to develop the Recogniser as a complimentary component of the matching engine. A self-learning method has been adopted to provide a training dataset that is needed to train the supervised learning model. Test results show the trained model is able to recognise product mentions from different tenders that have been collected from various e-procurement marketplaces. The entity recogniser in B2B context can help the semantic matching process in information retrieval systems of e-procurement marketplaces by providing a richer semantic search on heterogeneous procurement documents.

The proposed matching process has been tested using a heterogeneous set of e-catalogues and e-tenders. The test scenarios evaluated the application of the e-catalogue matching engine in searching in procurement e-marketplaces and its capability in improving the search performance. Four test cases have been defined to evaluate various features of the e-catalogue matching mechanism in main search scenarios that are possible in a procurement marketplace.

The proposed approach was implemented in a G2B2B e-procurement platform and the results of the e-catalogue matching mechanism were reported. The search results show the matching process is capable of matching various types of catalogues that come from different sources without restricting a data integration model. The proposed approach is not dependent on any assumption or underlying structure and is extendable to any new type of e-catalogue. The experimental results show that the proposed approach improves the similarity ratio between similar e-catalogues compared with the basic approach of Vector Space Model. Although the evaluation shows improvement in the matching ratios and number of retrieved instances, the most important value proposition of the proposed approach is its simplicity and practicality for implementation.

Beside the e-catalogue matching scenario that is mostly used by the buyers to find suppliers or in the other words helps the supplier to be seen by the buyer using their e-catalogues, another common search scenario has been tested to see how suppliers can improve their efficiency in finding business opportunities in e-procurement platforms using the content of their e-catalogues?

The business opportunity finder test-case evaluates the proposed e-catalogue matching engine in tender datasets retrieved from online public procurement portals. This test-case which evaluates different features of the matching mechanism such as the capability of detecting semantically related terms, exploiting available semantic resources and tolerating missing semantic resource, results shows improving the matching performance in tender search. This helps the suppliers to achieve better results in finding similar tenders to their e-catalogues as the main resource of business opportunities in procurement marketplaces. Consequently, suppliers can use the search engine to find tenders from various procurement resources using content of their e-catalogues.

7.3 Areas for Further Development and Research

This thesis has proposed a flexible approach for finding similar and related procurement documents in e-marketplaces. The matching mechanism provides the search results based on the product data contained in e-catalogues and other procurement documents. The research on e-catalogue matching and proposed solution can be extended in various aspects.

The syntactic matching layer is responsible for exploiting the information that can be figured out from the structure of the procurement documents. This process can be easily fine-tuned for the formats that are known for the system using boosting masks. Each boosting mask is a coefficient table for a known structure for adjusting the indexing weights. In future, these tables can be customized automatically for different structures using a learning mechanism based on searchers' profiles, search history and user feedback on the matching results.

The semantic matching layer is responsible for exploiting the information that can be understand from the content and meaning of the procurement documents. This process can be easily fine-tuned for the formats that are known for the system using boosting masks. This process finds the semantically related terms from procurement documents to the search query in an iterative process and grade them based on their semantic distances to the desired terms. The adjustment has been done based on the inverse geodesic distance. In future work, a learning mechanism can be used to find the optimum values and growth rates for these parameters.

Furthermore, the developed NER mechanism can be extended to a data linker system in order to disambiguate the extracted B2B Products using a B2B standard product classification system. Standard classification systems such as CPV, UNSPSC and eCl@ss can be employed to standardize the references that are used for describing goods and services in e-procurement documents. Linking the detected B2B-Product mentions using a common classification system for products and services will enable reliable and efficient search services for B2B e-marketplaces.

Business opportunity search as one of the main search scenarios provides the opportunity to evaluates the matching mechanism using large and every day growing data sets of procurement data. Public procurement portals provide a valuable resource of structured procurement documents that can be used in improving different features of the proposed solutions and also can be used for other research works in procurement. In future work the semantic matching method will be applied to procurement documents from more tendering marketplaces.

8

Bibliography

Aanen, S. S., Vandic, D., & Frasincar, F. (2015). Automated product taxonomy mapping in an e-commerce environment. *Expert Systems with Applications*, *42*(3), 1298–1313. http://doi.org/10.1016/j.eswa.2014.09.032

Ahn, J., Brusilovsky, P., Grady, J., He, D., & Florian, R. (2010). Semantic annotation based exploratory search for information analysts. *Information Processing & Management*, *46*(4), 383–402. http://doi.org/10.1016/j.ipm.2010.02.001

Ali, M. A., Shil, N. C., Nine, M. S. Q. Z., Khan, M. A. K., Mahedi, H., Ali, M. A., … Hoque, M. H. (2010). Vendor selection using fuzzy integration. *International Journal of Management Science and Engineering Management*, *5*(5), 376–382. http://doi.org/10.1080/17509653.2010.10671128

Alvarez-Rodríguez, J. M., Labra-Gayo, J. E., & De Pablos, P. O. (2014). New trends on e-Procurement applying semantic technologies: Current status and future challenges. *Computers in Industry*, *65*(5), 800–820. http://doi.org/10.1016/j.compind.2014.04.005

Alvarez, J. M., Labra, J. E., Calmeau, R., Marín, Á., & Marín, J. L. (2011a). I Nnovative Services To Ease the Access To the Public. *MeTTeG 20115th International Conference on Methodologies, Technologies and Tools Enabling E-Government*, 1–13.

Alvarez, J. M., Labra, J. E., Marin, A., & Luis Marin, J. (2011b). Semantic Methods for Reusing Linking Open Data of the European Public Procurement Notice. In *Extended Semantic Web Conference 2011 PhD Symposium*.

Alvarez, J. M., Labra, J. E., Cifuentes, F., Alor-hérnandez, G., Sánchez, C., & Luna, J. A. G. (2012). Towards a pan-european e-procurement platform to aggregate, publish and search public procurement notices powered by

linked open data: the MOLDEAS approach. *International Journal of Software Engineering and Knowledge Engineering*, *22*(3), 365–383. http://doi.org/10.1142/S0218194012400086

Benatallah, B., Hacid, M., Paik, H., Rey, C., & Toumani, F. (2006). Towards semantic-driven, flexible and scalable framework for peering and querying e-catalog communities. *Information Systems*, *31*(4–5), 266–294. http://doi.org/10.1016/j.is.2005.02.009

Beneventano, D., & Montanari, D. (2008). Ontological mappings of product catalogues. In *Ontology Matching Workshop (OM 2008) at the 7th International Semantic Web Conference* (pp. 244–249). Karlsruhe, Germany: CEUR-WS.org. http://doi.org/10.1.1.142.8596

Carmel, D., Efraty, N., Landau, G. M., Maarek, Y. S., & Mass, Y. (2002). An extension of the vector space model for querying XML documents via XML fragments. In *Proceedings SIGIR 2002 Workshop on XML and Information Retrieval* (pp. 14–25). Tampere, Finland. Retrieved from http://w3.cs.huji.ac.il/course/2002/sdbi/Papers/ir-xml/QuerybyXMLFragmentsFinal.pdf

Castells, P., Fernandez, M., & Vallet, D. (2007). An Adaptation of the Vector-Space Model for Ontology-Based Information Retrieval. *IEEE Transactions on Knowledge and Data Engineering*, *19*(2), 261–272. http://doi.org/10.1109/TKDE.2007.22

Chen, D., Li, X., Liang, Y., & Zhang, J. (2010a). A semantic query approach to personalized e-Catalogs service system. *Journal of Theoretical and Applied Electronic Commerce Research*, *5*(3), 39–54. http://doi.org/10.4067/S0718-18762010000300005

Chen, D., Li, X., Liang, Y., & Zhang, J. (2010b). Research on the Theory of Customer-Oriented E-Catalog Ontology Automatic Construction. In *2010 International Conference on E-Business and E-Government* (pp. 2961–2964). Guangzhou: Ieee. http://doi.org/10.1109/ICEE.2010.748

Chen, D., Li, X., & Zhang, J. (2010). User-oriented intelligent service of e-catalog based on semantic web. In *2010 The 2nd IEEE International Conference on Information Management and Engineering (ICIME)* (pp. 449–453). Chengdu: Ieee. http://doi.org/10.1109/ICIME.2010.5477872

Chen, L., Wang, F., Qi, L., & Liang, F. (2014). Experiment on sentiment embedded comparison interface. *Knowledge-Based Systems*, *64*, 44–58. http://doi.org/10.1016/j.knosys.2014.03.020

Council, T. H. E. (2002). REGULATION (EC) No. 2150/2002 OF THE EUROPEAN PARLIAMENT AND THE COUNCIL of November 2002 on waste statistics, *2002*(29), 1–562.

Distinto, I., D'Aquin, M., & Motta, E. (2016). LOTED2: An ontology of European

public procurement notices. *Semantic Web*, *7*(3), 267–293. http://doi.org/10.3233/SW-140151

Dorn, J., Grun, Ch., Werthner, H., & Zapletal, M. (2009). From business to software : a B2B survey, 123–142. http://doi.org/10.1007/s10257-008-0082-4

Du, T. C. (2009). Building an automatic e-tendering system on the Semantic Web. *Decision Support Systems*, *47*(1), 13–21. http://doi.org/10.1016/j.dss.2008.12.009

Eckhardt, A., Hreško, J., Procházka, J., & Smrſ, O. (2014). Entity linking based on the co-occurrence graph and entity probability. In *Proceedings of the first international workshop on Entity recognition & disambiguation - ERD '14* (pp. 37–44). New York, New York, USA: ACM Press. http://doi.org/10.1145/2633211.2634349

Elahi, A., & Rostami, A. (2012). Concept-based vector space model for improving text clustering. *Journal of Advanced Computer Science …*, *2*(3), 140–158. Retrieved from http://www.sign-ific-ance.co.uk/dsr/index.php/JACSTR/article/view/252

Esteban, G. (2015). Using the Semantic Web for the Integration and Publication of Public Procurement Data, *2*, 13–28. http://doi.org/10.1007/978-3-319-22389-6_2

European Commission. (2007). Commission Regulation (EC) No 213/2008. *Official Journal of the European Union*, *L 74*(1), 375.

European Dynamics SA. (2007). Electronic Catalogues in Electronic Public Procurement. In *DG Internal Market of the European Commission*. Marousi: Ã‚Â© European Communities.

Bengfort, B. (2012). A Survey of Stochastic and Gazetteer Based Approaches for Named Entity Recognition – Part 2 Approaches to Named Entity Recognition 2) Learning and Stochastic Approaches, (7), 1–9.

Fang Luo, Qizhi Qiu, & QianXing Xiong. (2011). Introduction to the product-entity recognition task. In *2011 3rd Symposium on Web Society* (pp. 122–126). IEEE. http://doi.org/10.1109/SWS.2011.6101282

Ghimire, S., Jardim-Goncalves, R., & Grilo, A. (2013). Framework for catalogues matching in procurement e-marketplaces. In *Information Systems and Technologies (CISTI), 8th Iberian Conference on* (pp. 1–6). Lisbon, Portugal: IEEE.

Ghimire, S., Jardim-Goncalves, R., Grilo, A., & Beca, M. (2013). Framework for inter-operative e-Procurement marketplace. In *Proceedings of the 2013 IEEE 17th International Conference on Computer Supported Cooperative Work in Design (CSCWD)* (pp. 459–464). Whistler, BC: IEEE. http://doi.org/10.1109/CSCWD.2013.6581006

Graux, A. H., Kronenburg, T., & August, P. (2012). *State of Play : Re-use of Public Procurement Data*.

Grilo, A., Ghimire, S., & Jardim-Goncalves, R. (2013). Cloud-Marketplace: New paradigm for e-marketplaces. In *Technology Management in the IT-Driven Services (PICMET), 2013 Proceedings of PICMET '13:* (pp. 555–561). San Jose, CA: IEEE. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6641817

Grilo, A., & Jardim-goncalves, R. (2013a). E-Marketplaces : A New Approach. In *International Proceedings of Economics Development and Research 59* (pp. 79–83). Singapore: IACSIT Press. http://doi.org/10.7763/IPEDR.2013.V59.17

Grilo, A., & Jardim-Goncalves, R. (2013b). Cloud-Marketplaces: Distributed e-procurement for the AEC sector. *Advanced Engineering Informatics*, *27*(2), 160–172. http://doi.org/10.1016/j.aei.2012.10.004

Grilo, A., Jardim-Goncalves, R., & Ghimire, S. (2013). E-Procurement in the Era of Cloud Computing. In J. Blooma (Ed.), *Proceedings of the 4th International Conference on Information Systems Management and Evaluation (Icime 2013)* (pp. 104–110). Ho Chi Minh City, Vietnam.

Guo, J. (2009). *Collaborative conceptualisation: towards a conceptual foundation of interoperable electronic product catalogue system design*. *Enterprise Information Systems* (Vol. 3). http://doi.org/10.1080/17517570802610362

Guo, J., & An, R. (2014). A Case Study on E-marketplace Basic Functions. *The Fourth International Conference on Business Intelligence and Technology, BUSTECH 2014*, (c), 25–30.

Hepp, M. (2008). GoodRelations: An Ontology for Describing Products and Services Offers on the Web. *Knowledge Engineering: Practice and Patterns*, *5268 LNAI*, 329–346. http://doi.org/10.1007/978-3-540-87696-0_29

Hepp, M., Leukel, J., & Schmitz, V. (2005). A quantitative analysis of eCl@ss, UNSPSC, eOTD, and RNTD: Content, coverage, and maintenance. *Proceedings - ICEBE 2005: IEEE International Conference on E-Business Engineering*, *2005*, 572–581. http://doi.org/10.1109/ICEBE.2005.15

Huang, J. Z., Feilong Tang, Yunming Ye, Huang, G., & Minglu Li. (2005). Ontology-based e-catalog matching for integration of GDSN and EPCglobal network. In *IEEE International Conference on e-Business Engineering (ICEBE'05)* (pp. 212–215). Beijing: IEEE. http://doi.org/10.1109/ICEBE.2005.92

Huang, S.-L., & Lin, C.-Y. (2010). The search for potentially interesting products in an e-marketplace: An agent-to-agent argumentation approach. *Expert Systems with Applications*, *37*(6), 4468–4478. http://doi.org/10.1016/j.eswa.2009.12.064

Icf - Ghk. (2014). *SMEs' access to public procurement markets and aggregation of*

demand in the EU. A study commissioned by the European commission, DG Internal Market and Services.

Interagency Procurement, & Working Group (IAPWG). (2006). *UN Procurement Practitioner's Handbook*. *IAPWG*.

Jap, S. D. (2007). The Impact of Online Reverse Auction Design on Buyer–Supplier Relationships. *Journal of Marketing*, *71*(1), 146–159. http://doi.org/10.1509/jmkg.71.1.146

Julashokri, M., Fathian, M., Gholamian, M. R., & Mehrbod, A. (2011). Improving Recommender System's Efficiency Using Time Context and Group Preferences. *INTERNATIONAL JOURNAL ON Advances in Information Sciences and Service Sciences*, *3*(4), 162–168. http://doi.org/10.4156/aiss.vol3.issue4.20

Kajan, E. (2012). *Handbook of Research on E-Business Standards and Protocols: Documents, Data and Advanced Web Technologies*. IGI Global.

Kajan, E., Dorloff, F.-D., & Bedini, I. (2012). *Handbook of Research on E-Business Standards and Protocols: Documents, Data and Advanced Web Technologies*. PA, USA: IGI Publishing Hershey.

Kannan, A., Givoni, I., Agrawal, R., & Fuxman, A. (2011). Matching Unstructured Product Offers to Structured Product Specifications Categories and Subject Descriptors. In *KDD '11 Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 404–412). Manchester Grand Hyatt, San Diego, CA: ACM. http://doi.org/10.1145/2020408.2020474

Kaptein, M., & Parvinen, P. (2015). Advancing E-Commerce Personalization: Process Framework and Case Study. *Kaptein,Maurits Parvinen,Petri* , *19*(March), 7–33. http://doi.org/10.1080/10864415.2015.1000216

Khormuji, M. K. (2014). Persian Named Entity Recognition based with Local Filters, *100*(4), 1–6.

Kim, D., Kim, J., & Lee, S. G. (2002). Catalog integration for electronic commerce through category-hierarchy merging technique. In *Proceedings Twelfth International Workshop on Research Issues in Data Engineering: Engineering E-Commerce/E-Business Systems RIDE-2EC 2002* (pp. 28–33). San Jose, CA: IEEE. http://doi.org/10.1109/RIDE.2002.995095

Kim, W., Choi, D. W., & Park, S. (2007). Agent based intelligent search framework for product information using ontology mapping. *Journal of Intelligent Information Systems*, *30*(3), 227–247. http://doi.org/10.1007/s10844-006-0026-8

Kwon, I.-H., Kim, C. O., Kim, K. P., & Kwak, C. (2008). Recommendation of e-commerce sites by matching category-based buyer query and product e-catalogs. *Computers in Industry*, *59*(4), 380–394.

http://doi.org/10.1016/j.compind.2007.10.002

Lampathaki, F., Mouzakitis, S., Gionis, G., Charalabidis, Y., & Askounis, D. (2009). Business to business interoperability: A current review of XML data integration standards. *Computer Standards & Interfaces*, *31*(6), 1045–1055. http://doi.org/10.1016/j.csi.2008.12.006

Lee, J., Lee, T., Lee, S., Jeong, O., & Lee, S. (2007). Massive Catalog Index based Search for e-Catalog Matching. In *The 9th IEEE International Conference on E-Commerce Technology and The 4th IEEE International Conference on Enterprise Computing, E-Commerce and E-Services (CEC-EEE 2007)* (pp. 341–348). Tokyo: IEEE. http://doi.org/10.1109/CEC-EEE.2007.64

Lee, T., Lee, I., Lee, S., Lee, S., Kim, D., Chun, J., … Shim, J. (2006). Building an operational product ontology system. *Electronic Commerce Research and Applications*, *5*(1), 16–28. http://doi.org/10.1016/j.elerap.2005.08.005

Leukel, J., Schmitz, V., & Dorloff, F. (2002). Exchange of Catalog Data in B2B Relationships-Analysis and Improvement. In *IADIS International Conference WWW/Internet 2002 (ICWI 2002)* (Vol. 2002, pp. 403–410). Lisbon, Portugal: Springer. Retrieved from http://pdf.aminer.org/000/259/144/exchange_of_catalog_data_in_b_b_relationships_analysis_and.pdf

Lipczak, M., Koushkestani, A., & Milios, E. (2014). Tulip : Lightweight Entity Recognition and Disambiguation Using Wikipedia-Based Topic Centroids. In *Proceedings of the first international workshop on Entity recognition & disambiguation - ERD '14*.

Liu, D.-R., Lin, Y.-J., Chen, C.-M., & Huang, Y.-W. (2001). Deployment of personalized e-catalogues: An agent-based framework integrated with XML metadata and user models. *Journal of Network and Computer Applications*, *24*(3), 201–228. http://doi.org/10.1006/jnca.2001.0132

Llorens, H., Saquete, E., & Navarro-Colorado, B. (2013). Applying semantic knowledge to the automatic processing of temporal expressions and events in natural language. *Information Processing & Management*, *49*(1), 179–197. http://doi.org/10.1016/j.ipm.2012.05.005

Luo, F., Xiao, H., & Chang, W. (2011). Product Named Entity Recognition Using Conditional Random Fields. *2011 Fourth International Conference on Business Intelligence and Financial Engineering*, 86–89. http://doi.org/10.1109/BIFE.2011.101

Lysons, K., & Farrington, B. (2006). *Purchasing and Supply Chain Management*. (K. Lysons, Ed.) (7th ed.). Financal Times Management.

Management Association, I. R. (2013). *Assistive Technologies: Concepts, Methodologies, Tools, and Applications*. *Assistive Technologies: Concepts, Methodologies, Tools, and Applications*. IGI Global.

Manning, C. D., Raghavan, P., Schütze, H., Prabhakar, R., & Schutze, H. (2008). *Introduction to Information Retrieval*. New York, USA: Cambridge University Press.

Manning, C. D., Raghavan, P., & Schutze, H. (2008). Evaluation in information retrieval. In *Introduction to Information Retrieval* (pp. 139–161). Cambridge: Cambridge University Press. http://doi.org/10.1017/CBO9780511809071.009

Marrero, M., Urbano, J., Sánchez-Cuadrado, S., Morato, J., & Gómez-Berbís, J. M. (2012). Named Entity Recognition: Fallacies, challenges and opportunities. *Computer Standards & Interfaces*, *35*(5), 482–489. http://doi.org/10.1016/j.csi.2012.09.004

Mehrbod, A., Zutshi, A., & Grilo, A. (2014a). A Vector Space Model Approach for Searching and Matching Product E-Catalogues. In J. Xu, V. A. Cruz-Machado, B. Lev, & S. Nickel (Eds.), *Proceedings of the Eighth International Conference on Management Science and Engineering Management* (Vol. 281). Lisbon, Portugal: Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-642-55122-2

Mehrbod, A., Zutshi, A., & Grilo, A. (2014b). Semantic and Syntactic Matching of e-Catalogues using Vector Space Model. In *Proceedings of the 11th International Conference on e-Business* (pp. 224–229). Vienna - Austria. http://doi.org/10.5220/0005115302240229

Mehrbod, A., Zutshi, A., Grilo, A., & Cruz-Machado, V. (2017). Evaluation of an E-catalogue Matching Mechanism in Public Procurement Notice Search. In J. Xu, A. Hajiyev, S. Nickel, & M. Gen (Eds.), *Proceedings of the Tenth International Conference on Management Science and Engineering Management* (pp. 1237–1247). Singapore: Springer Singapore. http://doi.org/10.1007/978-981-10-1837-4_101

Mehrbod, A., Zutshi, A., Grilo, A., & Jardim-Goncalves, R. (2015). Matching Heterogeneous e-Catalogues in B2B Marketplaces Using Vector Space Model. *International Journal of Computer Integrated Manufacturing, (IJCIM)*, *2*(Factories of the Future (FOF)). http://doi.org/10.1080/0951192x.2015.1107915

Melli, G., & Romming, C. (2012). An overview of the CPROD1 contest on consumer product recognition within user generated postings and normalization against a large product catalog. *Proceedings - 12th IEEE International Conference on Data Mining Workshops, ICDMW 2012*, 861–864. http://doi.org/10.1109/ICDMW.2012.104

Miyamoto, M. (2015). Application of competitive forces in the business intelligence of Japanese SMEs. *International Journal of Management Science and Engineering Management*, *10*(4), 273–287. http://doi.org/10.1080/17509653.2014.966794

Molander, P. (2014). PUBLIC PROCUREMENT IN THE EUROPEAN UNION: THE CASE FOR NATIONAL THRESHOLD VALUES. *JOURNAL OF PUBLIC PROCUREMENT*, *14*(2), 181–214.

Mukerjee, K., Porter, T., & Gherman, S. (2011). Linear scale semantic mining algorithms in microsoft SQL server's semantic platform. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '11* (p. 213). New York, New York, USA: ACM Press. http://doi.org/10.1145/2020408.2020447

Muñoz-soro, J. F., Esteban, G., Corcho, O., & Serón, F. (2016). PPROC , an Ontology for Transparency in Public Procurement. *Semantic Web*, *7: (3)*(Semantic Web for the Legal Domain), 295–309. http://doi.org/10.3233/SW-150195

Mynarz, J., Svátek, V., & Di Noia, T. (2015). Matchmaking Public Procurement Linked Open Data (Vol. 6428, pp. 405–422). http://doi.org/10.1007/978-3-319-26148-5_27

Nečaský, M., Klímek, J., Mynarz, J., Knap, Tt., Svátek, V., & Stárka, J. (2014). Linked data support for filing public contracts. *Computers in Industry*, *65*(5), 862–877. http://doi.org/10.1016/j.compind.2013.12.006

Obrst, L. (2003). Ontologies for semantically interoperable systems. In *Proceedings of the twelfth international conference on Information and knowledge management - CIKM '03* (p. 366). New York, New York, USA: ACM Press. http://doi.org/10.1145/956863.956932

Ordóñez de Pablos, P. (2012). *E-procurement Management for Successful Electronic Government systems*. IGI Global.

Pearcy, D. H., Parker, D. B., & Giunipero, L. C. (2008). Using Electronic Procurement to Facilitate Supply Chain Integration: An Exploratory Study of US-based Firms. *American Journal of Business*, *23*(1), 23–36. http://doi.org/10.1108/19355181200800002

Pedersen, K. V., Thomassen, G. W., Hoddevik, A., & Ciciriello, C. (2012). *PEPPOL Final Report*. Oslo; Agency for Public Management and eGovernment (Difi).

Piccinno, F., Ferragina, P., & Informatica, D. (2014). From TagME to WAT : a new Entity Annotator Categories and Subject Descriptors. *Proceedings of the First International Workshop on Entity Recognition & Disambiguation - ERD '14*.

Powers, D. M. W. (2011). Evaluation: From Precision, Recall and F-Measure to ROC., Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies*, *2*(1), 2229–3981. Retrieved from http://www.bioinfo.in/contents.php?id=51

Procházka, A., & Smrž, O. (2014). Entity Recognition Based on the Co-

occurrence Graph and Entity Probability. In *ERDC '2014* (pp. 37–44). Gold Coast, Australia: ACM. Retrieved from http://web-ngram.research.microsoft.com/ERD2014/Docs/submissions/erd14_submission_3.pdf

Putthividhya, D. P., Hu, J., & Ave, H. (2011). Bootstrapped named entity recognition for product attribute extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 1557–1567). http://doi.org/10.1016/j.pss.2010.03.014

Ramkumar, M., & Jenamani, M. (2012). E-procurement Service Provider Selection---An Analytic Network Process-Based Group Decision-Making Approach. *Service Science*, *4*(3), 269–294. http://doi.org/10.1287/serv.1120.0024

Roman, A. V. (2013). Public Policy and Financial Managment Through E-procurement: A Practice Oriented Normative Model For Maximaizing Transformative Impacts. *JOURNAL OF PUBLIC PROCUREMENT*, *13*(4), 447–475.

Schmitz, V., Leukel, J., & Dorloff, F. (2005). Do E-Catalog Standards Support Advanced Processes in B2B E-Commerce? Findings from the CEN/ISSS Workshop eCAT. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS-38 2005)* (Vol. 0, p. 162c–162c). Big Island, HI, USA: IEEE. http://doi.org/10.1109/HICSS.2005.209

Stolz, A., Rodriguez-Castro, B., Radinger, A., & Hepp, M. (2014). PCS2OWL: A Generic Approach for Deriving Web Ontologies from Product Classification Systems. In V. Presutti, C. D'Amato, F. Gandon, M. D'Aquin, S. Staab, & A. Tordai (Eds.), *The Semantic Web: Trends and Challenges SE - 43* (Vol. 8465, pp. 644–658). Springer International Publishing. http://doi.org/10.1007/978-3-319-07443-6_43

Tadelis, S. (2012). Public procurement design: Lessons from the private sector. *International Journal of Industrial Organization*, *30*(3), 297–302. http://doi.org/10.1016/j.ijindorg.2012.02.002

Teixeira, J., Sarmento, L., & Oliveira, E. (2011). A bootstrapping approach for training a ner with conditional random fields. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *7026 LNAI*(from 1997), 664–678. http://doi.org/10.1007/978-3-642-24769-9_48

Tekli, J., Chbeir, R., & Yetongnon, K. (2009). An overview on XML similarity: Background, current trends and future directions. *Computer Science Review*, *3*(3), 151–173. http://doi.org/10.1016/j.cosrev.2009.03.001

Toh, Z., Wang, W., Lan, M., & Li, X. (2012). An NER-based product identification and lucene-based product linking approach to CPROD1

challenge: Description of submission system to CPROD1 Challenge. *Proceedings - 12th IEEE International Conference on Data Mining Workshops, ICDMW 2012*, 869–871. http://doi.org/10.1109/ICDMW.2012.66

Turney, P., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, *37*, 141–188.

Urizar, M. (2013). *The Project Manager's Checklist for Building Projects*. Xlibris Corporation.

Valle, F., D'Aquin, M., Di Noia, T., & Motta, E. (2010). LOTED: Exploiting linked data in analyzing European Procurement notices. *CEUR Workshop Proceedings*, *631*, 52–63.

Vandic, D., & Milea, V. (2014). Semantic Web-Based Product Search. In J. Parsons & D. Chiu (Eds.), *Advances in Conceptual Modeling SE - 17* (Vol. 8697, pp. 150–159). Springer International Publishing. http://doi.org/10.1007/978-3-319-14139-8_17

Vandic, D., Nederstigt, L., & Aanen, S. (2014). Ontology Population from Web Product Information. In M. Indulska & S. Purao (Eds.), *Advances in Conceptual Modeling SE - 28* (Vol. 8823, pp. 263–272). Springer International Publishing. http://doi.org/10.1007/978-3-319-12256-4_28

Vandic, D., van Dam, J.-W., & Frasincar, F. (2012). Faceted product search powered by the Semantic Web. *Decision Support Systems*, *53*(3), 425–437. http://doi.org/10.1016/j.dss.2012.02.010

Vieira, H., da Silva, A., Cristo, M., & de Moura, E. (2015). A Self-training CRF Method for Recognizing Product Model Mentions in Web Forums. In A. Hanbury, G. Kazai, A. Rauber, & N. Fuhr (Eds.), *Advances in Information Retrieval SE - 27* (Vol. 9022, pp. 257–264). Springer International Publishing. http://doi.org/10.1007/978-3-319-16354-3_27

Wang, S., & Archer, N. (2007). Business-to-business collaboration through electronic marketplaces : An exploratory study $. *Journal of Purchasing & Supply Management*, *13*, 113–126. http://doi.org/10.1016/j.pursup.2007.05.004

Wei, W., Barnaghi, P. M., & Bargiela, A. (2008). Search with Meanings : An Overview of Semantic Search Systems. *Int. J. Communications of SIWN*, *3*, 76–82.

Widdows, D. (2008). Semantic vector products: Some initial investigations. *Second AAAI Symposium on Quantum Interaction*, (March).

Widdows, D., & Ferraro, K. (2008). Semantic Vectors: a Scalable Open Source Package and Online Technology Management Application. *LREC*, 1183–1190. Retrieved from http://repository.dlsi.ua.es/242/1/pdf/300_paper.pdf

Wu, B., Cheng, X., Wang, Y., Guo, Y., & Song, L. (2009). Simultaneous Product Attribute Name and Value Extraction from Web Pages. *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, 295–298. http://doi.org/10.1109/WI-IAT.2009.286

Wu, S., Fang, Z., & Tang, J. (2012). Accurate product name recognition from user generated content. *Proceedings - 12th IEEE International Conference on Data Mining Workshops, ICDMW 2012*, 874–877. http://doi.org/10.1109/ICDMW.2012.129

Yen, B., & Kong, R. (2002). Personalization of information access for electronic catalogs on the web. *Electronic Commerce Research and Applications*, *1*, 20–40.

Zhang, L. (2009). A Framework for an Ontology-based E-commerce Product Information Retrieval System, *4*(6), 436–443.

Zhang, Y., & Bhattacharyya, S. (2008). Analysis of B2B e-marketplaces : an operations perspective, 235–256. http://doi.org/10.1007/s10257-008-0096-y