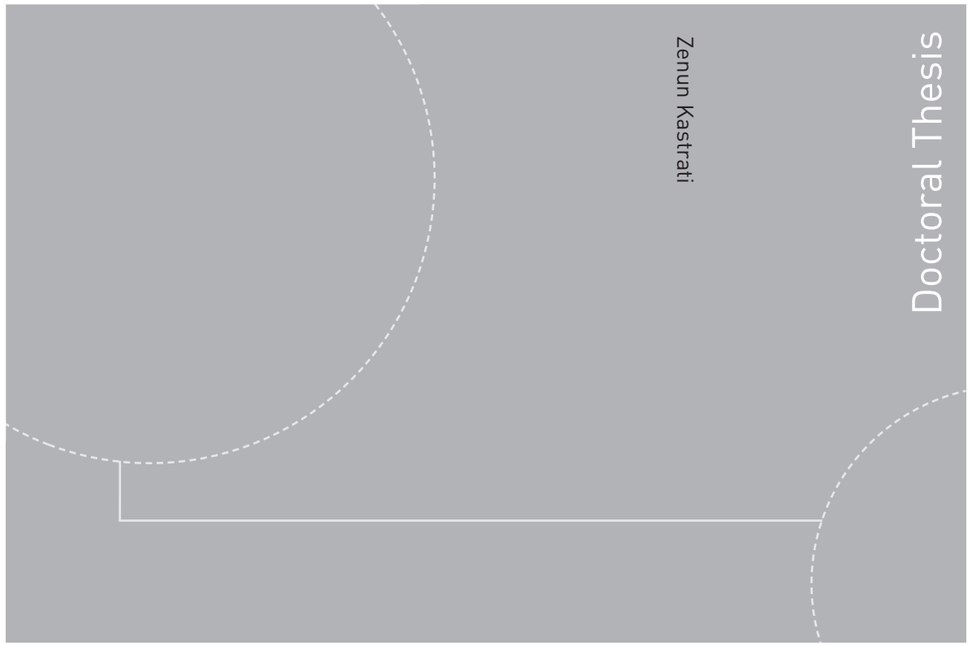
ISBN 978-82-326-2880-3 (printed version)
ISBN 978-82-326-2881-0 (electronic version)
ISSN 1503-8181



Doctoral theses at NTNU, 2018:44

Zenun Kastrati

Improving Document Classification Using Ontologies

NTNU
Norwegian University of
Science and Technology
Faculty of Information Technology
and Electrical Engineering
Department of Computer Science

NTNU
Norwegian University of
Science and Technology

Doctoral theses at NTNU, 2018:44

NTNU

NTNU
Norwegian University of
Science and Technology

Zenun Kastrati

Improving Document Classification Using Ontologies

Thesis for the degree of Philosophiae Doctor

Gjøvik, February 2018

Norwegian University of Science and Technology
Faculty of Information Technology
and Electrical Engineering
Department of Computer Science

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of Science and Technology

Thesis for the degree of Philosophiae Doctor

Faculty of Information Technology
and Electrical Engineering
Department of Computer Science

© Zenun Kastrati

ISBN 978-82-326-2880-3 (printed version)
ISBN 978-82-326-2881-0 (electronic version)
ISSN 1503-8181

Doctoral theses at NTNU, 2018:44

Printed by Skipnes Kommunikasjon as

Improving Document Classification Using Ontologies

Faculty of Information Technology and Electrical Engineering
Norwegian University of Science and Technology

Declaration of Authorship

I, Zenun Kastrati, hereby declare that this thesis and the work presented in it is entirely my own. Where I have consulted the work of others, this is always clearly stated.

Signed:

(Zenun Kastrati)

Date:

Summary

We are living in the age of internet where massive amount of information is produced from various digital resources on daily basis. The information of these resources is typically stored in unstructured textual format such as reports, news, e-mails, blogs, etc., therefore, a proper classification and organization of this huge amount of information is apparently needed. In this regard, an automatic classification, particularly ontology-based classification, plays an important role in helping people to classify and organize the information accordingly. The ontology-based classification system is an automatic system that utilizes the ontology in order to take advantages of organizing and classifying the knowledge in a more structural and formal way, thus providing better classification accuracy comparing to the traditional keyword-based classification system.

The performance of an ontology-based document classification system can be affected by several aspects involved in the entire classification process that generally is constituted of steps such as document collection and preprocessing, document representation, dimensionality reduction, and the classifier. It is almost impossible to address all these research aspects in order to obtain performance improvement in a single dissertation research work, therefore we selected to work on the aspects that we consider are either rarely studied or have a crucial role on the ontology-based classification system.

Document representation is one of the main aspects that affects the performance of ontology-based document classification, thus the first research aspect that we investigated is enriching document representation with semantics utilizing the background knowledge exploited by ontologies. The background knowledge derived from an ontology is embedded in a document using a matching technique. The idea behind this technique is mapping of terms that occur in a document with the relevant ontology concepts by searching only the presence of concepts labels in that document. Searching only the presence of concepts labels occurring in a document limits the capabilities of the classification system to capture and exploit the entire conceptualization involved in that document due to the semantic gap issue, the lack of an in depth-coverage of concepts, and the ambiguity problem. In this thesis, the focus is placed on the conceptual document representation, in which, a document is associated with a set of concepts not only by looking for the appearance of concept labels, but also through the acquisition of lexical information integrated (linked) to the ontology to enriching its coverage with new concepts. In this respect, an automatic ontology concept enrichment model is developed to enrich ontologies with new concepts in order to provide a broader coverage for document representation. The proposed model explores textual data and relies on semantic and contextual information of terms occurring in a discourse.

The performance of ontology-based document classification is highly dependent on the relevance of concepts that is indicated by weights. The weights reflect the discriminative power of concepts with respect to the documents and are typically computed through the frequency of occurrences of concepts in these documents. Thus, the second research aspect that we studied in this research work is enhancing the existing concept weighting scheme by introducing the notion of concept importance. Concept importance assesses the contribution of a concept in discriminating between documents depending on its position in the ontology hierarchy. In addition, we explored the possibilities to automatically evaluate the concept importance and a Markov-based approach is developed. Further, we aggregated concept importance and concept relevance in order to enhance the concept weighting

scheme and thus to improve the concept vector space representation model.

Lastly, the third research aspect studied in this dissertation is related to improving classification accuracy by taking the advantages of the ontology enrichment model, and the enhanced concept weighting scheme developed while studying the first and the second research aspect respectively. We proposed a document classification approach that relies on an ontology whose coverage is widen using the ontology enrichment model SEMCON and the weights of concepts are assessed through the new concept weighting technique composed of concept relevance and concept importance. Extensive experimental results demonstrated a considerable improvement of the classification effectiveness.

Acknowledgments

I would like to express my sincere appreciation to my main supervisor, *Assoc. Prof. Dr. Sule Yildirim Yayilgan* for her invaluable advices, guidance, and positive disposition throughout my PhD work. I also would like to extend my sincere gratitude to my co-supervisor, *Prof. Dr. Rune Hjelsvold* for his invaluable advices and generous support. They have always provided their expertise, dedication, and invaluable time whenever I needed them. I truly appreciate and they will always have my sincere admiration and respect. I also would like to thank the members of PhD review committee for their suggestions and valuable comments.

I would like to acknowledge the Academic Exchange for Progress (AEP) project for financial support during my research work in the Norwegian University of Science and Technology. The AEP project was part of the Higher Education, Research, and Development (HERD) program financed by the Ministry of Foreign Affairs of Norway.

I would like to thank all of the Norwegian University of Science and Technology members for creating a spirit and environment of cooperation and research. A special thanks to *Dr. Ali Shariq Imran* for his suggestions, motivations, and continuous support and cooperation. I am also thankful to *Dr. Stein Runar Olsen* for his generous support on administrative stuff throughout my doctoral studies.

Last but not least, a special thanks to my friends and family who have accompanied me throughout the long PhD journey and have created a relaxing environment and lots of motivation with their incredible patience and encouragement.

Contents

List of Figures

List of Tables

Part I

Introduction

This part presents an overview of this thesis. In Chapter 1, we provide some background information on ontology-based classification, and then we discuss our motivations and research questions that have been addressed in this thesis. Chapter 2 investigates the state of the art in context to the research aspects studied in this thesis. Our contributions and their connections to the published articles are presented in Chapter 3.

Chapter 1

Introduction

This chapter presents a synopsis of the research work done during my PhD. First, it provides some background information about ontology-based classification and a discussion about our motivations. Next, it outlines the objective and research questions followed by the list of published articles and their connections to the research questions. Finally, it depicts the structure of this thesis.

1.1 Background and Motivation

Nowadays, the web is the main source of information which is growing rapidly. This information is coming from various unstructured resources in forms of emails, reports, news, blogs, views, among others. Accessing and retrieving massive amount of such resources is not an easy task, therefore, the need of an automatic retrieval of useful knowledge in order to assist the human analysis is apparent. Automatic text document classification plays a vital role in this regard.

Automatic text document classification also known as text categorization is the process of automatically assigning a text document from a given domain to one or more class labels from a finite set of predefined categories.

The first step in text document classification process is the preprocessing where a document has to be converted from a full text version to a document as a vector of features. The vector space model (VSM) is one of the simplest and most common models for representing documents and is widely used in document classification [75]. In this model, a document is typically represented as bag of words where each word/term is represented as a dimension in a vector space and independent to other terms in the same document. Numeric values are assigned to each term in order to show the relevance of that term for distinguishing a document from the other documents. To compute these numeric values, the vector space model uses the *tf***idf* weighting scheme which is composed of two main factors: 1) the document specific statistic factor called term frequency *tf*, which shows the importance of a term in a particular document, and 2) the global statistic factor called inverse document frequency *idf*, which indicates how widely a term is distributed over the collection of documents.

Even though vector space document representation model has proved to be very simple and commonly used in the domain of text classification, it however has some limitations [6, 7, 150]. The main limitation of this model is the ignorance of dependencies of terms, i.e. grammatical relations, their hierarchical structure, i.e. taxonomic and non-taxonomic, and ordering in a text document [93, 94]. In order to address the limitations of VSM, several approaches have been proposed in the literature. These approaches make a step away from string literals (keyword) based representation towards meaning (semantic) based document representation. This content based document representation is achieved by using the background knowledge constructed from thesaurus or ontologies [14, 88, 104, 137, 151]. In particular, WordNet [44] has been widely used to construct the background knowledge for improving representation of documents [64, 78, 88]. WordNet is a lexical database which groups terms into set of synonyms called synsets and each of these sets is considered one concept. These concepts are linked together using semantic relationships such as meronyms, hypernyms, hyponyms, synonyms, etc. Accordingly, this background

knowledge gathered from WordNet is utilized as a means to enhance the representation of documents.

Wikipedia is also another background knowledge resource which has been used to improve document representation [49, 78, 99, 151]. Wikipedia describes concepts by articles and each concept belongs to at least one category. Semantic relations between concepts, namely, synonymy (equivalence), hyponymy/hypernymy (hierarchical), and associative, are described via hyperlinks used between articles. Wikipedia also deals with polysemous concepts. It provides disambiguation pages which cover all possible meanings associated with the corresponding concept. Therefore, this background knowledge derived from Wikipedia is used to add semantic information into documents, and thus to enhance documents representation.

Ontologies have recently been emerged as a means which takes advantages of organizing the knowledge in a more structural and formal way comparing to knowledge organized in thesaurus (WordNet, Wikipedia). Ontologies consist of a set of concepts and relations which link these concepts. Relations of ontologies are composed of taxonomy relations and non-taxonomy relations. The taxonomy relation of concepts forms a hierarchical tree-based structure and it is indicated as *is-a* relation. Non-taxonomy relation represents partonomy (part-whole) relation which divides concept as a whole into different parts, e.g. *Norway* and *Sweden* are part of *Europe*. This wide coverage of concepts and relationships provided by ontologies is the most common background knowledge used in literature [15, 102, 130, 131, 154] to embed the semantics into documents in order to improve documents representation.

The background knowledge derived from an ontology or thesaurus is embedded in a document using a matching strategy. This strategy maps terms occurring in a document with relevant concepts. This term to concept mapping is an exact matching which is achieved by searching only the concepts (concepts labels) that explicitly occur in a document [150]. The relevance of concepts is indicated by weights assigned to them (concepts). These weights reflect the discriminative power of concepts with respect to the documents and are computed automatically using the frequency of occurrences of concepts in each document.

The next step involved by text document classification is employing of one of the techniques from machine learning to train a classifier and generate a predictive model for classifying the unlabelled documents.

The final step of the document classification process is classifying of new unlabelled documents into the appropriate categories. To achieve this task, unclassified documents are primarily represented as concept vectors which are then used as input to the predictive model built by the machine learning algorithms to classify the documents appropriately.

Searching only the presence of concept labels provides limited capabilities for capturing and exploiting the whole conceptualization involved in user information and content meanings. These capabilities limitations occur due to the following reasons: 1) semantic gap issue - the domain ontology used for indexing may have different focus (intention) and does not cover parts (content aspects) of the document that are of interest to some users [11], 2) an in depth coverage of concepts is often not available, and 3) term to concept matching is a many to many mapping due to linguistic forms such as polysemy and synonymy [46] and this phenomenon leads to an ambiguity issue.

The weighting scheme employed by the existing approaches represents also a limitation for conceptual document representation. The weights of concepts are basically computed using the *tf*idf* weighting scheme which reflects only the relevance of concepts [24, 45]. It is a well-known fact that some concepts are better at discriminating between documents than the others, which means that various concepts of an ontology contribute differently. The contribution depends on the position/location of concepts in the ontology hierarchy and it is indicated by its corresponding importance.

In this research, emphasis is given to the conceptual document representation where a

document is associated with a set of concepts not only by checking the presence of concept labels, but also by identifying and extracting lexical information attached to the ontologies. The need for integrating (attaching) ontologies and lexical information is a main issue for the next generation tools envisaged by the semantic web [119] and depending on the final result we intend to achieve, the integration can be used to enrich the coverage of an ontology or to build a system which covers properties of an ontology and a lexical information [110]. In this research, we aim to link ontologies and lexical information in order to enrich the coverage of an ontology with new concepts and therefore an emphasis is placed on exploring and analysing the approaches and techniques to deal with ontology enrichment issue. In this regard, we pay close attention to the strategies and techniques to deal with the word sense disambiguation problem which occurs due to the fact that a term may appear within more than one concept labels. In this thesis, we are also looking at models for enhancing the existing concept weighting scheme using the concept importance which provides the capabilities to reflect the contribution of concepts in an ontology.

1.2 Objective and Research Questions

This research work aims to explore new methods and techniques for conceptual representation of documents by using ontologies as a background knowledge in order to improve text document classification. So with this background, we formulated the main objective as follows:

> **Main Research Objective:** *Improve text document classification effectiveness using ontologies*.
>
> The main goal of this research is to improve text document classification performance in terms of effectiveness using ontologies as background knowledge. Ontologies provide a broad coverage of concepts which enable to derive the semantic representation of documents, therefore, we aim to utilise concepts of ontologies as a means to improve document classification effectiveness.
>
> To achieve this main objective, the theme of the research is divided into two sub-objectives:
>
> **Objective 1:** *Explore and analyse new strategies and techniques for enriching ontologies with new concepts*.
>
> The objective is to explore new possibilities for conceptual document representation in order to overcome the limitations of state of the art document representation techniques (recall Section 1.1). Moreover, the focus of this objective is to explore and develop new strategies and techniques for identification and extraction of lexical information attached to the ontology concepts, which, in literature, is referred as ontology concept enrichment.
>
> **Objective 2:** *Explore and improve concept vectors with new concept weighting scheme capable to consider and assess automatically the importance of concepts of the ontology*.
>
> The core of this objective is to investigate the contribution of ontology concepts in terms of concept's discriminating power. This contribution is represented by concept importance. Furthermore, an emphasis is placed on approaches to compute automatically concept importance which then can be aggregated with concept relevance in order to enhance the concept weighting scheme used in concept vectors.

According to the objectives above, the following research questions were raised:

> **Research Question 1:** *Can the text document classification effectiveness be improved by enriching the ontologies used as the background knowledge*?

This research question is the core of this research work and the main challenge of this question is to determine whether the performance of text document classification can be improved by enriching the ontologies used as the background knowledge, and by enhancing the concept vectors through incorporating concepts importance into weighting scheme.

Research Question 2: *How can we use semantic and contextual information of terms within a discourse for enriching an existing ontology with new concepts? What influence do they have on the quality of ontology enrichment?*

The main challenge of this research question is exploring and developing an effective method for enriching ontologies, which, besides the contextual information, would also take into account the semantic information of terms appearing within a discourse. Moreover, in this research question an emphasis is being placed on investigating the contribution of semantic and contextual information on the quality, i.e. precision, of the ontology enrichment, and to what extent each of these components influences the quality.

Research Question 3: *What is the impact of statistical features on the performance of the enrichment of the ontology? How and to what extent the disambiguation techniques affect the quality of enriching an ontology*?

This research question explores the contribution of new statistical features used for deriving the context and evaluates to what extent such a contribution affects the performance of the enrichment of the ontology. It has also placed an emphasis on exploring the techniques for disambiguating the meaning of terms and on the evaluation of the ontology enrichment quality and classification performance using these disambiguation techniques.

Research Question 4: *Can we consider and assess automatically the importance of the concepts of an ontology in order to improve the concept vectors with new weighting scheme*?

The study related to this question is focused on exploring and developing a model which considers and is capable to automatically estimate the importance of concepts of an ontology. This research question also addresses the issue of improving concept vectors through an enhanced concept weighting scheme that aggregates the concept relevance and the concept importance.

1.3 List of Publications

This section provides a list of research articles which are produced as part of this research work. These articles have been published in peer-reviewed international conferences and journals.

A1: [69] Kastrati, Z., Imran, A., and Yayilgan, S., "*SEMCON: Semantic and Contextual Objective Metric*", in the 9^{th} IEEE International Conference on Semantic Computing (ICSC'15) (2015), IEEE, pp. 65-68.

A2: [70] Kastrati, Z., Imran, A., and Yayilgan, S., "*SEMCON - A Semantic and Contextual Objective Metric for Enriching Domain Ontology Concepts*", International Journal on Semantic Web and Information Systems (IJSWIS) (2016), vol. 12, issue 2, pp. 1-24.

A3: [72] Kastrati, Z., Imran, A., and Yayilgan, S., and Dalipi, F., "*Analysis of Online Social Networks Posts to Investigate Suspects Using SEMCON*", in the 17^{th} International Conference on Human-Computer Interaction (HCI'15) (2015), Springer, pp. 148-157.

A4: [67] Kastrati, Z., and Imran, A., "*Adaptive Concept Vector Space Representation using Markov Chain Model*", in the 19^{th} International Conference on Knowledge Engineering and Knowledge Management (EKAW'14) (2014), Springer, pp. 203-208.

A5: [68] Kastrati, Z., Imran, A., and Yayilgan, S., "*An Improved Concept Vector Space Model for Ontology Based Classification*", in the 11^{th} International Conference on Signal Image Technology & Internet Systems (SITIS'15) (2015), IEEE, pp. 240-245.

A6: [74] Kastrati, Z., Yayilgan, S., and Hjelsvold, R., "*Automatically Enriching Domain Ontologies for Document Classification*", in the 6^{th} International Conference on Web Intelligence, Mining and Semantics (WIMS'16) (2016), ACM, pp. 1-4.

A7: [73] Kastrati, Z., and Yayilgan, S., "*Supervised Ontology-Based Document Classification Model*", in the International Conference on Compute and Data Analysis (ICCDA'17) (2017), ACM, pp. 245-251.

A8: [71] Kastrati, Z., Imran, A., and Yayilgan, S., "*A Hybrid Concept Learning Approach to Ontology Enrichment*", IGI Global 2017, ch. Innovations, Developments, and Applications of Semantic Web and Information Systems.

An overview of the published articles and their relationships to the research questions is shown in Figure 1.1.

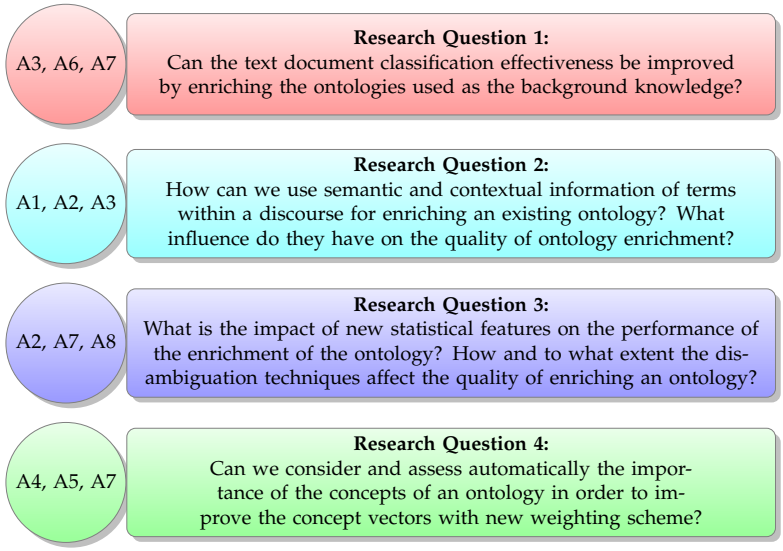


Figure 1.1: An overview of articles and their relationships to the research questions

1.4 Structure of the Thesis

This dissertation consists of the three major parts:

- **Introduction**: The remainder of this first part includes two more chapters: Chapter 2 that contains a summary of state of the art, and Chapter 3 that describes the contributions of this thesis.
- **Main Part**: Contains the published articles A1-A7 listed above, and the conclusions. Articles are categorized into three main research aspects: 1) Ontology concept enrichment, 2) Concept weighting scheme, and 3) Improving document classification. This part is structured according to these research aspects.

- **Appendix**: Includes the Chapter 12, in which, the ontology concept enrichment research aspect is also addressed. The chapter is an enhanced paper of the research article A2, and this is the reason that we do not cover it in the Main Part of this dissertation. We modified the reference and the literature in order to include the most current research findings related to the subject of the article, and augmented some new opinions for the discussion and contribution of the issues in the original article. The title and the abstract is also modified reflecting the enhanced content.

In more detail, the main part consists of the following:

- **Part II**: This part tackles the ontology concept enrichment and it contains three chapters. Chapter 4 shows the SEMCON model developed for enriching ontologies while Chapter 5 gives an extension of the model with additional experiments and comparisons. An application of SEMCON model on analysing OSNs is presented in Chapter 6.

- **Part III**: The focus of this part is on the concept weighting scheme and it is composed of two chapters. Chapter 7 describes an automatic approach based on Markov model for computing concept importance while in Chapter 8 we show the implementation of this approach on concept vectors in order to improve them with new concept weighting scheme.

- **Part IV**: This part focuses on the approaches for improving document classification performance and it includes two chapters. In Chapter 9, we present a document classification utilising an enriched ontology while Chapter 10 contains a ontology-based classification model which employs an enriched ontology and a new concept weighting scheme.

- **Part V**: This final part includes the chapter about the conclusions and future work.

The tree diagram illustrated in Figure 1.2 shows the structure of main part of this thesis. Ellipses in the tree diagram represent the research aspects addressed by this thesis while rectangles show the main features which have been tackled for each research aspect. The figure would serve as a guide to aid the reader, and it will be shown in the beginning of each part along with the chapters that are included in that part.

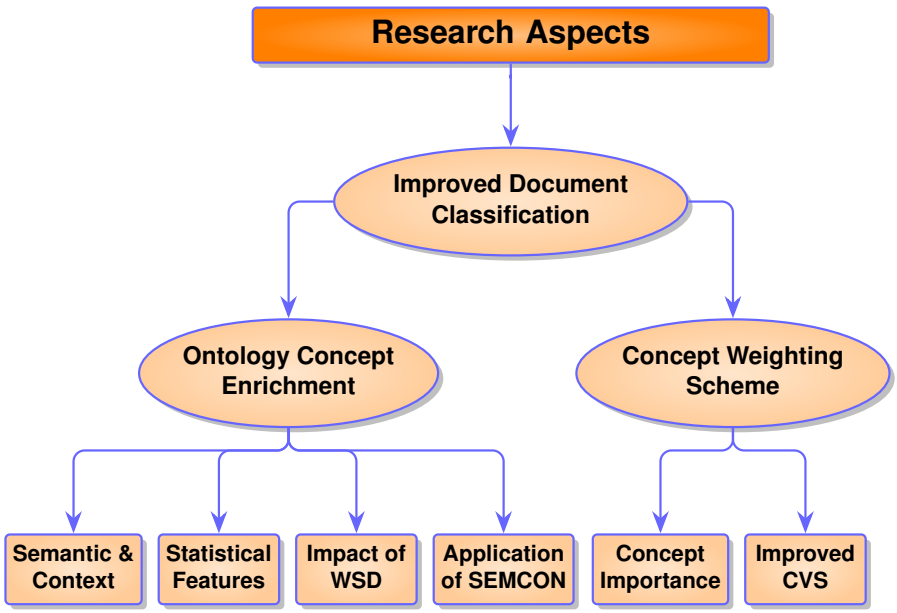


Figure 1.2: Tree diagram of the main research aspects addressed in this thesis

Chapter 2

State of the Art

In this chapter, we provide an overview of the state of the art research work done in the field of text document classification with a focus placed on the existing techniques of indexing that use ontologies as a background knowledge. This chapter provides a more comprehensive overview of the state of the art than that provided in the research articles presented in Part II. First, we provide an introduction of the ontology-based document classification approaches related to our research work. Then, we explore the techniques to identify and extract the lexical information attached to the ontology concepts; a process known as ontology concept enrichment. Finally, we provide a description of the concept weighting approaches employed for assessing the weight of concepts.

2.1 Introduction

The term ontology originates from the field of philosophy. Philosophers described the ontology as the study of existence. More than two thousand years ago, Aristotle was the first who developed an ontology for his classification system of categories, which is still relevant for defining nowadays ontological classification systems.

Ontology is defined as a fundamental form of representation of knowledge about the real world and it has been growing into popular research in computer science. In the context of computer science, an ontology is defined as a set of representational primitives which are used to model the knowledge of a particular domain or a discourse [53]. These representational primitives are typically composed of classes, attributes (properties) and relationships among these classes.

There is no single definition for ontology but the contribution provided by Gruber [52] is actually the first credible attempt at defining the notion of ontology. Gruber defines the ontology as "*a specification of conceptualization*". Gruber's definition is widely accepted among researchers; however, one objection about it is the general nature of the term '*specification*', which allows different interpretation starting from simple one (simple glossaries) to more advanced (logical theories of predicate calculus). Later, Borst [16] modified the Gruber's definition as "*ontology is a formal specification of a shared conceptualization*". Inclusion of the two words '*shared*' and '*formal*' in the modified definition makes it more explicit emphasizing that conceptualization should express a shared view between several parties and should be expressed in a (formal) machine readable format.

Ontology concepts and their relationships in a domain are commonly described using a 5-tuple based structure [91]. This 5-tuple ontology structure is formally defined as:

$$O = (C, R, H, rel, A) \tag{2.1}$$

where:

- C is a non-empty set of concepts
- R is a set of relation types
- H is a set of taxonomy relation of C
- rel is a set relationship of C with relation type R, where $rel \subseteq C \times C$

- A is a set of description of logic sentences.

Ontologies vary in their coverage and level of details and based on this variation they generally can be divided in 3 different types: upper (top-level) ontologies, lexical ontologies, and domain ontologies. These types of ontologies create conceptualization by defining vocabularies organized by formal relationships among concepts.

Upper ontologies also known as top-level ontologies are ontologies which consist of concepts that are universal, generic, and abstract, and are common across all domain areas. These ontologies provide a structure and a set of concepts which can be used as a starting point to develop ontologies for a particular domain, e.g. education, science, finance, etc. SUMO (Suggested Upper Merged Ontology) [114], Cyc [82] and its free version OpenCyc, and DOLCE (Descriptive Ontology for Linguistic and Cognitive Engineering) [50], are some prominent examples of the upper ontologies.

Lexical ontologies are the type of ontologies which describe the linguistic knowledge. The lexical senses (meanings) are defined by ontological structures (concepts). WordNet - an English based system [44], and its counterparts, HowNet - a Chinese-English bilingual system [33], and EuroWordNet - a multilingual lexical database for European languages based on WordNet [149], are some examples of lexical ontologies.

Domain ontologies are the third types of ontologies which are developed for a specific domain area. These ontologies consist of concepts that are less abstract but more specific comparing to concepts of upper ontologies that usually are more generic and common over several domains. Domain ontologies are one of the most important [27] and commonly used types of ontologies because nowadays most of the application ontologies are developed for particular target domain.

Ontologies are used in a wide range of applications and text document classification (also known as categorization), in particular, is one among these applications where ontologies play a vital role. Ontologies provide knowledge that is organized in a more structural and semantic way, therefore, ontology-based document classification systems utilize ontologies to derive and exploit the semantic aspect of documents. Consequently, ontologies enable to move from a document evaluation based on terms to an evaluation based on concepts, thus moving from lexical to semantic interpretation.

Ontology-based document classification has become increasingly important and attractive research topic in many areas, dealing with enrichment of document and category representation, and classification of documents in real time and without training corpus. Document representation is enriched by adding the semantics derived from ontologies into documents. An example of document classification which utilizes ontologies and relation between documents as a background knowledge in order to enrich the document representation is presented in [109]. A set of binary $T \times T$ matrices that contain all relations between terms, i.e. hyponyms, hypernyms, hyponyms of hyponyms, etc., is defined as a background knowledge. The hyponymy and associative relations are extracted from General Finish Ontology - YSO [65] and they are used to add ontology information to the terms appearing in the documents. This way, the classification model is capable to extend the traditional bag of words classifier with new relations utilizing the background knowledge exploited by the ontology.

Similarly, the approach presented by Camous et al. [21] exploits the ontology to enrich document representation but its focus is on enriching documents from the biomedical domain. More concretely, they presented a domain-independent classification approach which uses the relations between concepts from Medical Subject Headings - MeSH ontology [159] for enriching the existing MeSH based representation of documents. To acquire new concepts that are semantically close to the initial ontology concepts, a semantic similarity measure is used. The semantic similarity is calculated by simply counting edges (relations) between concepts in the MeSH hierarchy. The authors assume that all edges between concepts in the hierarchy correspond to the same semantic distance. The edges between concepts (descriptor or qualifier) in MeSH hierarchy are two types: the broader-

than type which is close to hyponomy and holonomy relationships used in WordNet, and narrower-than type which is close to WordNet's hypernymy and meronymy relationships.

There are some ontology-based classification approaches that use background knowledge derived from lexical ontologies to enhance document representation. These approaches are focused primarily on exploiting background knowledge from WordNet to enrich semantically the document representation. For instance, Nasir et al. [103] presented a semantically enriched document representation for text classification which is based on a semantic relatedness measure called Omiotis [145]. Omiotis is constructed from the word thesaurus and lexical ontology WordNet and it takes into account all of the available semantic relations in WordNet. Authors report that their approach provides significant improvements across different text classification methods and different data sets. Later, Nasir et al. [104] presented a similar approach of enriching document representation but with some differences. In addition to the background knowledge gathered by WordNet, they used the background knowledge derived by Wikipedia, and by large text corpora. Additionally, they considered, besides Omiotis that uses WordNet, two other semantic relatedness measures: one knowledge-based called Wikipedia Linked-based Measure - WLM [100] that uses Wikipedia, and one corpus-based called PoitWise Mutual Information - PMI [146] that uses word co-occurrence trained on SemCor corpus.

Other ontology-based approaches [131, 139, 140, 141] concern with real time document classification. These approaches do not require a training set or a learning process to train the classifier but they generally rely on the computation of similarity between the terminology information extracted from text documents and the ontology categories. For example, the ontology-based document classification approach presented in [139, 140, 141] basically involves two phases: 1) finding relevant terminology (key vocabulary) in the documents, and 2) mapping the vocabulary into a node (concept) in the concept hierarchy. The flow diagram of this classification approach is illustrated in Figure 2.1.

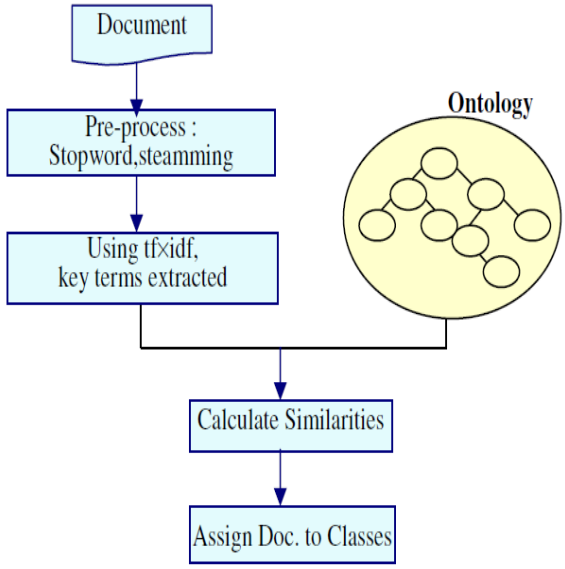


Figure 2.1: Document classification using Ontology [139]

The first phase of the approach shown in Figure 2.1 is concerned with document preprocessing where removing of stop words and stemming of words are performed and the

relevance of terms is computed using *tf*idf*. In the second phase, the document is classified into the appropriate category by mapping the given document into a concept with the highest similarity value which is computed using Equation 2.2

$$Sim(Node, d) = \frac{\sum_{i=0}^{N} \frac{freq_{i,d}}{max_{i,d}}}{N} \times \frac{V_d}{V} \tag{2.2}$$

where, N denotes the frequency of a concept, $freq_{i,d}$ is the frequency of property (feature and attribute) i that is matched in document d, and $max_{i,d}$ is the frequency of the property that is matched the most in document d. V indicates the number of constraints, that is, type of associations allowed between concepts (i.e. *has-a*, *part-of*), and V_d the number of constraints satisfied by document d.

Another ontology-based approach which also relies on real time classification of documents is presented in [131]. This approach exploits the background knowledge represented in a domain ontology and it uses ontology concepts, relationships between these concepts, and the taxonomy of categories represented in the ontology. The classification of a given document into the appropriate category is achieved by transforming the text document into a graph structure and employing then entity matching and relationship identification. This approach consists of 4 modules: 1) a preprocessing step which involves lemmatization, stemming, and removing of stop-words, 2) a thesaurus which indicates when a word occurring in the text is present in the ontology, 3) a set of ontology terms tagged with its corresponding classification label, and 4) a thesaurus crawling algorithm which evaluates the matching degree of text words with a corresponding ontology term. The process flow diagram of this ontology-based classification approach is illustrated in Figure 2.2.

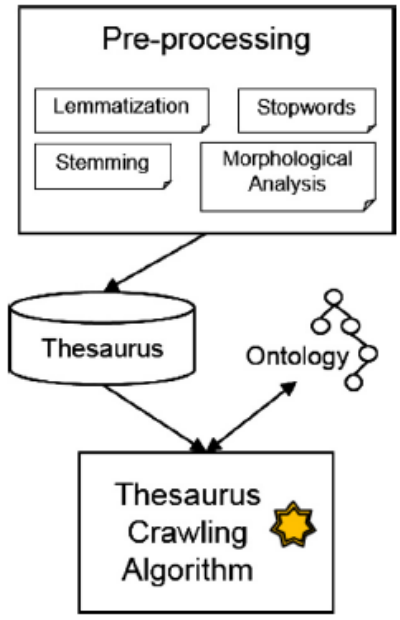


Figure 2.2: Ontology-based classification process flow diagram [131]

2.2 Ontology enrichment

Ontology enrichment is a process which aims to improve an existing ontology by updating it with new concepts. Ontology enrichment using textual data has been an attractive research field recently. This attraction produced a wide variety of approaches which based on relevance to the concept enrichment task can be grouped into 2 major categories: 1) Statistical-based approaches, and 2) Linguistic-based approaches.

Statistical-based ontology enrichment approaches [8, 26, 28, 37, 89, 122, 158] rely on the statistical features of terms such as term frequency (*tf*) or term frequency inverse document frequency (*tf*idf*), and term collocations, to identify and extract concepts within the textual data. The horizontal (non-taxonomic) relationships represented by dependencies between concepts such as synonymy, meronymy, antonomy, etc., are computed using lexical co-occurrence statistics [28, 158]. For example, DOODLE [134], and its extended version, DOODLE II [158] is a statistical-based learning approach that relies on lexical co-occurrence statistics. A machine-readable dictionary and domain-specific texts are used as input to the system to build concepts along with taxonomic and non-taxonomic relationships of domain ontologies. DOODLE II deals with the non-taxonomic relationships represented by dependencies between concepts such as synonymy, meronymy, antonymy, attribute-of, and possession. These non-taxonomic relationships are extracted from domain specific texts using lexical co-occurrence statistics based on WordSpace [61]. WordSpace is a multi-dimension vector space which is composed of a set of word vectors. The inner product between two word vectors represents the semantic relatedness between those two words. Consequently, words that occur together frequently (their inner product is above some upper bound) can have non-taxonomic conceptual relationships.

Other statistical approaches are concerned with batches of terms where meaning of terms is represented by term co-occurrences and the frequency of the co-occurrences [8, 26, 37, 122]. The occurrence of two or more terms within a discourse is known as collocation [63]. Learning utilizing term collocations and statistical features (frequency) of collocations is the most addressed technique in statistical concept learning approaches. Location and extraction of lexical information of a concept is achieved using correlation between terms and a given concept within a window size. SYNOPSIS [37] is an example which follows this approach to acquire the lexical information for a given concept. It is an automatic system that builds a lexicon for each specific term known as criterion. To identify and locate the lexicon terms within a document, it (document) is initially split into several passages. The similarity between terms and the user criterion is computed using the relative position in a window size, that is, the frequency of occurring of grammatical terms, i.e. common nouns, between a given term and the user criterion. Based on the assumption that the most frequent term is more likely to characterize the criterion, a grammatical term with the highest similarity is used then to built its lexicon. Later, Ranwez et al. presented an adaptation of SYNOPSIS, called CoLexIR [122]. CoLexIR, which stands for Conceptual and Lexical Information Retrieval, employs the same approach as SYNOPSIS but rather than building lexicon of terms, it builds automatically the lexicon of concepts of an ontology.

Arabshian et al. [8] presented a semi-automatic ontology learning system called LexOnt, which uses the Programmable Web directory as the corpus, external domain knowledge such as Wikipedia and Wordnet, and the current state of the ontology, to suggest relevant terms that may be incorporated within the ontology. To accomplish this, LexOnt generates initially a list of terms and phrases obtained by *tf*idf* algorithm, and significant phrase generation which is a two-phase process: the first phase which determines a list of collocations, and the second phase which filters out only unique collocations from the list obtained in the first step. Next, the generated terms are compared to external domain knowledge such as Wikipedia, Wordnet, and the current state of the ontology in order to obtain only the significant terms. More concretely, significant terms are considered only those terms that are matched to the Wikipedia page description of the category. For each of these significant terms, the synonymous terms are found using WordNet. Finally, the system searches if

any of the generated terms lexically match terms that have been assigned manually in the ontology and labels these terms to indicate that they have already been created.

Brunzel [20] presented a system called EXTREEM (Xhtml TREE Mining) that also relies on statistical approach but rather than taking unstructured data as input, the system exploits the structure of Web documents for acquisition of the relevant terminology to enriching an ontology. Web content is constituted by markups that represent textual data marked-up by tags. The system exploits these markups and generates a collection of text spans in which a frequency statistic (frequency of occurrences) is employed to select the candidate terms for enriching the ontology

In contrast to statistical approaches which depend on statistical features of terms and their co-occurrences, linguistic approaches rely on linguistic components, i.e. noun phrases, and involve natural language processing (NLP) techniques such as syntactic [55, 105, 106], morpho-syntactic [135, 161], and lexico-syntactic patterns analysis [1, 18, 38, 60, 101], to acquire concepts from textual data. The NLP based technique of lexico-syntactic pattern analysis is one of the most commonly used linguistic approach and the first lexico-syntactic patterns are introduced and explored by Hearst in [60]. These patterns represented in form of regular expressions are used for acquisition of ontological knowledge from English textual data. Table 2.1 shows the list of lexico-syntactic patterns proposed by Hearst.

Table 2.1: List of lexico-syntactic patterns introduced by Hearst

No	Lexico-syntactic patterns
1	NP_0 such as {NP, }* {(and|or)} NP_n
2	such NP as {NP, }* {(or|and)} NP
3	NP {, NP}* {,} or other NP
4	NP {, NP}* {,} and other NP
5	NP {,} including {NP,}* {(or|and)} NP
6	NP {,} especially {NP,}* {(or|and)} NP

The patterns proposed by Hearst demonstrated to be successful on identification and acquisition of ontological knowledge, that is, a set of relationships such as hypernym, but this technique was limited to a small number of patterns. A list with a larger number of patterns is used by Etzioni et al. [38, 39] in the proposed lexico-syntactic based approach called KnowItAll. The list of lexico-syntactic patterns employed by KnowItAll is comprised of some patterns adapted from Hearst's patterns and by some other patterns which are developed independently by the authors. Table 2.2 shows the list with some of the lexico-syntactic patterns used by KnowItAll.

Table 2.2: A sample of lexico-syntactic patterns used by KnowItAll

No	Lexico-syntactic patterns
1	NP1 {","} "such as" NPList2
2	NP1 {","} "and other" NP2
3	NP1 {","} "including" NPList2
4	NP1 "is a" NP2
5	NP1 "is the" NP2 "of" NP3
6	"the" NP1 "of" NP2 "is" NP3

KnowItAll is a domain-independent system which selects the relevant concepts for enriching an ontology using the developed patterns and by evaluating concept plausibility using Turney's PMI-IR algorithm [146] - a version of the pointwise mutual information statistical measure.

ABRAXAS [18, 66] is another approach which depends on lexico-syntactic pattern analysis to perform concept and relation extraction for enriching ontologies. It uses three re-

sources, namely, a corpus of texts, a set of lexico-syntactic learning patterns, and an input ontology. The input ontology is needed for deriving a set of lexico-syntatic patterns from which all co-occurrences of the subject-object pairs found in this ontology are spotted. These derived patterns are then applied to the corpus in order to acquire new ontology concepts. In addition to this, syntactic and semantic similarity measures are employed to cluster the subject/objects of the concepts occurring in the corpus. This step produces a ranked set of candidate concepts from which only the top most concept is selected to be added into the ontology.

There is a strand of approaches that exploit the documents from the medical domain through the lexico-syntactic pattern analysis to acquire the relevant terminology. Such an example is presented in [144], in which, a simple lexico-syntatic pattern of the form $Noun\_\{and, or, but\}\_Noun$, and the Resnik similarity algorithm [123], are used to locate and extract pairs of noun terms from the Oshumed corpus [62]. The lowest common ancestor of each pair of noun terms that reflects the correct medical meaning of these two nouns is used to enriching WordNet concepts. Ben Abacha and Zweigenbaum [1] describe a similar lexico-syntactic pattern-based approach for recognition of medical concepts and relations linking those concepts. They present a platform called MeTAE (Medical Texts Annotation and Exploration) which is composed of two main parts. The first part deals with the identification and extraction of medical entities using an enhanced MetaMap [9]. MetaMap is a tool which allows effective mapping of biomedical text to the UMLS (Unified Medical Language System) concepts. The second part deals with the extraction of semantic relations that exist between concepts identified in the first part.

In addition to the above mentioned approaches, there are some other linguistic approaches that rely on natural language processing techniques concerning morpho-syntactic analysis. An example of morpho-syntactic based approach employed as learning technique is HASTI [135]. HASTI is an automatic system for building ontologies utilizing a combination of morpho-syntactic and semantic analysis approach. The input to the system is unstructured data kept in the form of natural language texts in Persian. The initial ontology in HASTI is a small kernel (small ontology) with a very small lexicon at the beginning but it grows gradually by extracting new terms. These extracted terms along with their conceptual relationships, taxonomic and non-taxonomic, are used on top of the existing kernel to build the ontology.

OntoCmaps [161] is another approach which depends on natural language processing technique to extract information. It is a domain independent ontology learning system that is based on two main steps: a knowledge extraction step, and a knowledge filtering step. The knowledge extraction step relies on patterns to extract candidate concepts from texts. These patterns are mainly syntactic patterns which use a dependency grammar formalism and part-of-speech tagging. Stanford Parser [76] is used to perform the dependency analysis where a set of grammatical relations that link each pair of related words in a sentence is obtained. 31 different syntactic patterns are used by system to identify and extract ontological knowledge. The knowledge filtering step is used to filter relevant concepts among the candidate ones obtained by the first step.

A system for enriching an ontology by populating it with new concepts is also described in [127, 126]. The system employs a linguistic learning approach and is composed of 4 sequential phases. The first phase concerns with a morphologic and syntactic analysis, in which, tokenizing, tagging, lemmatizing, and parsing, are performed using the GATE framework [30]. In the second phase are identified and acquired name entities while in the third phase the system classifies name entities as instances, attributes, and relationships of the ontology. The last phase checks the consistency of the enriched ontology using OWL-DL reasoners such as Hermit or Pellet [138]. An extended OnTour ontology, and two datasets from hotels and restaurants domain, have been used for experimenting.

Ontology learning system based on syntactic analysis described by Hahn and Romacker in [55] utilizes technical documents in German language taken from test reports from the

information technology domain and medical finding reports. They developed the system named SynDiKATe whose approach to learning new concepts is based on syntactic analysis in performed in two different levels: sentence and text level. The syntactic analysis results are captured in a dependency graph constituted of vertices that represent terms, and edges that represent relations between these terms. Later, Hahn et al. brought an extension of SynDiKATe system called medSynDiKATe [56] that is designed to automatically identify and acquire medical knowledge from medical finding reports. The system takes the advantages of using various textual resources required for text understanding with a focus being placed on grammar and domain knowledge. Additionally, the system puts an emphasis on finding alternative ways to support knowledge acquisition to foster the scalability of the system. An automatic and semi-automatic concept learning approach are employed and fully embedded in the text understanding process of the system.

Navigli and Velardi [105, 106] have also acknowledged the problem of ontology enrichment based on natural language processing utilizing syntactic analysis. They proposed a new automatic approach for enriching a core ontology with the concepts and properties of a domain glossary. This approach is applied in the cultural heritage domain using a core ontology called CIDOC-CRM [29]. Resources such as the AAT art and architecture glossary, WordNet, and the Dmoz taxonomy for identification named entities, are used to enrich the CIDOC. To accomplish enriching task, this method involves several steps. The first step is part of speech analysis where a given document is processed with the TreeTagger capable to capture named entities of locations, organizations, persons, numbers, and time expressions. The next step is annotation of documents using regular expressions enriched with syntactic and semantic constraints. Syntactic constraints are defined by matching the lemma of the word with a regular expression such as Verb-Preposition-Noun, e.g. *Composed-Of-Stone*, while semantic constraints represent matching of words with concepts of formal core ontology CIDOC-CRM. The final step is formalisation of vocabularies to enrich CIDOC-CRM ontology.

Acquiring terminology for ontology population and enrichment utilizing natural language processing techniques is examined by Valarakos et al. in [148]. The researchers placed the focus on the maintenance of the ontological concepts and their lexicographic variants. To achieve this, new lexical variants are initially identified and attached to concepts of a domain ontology and then non-taxonomic relationships between concepts are acquired. This process referred as ontology population and enrichment is shown in Figure 2.3. It is an incremental process which consists of the following modules:

- Ontology-based semantic annotation module which uses concepts of the domain ontology to automatically annotate a domain specific corpus
- Knowledge discovery module which aims to locate new ontological concepts
- Knowledge refinement module whose concerning is the identification of lexicographic variants of each concept using a partition-based clustering algorithm called COCLU [147]. For example, 'Pentium 2' can have different variants, such as 'Pentium II', 'P II' or 'Intel Pentium 2'
- Validation and insertion module where a domain expert validates the candidate concepts that have been attached in the ontology.

While the above research all rely on either statistical or linguistic learning approach, other research take the advantages of both approaches for ontology enrichment. TEXT-TO-ONTO [26, 89] is an example which relies on learning by term collocations and co-occurrences technique with a basic linguistic processing technique. Textual data kept in either structured, semi-structured, or unstructured format can be exploited through the frequency of term co-occurrences to locate and acquire horizontal (non-taxonomic) relations using background knowledge gathered by a lexicon and a taxonomy.

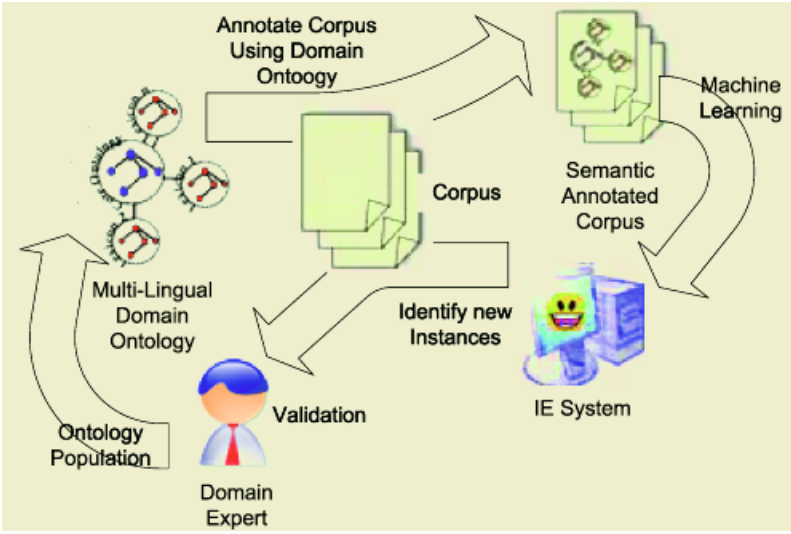


Figure 2.3: Ontology population and enrichment flow diagram [148]

WEB→KB [28] is another ontology learning system that relies on statistical and linguistic approach to identify and acquire concepts from the world wide web documents. To achieve this, the system is primarily trained using two training sets: 1) a set of concepts and relations that are interesting for creating the knowledge base, and 2) a set of hypertexts with labelled regions that are instances of these concepts and relations.

The work by Mima et al. [101] relies on lexico-syntactic pattern analysis and statistical information to identify and acquire the terminology for enriching ontologies. To achieve this, they developed the ATRACT system. ATRACT, which stands for Automatic Term Recognition And Clustering of Terms, is an approach used for terminology recognition and clustering from the domain of molecular biology. Terms included in documents represented in HTML/XML format are identified and extracted using the C/NC-value method [47]. C/NC-value is a method for the automatic extraction of multi-word terms, which combines linguistic (lexico-syntatic patterns) and statistical information (frequency of occurrences of terms). These terms are then clustered based on the context in order to form the concepts.

In contrast to the above mentioned approaches which utilize textual data, Castano et al. [22, 23] present an ontology enrichment system, namely BOEMIE, that is able to identify and extract concepts from a variety of modalities, including texts, images, and videos. The acronym BOEMIE stands for Bootstrapping Ontology Evolution with MultImEdia Information. The system requires an initial ontology to be enriched and a collection of documents from which new concepts are identified and extracted. Due to the multi-modal nature of BOEMIE, it separates the concepts into mid-levels concepts and high-levels. Mid-level concepts represent primitive concepts that can be mapped directly to the objects (documents) while high-level concepts refer to composite concepts which can not be mapped directly to objects and they usually are build on top of the primitive ones. The ontology enrichment process of BOEMIE, illustrated in Figure 2.4, consists of the following tasks:

- Concept learning which aims to propose new concepts and relationships by exploiting similarities found through clustering.
- Concept enhancement which is responsible for improving a concept identified in pre-

vious task, through knowledge gathered from external sources such as external domain ontologies or taxonomies.

- Concept definition where an ontology expert must approve whether a new concept (relation) is ready for adding into the ontology or its definition must be revised.
- Concept validation whose role is consistency checking. It tries to detect possible inconsistencies which may occur due to the addition of the new concept/relation to the ontology.
- Concept assimilation is the last task of ontology enrichment. It takes care for the changes in the ontology structure that are required to add the newly formed concept/relation into the ontology.

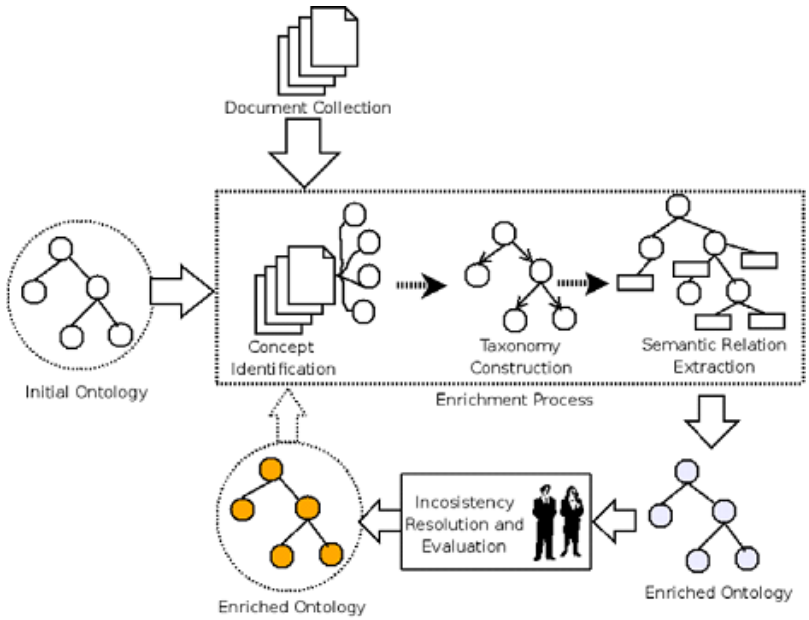


Figure 2.4: The ontology enrichment process [23]

A summary and comparison of the research described above in this section is presented in Table 2.3, which constitutes of six dimensions that represent the major distinguishing features among the ontology enrichment research. The very first column of the table contains the reviewed approaches while the following columns indicate the evaluated features of these approaches. The Learning Target column denotes the elements learned from the enrichment process and it can take the values of *Concepts* and *Relations*. The Learning Approach column describes the techniques used by the approach to identify and extract concepts/relations and its values can be either *Statistical*, *Linguistic*, or both *Statistical* and *Linguistic*. The Data Source column indicates which type of input data are supported by the approach and the values of *Struct* for structured, *Semi* for semi-structured, or *Unstru* for unstructured, can be taken of this column. The next column, WSD, describes whether the approach addresses the disambiguation issue and it can take the values of *Yes* or *No*. The following column namely Auto shows if the approach is automatic or manual, and its

values can be either *Yes* for automatic, *Semi* for semi-automatic. The final column, Domain Specific, describes if the approach is domain specific or independent and it can take the values of *Yes* for domain specific and *No* for domain independent.

Table 2.3: Summary of the related ontology enrichment researches

Approach	Learning Target	Learning Approach	Data Source	WSD	Auto	Domain Specific
DOODLE II	Concepts Relations	Statistical	Semi Unstru	No	Yes	Yes
WEB→KB	Concepts Relations	Statistical Linguistic	Semi Unstru	No	Yes	Yes
SYNOPSIS	Concepts	Statistical	Unstru	No	Yes	Yes
CoLexIR	Concepts	Statistical	Semi Unstru	No	Yes	Yes
TEXT-TO-ONTO	Concepts Relations	Statistical Linguistic	Struct Semi Unstru	No	Yes	No
LexOnt	Concepts Relations	Statistical	Unstru	No	Semi	Yes
XTREEM	Concepts Relations	Statistical	Semi Unstru	No	Yes	No
ABRAXAS	Concepts Relations	Linguistic	Semi Unstru	No	Yes	Yes
KnowItAll	Concepts	Linguistic	Semi Unstru	No	Yes	No
ATRACT	Concepts	Linguistic Statistical	Semi Unstru	Yes	Yes	Yes
[144]	Concepts	Linguistic	Unstru	Yes	Semi	Yes
MeTAE	Concepts Relations	Linguistic	Unstru	No	Yes	Yes
HASTI	Concepts Relations	Statistical	Semi Unstru	Yes	Yes	Yes
OntoCmaps	Concepts Relations	Linguistic	Unstru	Yes	Semi	No
[126, 127]	Concepts Relations	Linguistic	Unstru	Yes	Yes	Yes
SynDiKATe	Concepts	Linguistic	Unstru	No	Yes	Yes
medSynDiKATe	Concepts	Linguistic	Unstru	No	Semi	Yes
[105, 106]	Concepts Relations	Linguistic	Unstru	Yes	Yes	Yes
[148]	Concepts	Linguistic	Unstru	Yes	No	Yes
BOEMIE	Concepts Relations	Statistical Linguistic	Struct Semi Unstru	Yes	Yes	No

As can be seen from Table 2.3, most of the approaches (12) put the focus on learning both concepts and relations through the ontology enrichment process while the some of the other approaches (8) focus on learning only concepts.

From the concept learning approach perspective, the linguistic-based technique is employed in the most of the approaches (10), followed by the statistical-based which is used by six approaches. Four systems employ both the linguistic and statistical learning technique.

The unstructured textual data is used as input in all the approaches. Ten of the ap-

proaches support also the semi-structured data, and just two, TEXT-TO-ONTO and BOEMIE, support the structured one.

The disambiguation issue is ignored by twelve of the approaches and it is addressed only in nine of them. However, from these nine approaches, it is only [105, 106] that deal explicitly with the disambiguation issue where a word sense disambiguation algorithm called Structural Semantic Interconnections - SSI, is used to find the correct meaning of concepts.

Most of the approaches are automatic (four are semi-automatic) and only few of them (4) are domain independent.

2.3 Weighting Scheme - Concept Importance

Similar to the classical vector space model which assigns weights to keywords appearing in a document, the concept vector space model assigns weights to concepts. These weights reflect the relevance of concepts on representing the document meaning and are usually computed using a modified *tf*idf* algorithm [128]. The modified version of this algorithm relies on the frequency of occurrences of concepts which is primarily defined as the number of times the label of the given concept occurs in that particular document. The main drawback of this algorithm is that it does not consider the importance of concepts reflected by the number of ties (relations) a concept has to other concepts in an ontology.

Recently, there have been some efforts in the research domain to take into account the discriminating feature of ontology concepts indicated by their position depicted in the ontology hierarchy. The common aspect of these efforts is computing of concepts weight empirically through trial and error by conducting experiments thus keeping these weights fixed. For instance, the research shown in [120] follows the idea that the higher the concept in the ontology tree, it is less abstract and gives less contribution. Following this idea, they used different weights for concepts depending on the position where they occur in the ontology hierarchy. The first weight was assigned to concepts which are occurring as classes, second weight for concepts occurring as subclasses and the third weight for concepts occurring as instances. Finally, after an empirical analysis through trial and error by conducting experiments, the value of 0.2 is set for the concepts which occur as classes, 0.5 for concepts occurring as subclasses and 0.8 when concepts occur as instances.

A similar approach to [120] is proposed later by the researchers in [87], in which, the weight of concepts are computed based on the weights set to taxonomic and non-taxonomic relationships acquired by the MeSH ontology. Four types of relationships, namely, identity, synonymy, hypernymy, and meronymy, were taken in consideration and different weights are set to each of them. More precisely, for ontology relationships such as identity and synonymy, the weight is set to 1.0. The weight of 0.7 is set for hypernymy ontology relationships with a decrease step of 0.1 in backward direction at each level of the taxonomy and the weight of 0.8 is set for meronymy ontology relationship with a decreased step of 0.01 at each upper level. Based on these weights of relationships and the number of occurrences of concepts, the weight of the concept is defined formally in Equation 2.3

$$W(c_i) = \frac{\sum_{j \in R} Freq_j \times weight_j}{N} \tag{2.3}$$

where, $W(c_i)$ denotes the weight of the particular concept, $Freq_j$ denotes the frequency of the particular relationship, i.e., identity, synonymy, hypernymy and meronymy. $weight_j$ represents the weight set for the particular relationship, and N represents the frequency of occurrences of all concepts.

A simple example which illustrates the way of assigning weights to different relations for the MeSh ontology concept *Hepatitis* shown in Figure 2.5, is given in Table 2.4.

In addition to the approaches outlined above, there are also some other approaches [35, 42, 54, 116, 117] that focus on layers of ontology tree to assess the weights of concepts.

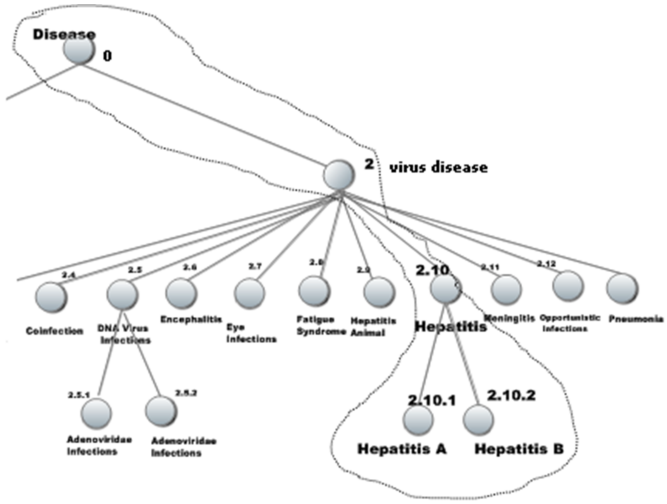


Figure 2.5: Selection of path for the concept *hepatitis* [87]

Table 2.4: Relationships of concept *Hepatitis* and their weights

Relationship	Concept	Weight
Identity	Hepatitis	1
Synonymy	Hepatitis Animal, Hepatitis Human	1
Hypernymy	Virus Disease	0.7
	Diseases	0.6
Meronymy	Hepatitis A, Hepatitis B	0.8

Consequently, the weight of each concept is computed by counting the length of path from the root concept (node) to the given concept. Fang et al. [42] compute weight of concepts using layers of ontology graph defined by counting the length of path which have *is-a* relations starting from the very top node. More formally, weight of a concept is computed using the Equation 2.4.

$$Weight(c) = \frac{1}{(layer)^{\frac{1}{4}}} \tag{2.4}$$

Similarly, researchers in [54] used path of ontology tree to define weights of concepts. They defined a concept's weight (Equation 2.5) as the fraction of path's length of current concept *h* and path's length of the branch including current concept *H*.

$$W(c) = \frac{h}{H} \tag{2.5}$$

Weight of non-leaf concepts which are located at the same layer in the ontology tree is computed using Equation 2.6.

$$W^{'}(c) = \frac{W(c)}{K^{n}} \tag{2.6}$$

where, *n* denotes the distance between the current concept and the concept with the longest path in the branch, and *K* is a constant with value set to 2.

A concept weighting scheme relying on concept's path length to compute weight of concepts is proposed in [5]. The weighting scheme is defined by using *tf*idf* along with

some information (path length) from the domain ontology. The idea of using ontology is to consider the relationships between concepts and semantics which can be ignored using only *tf*idf*. This concept weighting scheme is mathematically defined in Equation 2.7.

$$W_i' = W_i + \sum_j [-Log_{10}(E_{ij}) * W_j] \tag{2.7}$$

where,
W_i represents the value of *tf*idf* calculated based on the Equation 2.8.

$$W_i = \frac{f_i}{N_d} * log_{10}(\sum_{d,i} \frac{N_d}{df_i}) \tag{2.8}$$

E_{ij} denotes the information from the domain ontology reflected by the weight of the path from concept *i* to concept *j* in the ontology. If there is no path between these two concepts in the ontology, then the E_{ij} would be zero and the weighting scheme would be consisted of only *tf*idf*. The logarithm is used to increase the effect of weights of the ontology concepts on the final weights.

The work by Pereira and Tettamanzi [116, 117] uses also path length to compute the weight of concepts but considering only leaf concepts of the ontology. They assume that more general concepts, such as super-classes, are implicitly taken into account through the use of leaf concepts by distributing their weights to all of their subclasses down to the leaf concepts in equal proportion. Mathematically, the weights of concepts are computed using Equation 2.9.

$$N(c) = occ(c) + \sum_{c \in Path(c,\ldots,T)} \sum_{i=2}^{length(c)} \left(\frac{occ(c_i)}{\prod_{j=2}^{i} |children(c_j)|}\right) \tag{2.9}$$

where, *N(c)* is the weight of concept *c* computed using frequency of implicit (when a concept e.g. *dog*, is referred in a document by its super-class, e.g. *animal*) and explicit (when a concept is referred (mentioned) directly in the document) occurrences of concept *c*, and *occ(c)* shows the number of occurrences of lexicalizations of concept *c*.

The approach in [116, 117] computes the weights of concepts for domain specific ontologies and do not consider all possible concepts of ontologies. They apply a cut at a given specificity level considering only leaf concepts. Alternatively, Dragoni et al. [35, 34] presented an approach which is adapted for more general purpose ontologies and it considers all independent concepts contained in a given ontology. The authors report that by doing this, the weight associated to each concept is more precise and there is no need to apply the cut. This way, the final weight of a concept is defined by the depth of concept *c* in the ontology graph, frequency of occurrences in the document, and frequency of occurrences in the whole set of documents (corpus). Consequently, these two frequencies also rely on the number of ancestors (parent) and descendants (children) of concept *c*. The example given in Figure 2.6 illustrates computation of the importance of an ontology concept which is proportional to the number of children that all of its parents have.

In contrast to the approaches outlined above which assigns weights to concepts of an ontology, the approach presented by Ni et al. [107] concerns with assigning weights to concepts acquired from the knowledge base resource known as DBpedia [13]. The approach employs two assignment methods to compute and assign weights to a concept: a global method and a local method.

The global method which relies on graph-based weights, consists on the strength of semantic relationships among concepts measured through a modified closeness centrality property. The modified closeness defines the distance between two concepts as the inverse of the weight of the edge between them in the concept graph instead of using the shortest

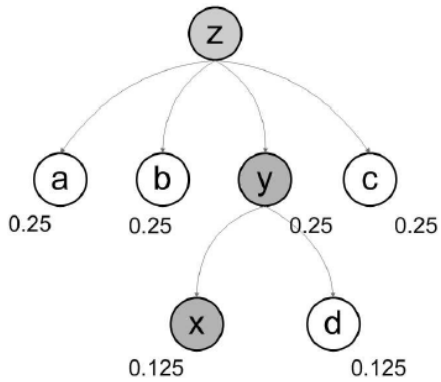


Figure 2.6: Ontology representation and importance for a concept

path between these two concepts. The formal definition of distance between two concepts, c_1 and c_2 is given in Equation 2.10.

$$dis(c_1, c_2) = \frac{1}{\lambda_1 \cdot ctxt(c_1, c_2) + \lambda_2 \cdot cat(c_1, c_2) + \lambda_3 \cdot struct(c_1, c_2)} \tag{2.10}$$

where, *ctxt*(c_1,c_2) represents context association, i.e. how often two concepts share context, *cat*(c_1,c_2) denotes category association, i.e. grouping similar concepts, *struct*(c_1,c_2) denotes structure association, i.e. Wikipedia infoboxes, and λ_i denotes weight parameters for the three types of associations.

Subsequently, the closeness centrality of a concept *c* is defined in Equation 2.11

$$centrality(c) = \frac{1}{|V|} \times \sum_{c_j \in V} \frac{1}{dist(c, c_j)} \tag{2.11}$$

where, *V* denotes a set of concepts in the concept graph.

The local method relies on content similarity between the Wikipedia page of the concept and the given document computed through Information retrieval techniques, specifically, using the simple *tf*idf* measure of document similarity.

Chapter 3

Contributions

This chapter presents a set of results, described as contributions, which are produced by this research work. It also depicts the relationships of these contributions in context of the published articles listed in Section 1.3.

3.1 Contributions of this Research

The main contributions of this research work are as follows:

Contribution 1: *An effective approach for enriching an ontology with new concepts utilising contextual and semantic information.*

The ontology enrichment problem statement is not new in the research community but according to our observation, most of the state of the art ontology enrichment research was performed using only contextual information derived from distributional property of terms such as term frequency (*tf* or term frequency inverse document frequency *tf*idf*), and terms co-occurrence analysis. In the research article A1, we have investigated the possibility to perform enriching of concepts of an ontology not only using contextual but also semantic information of terms occurring in a discourse. We have developed an ontology enrichment model called SEMCON which combines both contextual and semantic information of terms. Contextual information of a term is defined by its surroundings, that is, the part of a text or statement (passage) in which that particular term occurs and it is computed through the cosine distance between the feature vectors. The feature vectors are constituted of values derived by the frequency of occurrences of terms in corresponding passages, and the new introduced statistical features such as font types and font sizes. The semantic information on the other hand is defined through a semantic similarity measure based on the lexical database WordNet.

Subjective and objective experiments were conducted in the research article A1 and A2 to validate our proposed approach. The subjective survey was conducted by publishing online a questionnaire from which the subjects had to pick up 5 terms from a list of terms that were most semantically related to a particular concept. Using Borda Count method, a list consisted of the top 10 terms was obtained from the survey, which was then taken as the ground truth for evaluating the effectiveness of our SEMCON model, and the two other referent objective methods, namely *tf*idf*, and χ^2. The objective experiment was extended in the research article A2, in which, new objective method called Latent Semantic Analysis - LSA was used in addition to the two previous objective methods for comparing results obtained from SEMCON.

Contribution 2: *Introducing of new statistical features for deriving the context of a discourse.*

We have come up with a novel technique for deriving the context of a discourse (article A1 and A2). In addition to the term frequency, we have introduced for the first time in the research community two new statistical features namely term font sizes and term font types, for deriving the context. A linear increase model is adopted to

set different values for various font types and font sizes. The idea of using linear model is to keep the effect of each variable the same for all values of the other variables, e.g. the effect of bold font type terms is the same for every value of underline or italic font type terms.

An investigation on the impact of these new statistical features on the performance of the enrichment of the ontology in terms of precision is also performed in the research article A2. The findings showed that the proposed statistical features have a considerable impact on the performance of ontology enrichment.

Contribution 3: *The influence of each of contextual and semantic components on the performance of ontology enrichment.*

It is shown (article A1) that using both contextual and semantic information do contribute on improving the performance of ontology enrichment SEMCON model but there is a need to examine the effect that each of these components may have on constituting the performance improvement. To achieve this, in the research article A2, we conducted an empirical investigation of the impact of each of contextual and semantic components on the overall task of ontology enrichment process. Experiments were conducted using various settings of weight parameter w, and based on the empirical analysis of the dataset, we found that a balanced weight between these two components gives the best performance in terms of precision.

Contribution 4: *Application of the proposed SEMCON approach.*

We have applied the proposed SEMCON approach to analyse Online Social Networks (OSNs). More concretely, we proposed a model for automated social network analysis for identifying criminal activity and possible suspects, with a special focus on analysing Facebook posts (article A3). Users' data such as posts, feeds, and comments retrieved by the acquisition module through a dedicated Facebook crawler, have been exploited semantically and contextually using the ontology enrichment objective metric SEMCON. The final output of this automated network analysis model is a probability value of a user being a suspect computed through cosine similarity measure by comparing the terms obtained from the SEMCON and the concepts of criminal ontology. The model is evaluated empirically through an experiment conducted using the public information of 20 Facebook users.

Contribution 5: *Improving concept vectors with new concept weighting scheme.*

Concept importance shows how important a concept is in an ontology and this is reflected by the number of ties (relations) a concept has to other concepts. We explored the possibilities to automatically compute concept importance and a Markov-based approach has been introduced in the research article A4. Moreover, an improved concept vector space model (iCVS) which takes into account the importance of ontology concepts is presented in the research article A5. Concept importance computed using the approach presented in the research article A4 is aggregated with concept relevance computed using the frequency of concept occurrences in the dataset in order to enhance the concept weighting scheme. Experiments conducted on a real dataset showed that our iCVS proposed model yields higher classification accuracy comparing to the traditional concept vector space (CVS) model, ultimately giving better document classification effectiveness.

Contribution 6: *The impact of disambiguation on the quality of ontology enrichment.*

The impact of Word Sense Disambiguation on enriching concepts of an ontology by employing two techniques, namely, First sense heuristic and Maximizing semantic similarity, is investigated in the research article A7. Experiments are conducted and the observation showed that different terms are retrieved as the relevant terms for

enriching a particular concept when these two disambiguation techniques are employed by our proposed ontology enrichment model. Hence, using disambiguation techniques yields different classification performances and the accuracy of some classifiers is more affected (Naive Bayes) than the accuracy of the others (SVM or Decision Tree) when the model applies these disambiguation techniques.

Contribution 7: *Improving document classification effectiveness*. In the research article A6, we applied the proposed SEMCON approach to enriching a baseline ontology with new concepts and investigated how and to what extent the ontology enrichment impacts the classification performance. To achieve this task, we used the Top-N terms obtained from SEMCON as the most relevant terms for enriching each concept of the baseline ontology. The results obtained by experiments conducted on the documents from the funding domain using Decision Tree classifier showed that the average F1 measure is improved from 65.5% to 77.1% when concepts of the baseline ontology are enriched with 4 new terms. Furthermore, the research article A7 extends the document classification approach presented in the article A6 in two aspects: 1) in addition to ontology enrichment, the model employs a new concept weighting scheme that aggregates concept importance and concept relevance, and 2) it investigates the impact of disambiguation techniques on the quality of ontology enrichment and classification effectiveness. The experiments conducted on a real dataset and real ontology using three different classification algorithms showed that a considerable improvement is achieved by our proposed classification model when the disambiguation issue and the new concept weighting scheme is considered.

The contribution of this research work have been published in peer-reviewed international conferences and journals and an overview of the published articles and their relationships to the contributions of this work is shown in Figure 3.1.

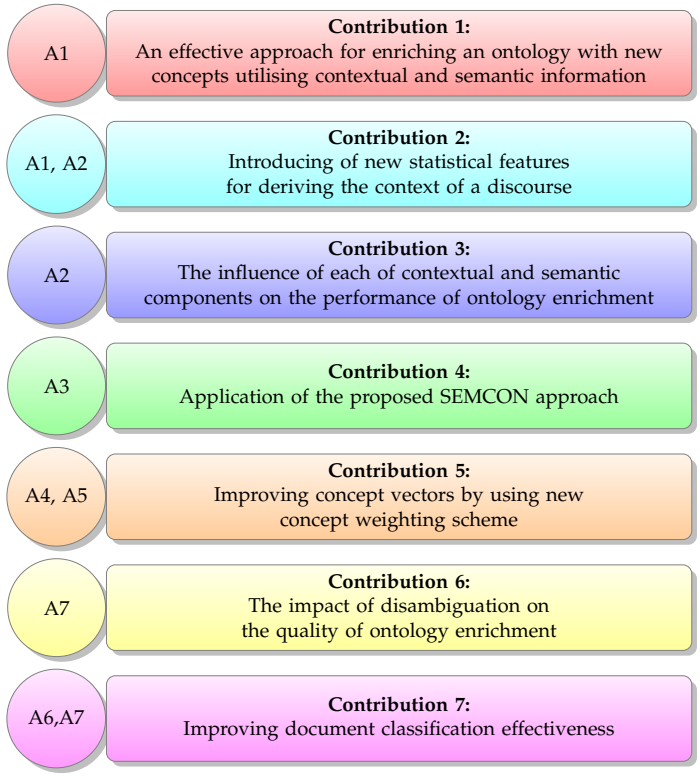
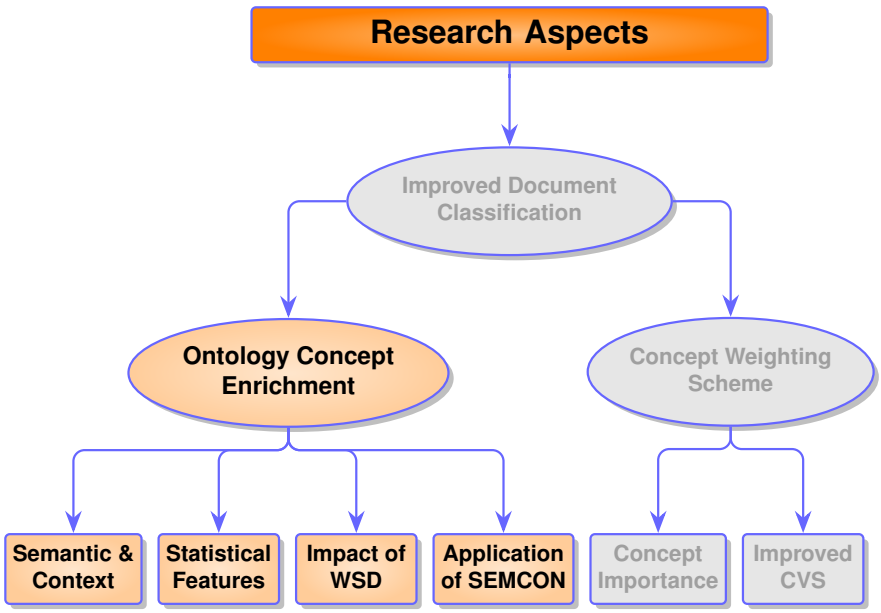


Figure 3.1: An overview of published articles and their relationships to the contributions of this thesis

Part II

Ontology Concept Enrichment



This part includes the first research aspect addressed in this dissertation 'Ontology concept enrichment'. The research work presented in this part aimed to answer the second and the third research question (first part) listed in Section 1.2.

This part contains three chapters constituted by three published research articles. Chapter 4 presents a model called SEMCON capable to enrich an ontology with new concepts utilizing semantic and contextual information. New statistical features are introduced to derive the contextual information of a discourse.

Chapter 5 extended the previous work with an investigation of the impact of the contextual and semantic information, and the statistical features on the enrichment of the ontology. A thorough discussion and analysis along with a more extensive evaluation of the proposed ontology enrichment SEMCON model is also presented in this chapter.

Chapter 6 gives an application of SEMCON approach which is used as the main module of a model capable to analysing Online Social Networks, with a special focus on Facebook. Users's data such as post, comments, feeds, etc, are retrieved using a dedicated web crawler and exploited semantically and contextually through the SEMCON to predict suspect users.

A1: SEMCON: Semantic and Contextual Objective Metric

Publication details
Kastrati, Z., Imran, A., and Yayilgan, S., "*SEMCON: Semantic and Contextual Objective Metric*", in the 9^{th} IEEE International Conference on Semantic Computing (ICSC'15) (2015), IEEE, pp. 65-68.

Chapter 4

SEMCON: Semantic and Contextual Objective Metric

Abstract

This paper proposes a new objective metric called SEMCON to enrich existing concepts in domain ontologies for describing and organizing multimedia documents. The SEMCON model exploits the document contextually and semantically. The preprocessing module collects a document and partitions that into several passages. Then a morpho-syntatic analysis is performed on the partitioned passages and a list of nouns as part-of-speech (POS) is extracted. An observation matrix based on statistical features is then computed followed by computing the contextual score. The semantics is then incorporated by computing a semantic similarity score between two terms - term (*noun*) that is extracted from a document and term that already exists in the ontology as a concept. Eventually, an overall objective score is computed by adding contextual score with semantic score. Subjective experiments are conducted to evaluate the performance of the SEMCON model. The model is compared with state-of-the-art *tf*idf* and χ^2 (Chi square) using F1 measure. The experimental results show that SEMCON achieved an improved accuracy of 10.64% over the *tf*idf* and 13.04% over the χ^2.

4.1 Introduction

Domain ontologies are a good starting point to model in a principled way the basic vocabulary - concepts of a given domain. However, in order for an ontology to be actually usable in real applications, it is necessary to enrich concepts in ontology with available lexical resources of this particular domain. Concepts enrichment means adding new concepts without dealing with their ontological relations and types. Moreover, the ontology structure will remain the same but its concepts will be enriched with their synonyms and homonyms.

Recently, the population of the ontology with lexical data known as onto-terminology [125] has been the subject of research. In this regard, researchers in [37] proposed a new approach named *Synopsis* to automatically building a lexicon for each specific term called criterion. The authors used the assumption that terms appearing closer to a given criterion are more correlated to this criterion. The correlation is simply computed by only counting the number of grammatical terms between a given term and the user criterion. An adaptation of this approach is presented by researchers in [122]. They used the same methodology to build automatically the lexicon of an ontology concept in contrast to building a lexicon for a term. In order to do this, they built an information retrieval system called *CoLexIR* which automatically identifies all parts of a document that are related to a given concept. The issue of enriching the ontology concepts is also treated in [108] where researchers proposed a new methodology to enrich the upper-level ontology SUMO (Suggested Upper Merged Ontology) with the lexical data from the WordNet lexical database.

These approaches, using the the co-occurrence of terms, take into account only the contextual aspects of the domain in their learning process and do not consider the semantics. Therefore, this paper proposes a new approach namely SEMCON, which combines the contextual information and semantic information in the learning process of enriching the

ontology concepts. Furthermore, in addition to frequency of occurrences of common noun terms, new statistical features such as term's font size and term's font type are introduced in this paper to build the observation matrix.

The rest of the paper is organized as follows. Section 4.2 describes the proposed SEMCON model in detail. In section 4.3, we describe the subjective and objective experiment and we compare the subjective results with the results obtained by SEMCON model. Lastly, section 4.4 concludes the paper.

4.2 SEMCON

This section describes the proposed SEMCON model to enrich concepts *c* of a domain ontology with new terms *t*. The model, illustrated in Figure 4.1, consists of 4 modules which are explained in the following subsections.

4.2.1 Preprocessing

This module first collects a document and partitions that into subsets of text known as passages. Each passage is treated as independent document in this paper.

Then a morpho-syntatic analysis is performed on the partitioned passages and the potential terms obtained can either be a noun, verb, adverb or adjective. These are different parts-of-speech (POS) of a language. It is a well established fact that nouns represent the most meaningful terms in a document [84], thus the focus of this paper is on extracting only common noun terms t_n for further consideration.

4.2.2 Observation Matrix

The second module of SEMCON deals with calculation of the observation matrix. The observation matrix is formed using the frequency of occurrences of each term t_n, their font type (*bold*, *underline*, *italic*), and their font size (*title*, *level* 1, *level* 2) as given in Equation 4.1. Using of font type and font size of a term is inspired from the representation of tags in the tag cloud. The font size and position of terms are found to be amongst the very important factors in the information finding process [58]. For instance, the bigger the font size is, the more important a term is in the given context.

$$O_{i,j} = \sum_{i \in t_n} \sum_{j \in p} (\alpha * Freq_{i,j} + \beta * Type_{i,j} + \gamma * Size_{i,j}) \tag{4.1}$$

where, t_n and p denotes the set of terms and set of passages respectively. α, β, γ are some constants set as 1 in our case. $Freq_{i,j}$ denotes the frequency of occurrences of term t_{ni} in passage p_j, $Type_{i,j}$ denotes term's font type t_{ni} in passage p_j, and $Size_{i,j}$ denotes term's font size t_{ni} in passage p_j.

We assumed that terms occurring in bold have more influence/effect on the readers than underline and than italic. According to this assumption, we computed the font type of a term t_n as given in Equation 4.2.

$$Type(t_n) = 0.75 * B + 0.5 * U + 0.25 * I \tag{4.2}$$

Font size of a term t_n is calculated using Equation 4.3.

$$Size(t_n) = 1.0 * T + 0.75 * L_1 + 0.50 * L_2 + 0.25 * L_3 \tag{4.3}$$

where *T* indicates title font size, L_1 indicates level 1 font size, L_2 indicates level 2 font size, L_3 indicates level 3 font size, *B* indicates the bold font type, *U* indicates underlined font type, and *I* indicates the italic font type.

The computation of each term's font size in the observation matrix is performed using the font sizes from a master slide in PowerPoint presentations where the level 1 font size is set to 28 pt, level 2 is set to 24 pt and level 3 is set to 20 pt.

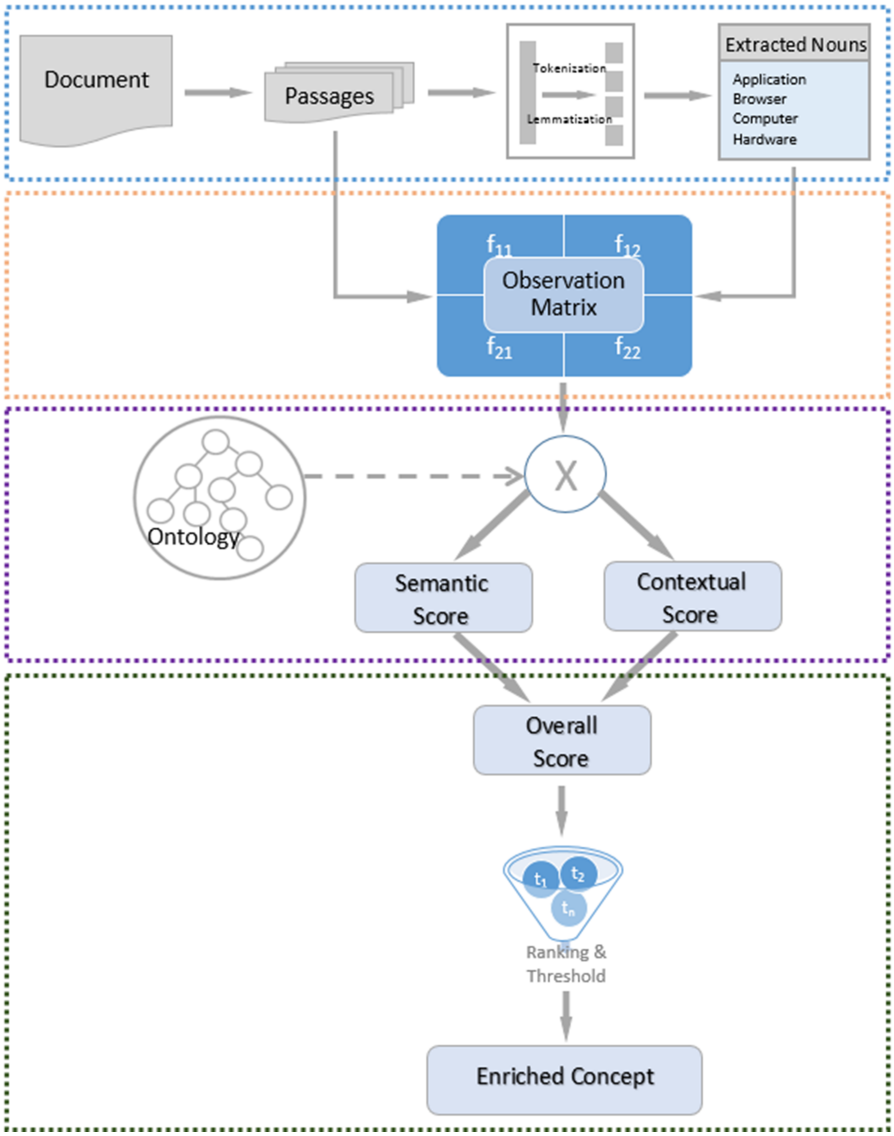


Figure 4.1: Block diagram of SEMCON model.

4.2.3 Contextual and Semantic Similarity

The observation matrix is used as input to compute the contextual and semantic similarity between two terms in order to match a term extracted from a passage with a concept in the ontology.

Term to term contextual distance, given in Equation 4.4, is computed using the cosine

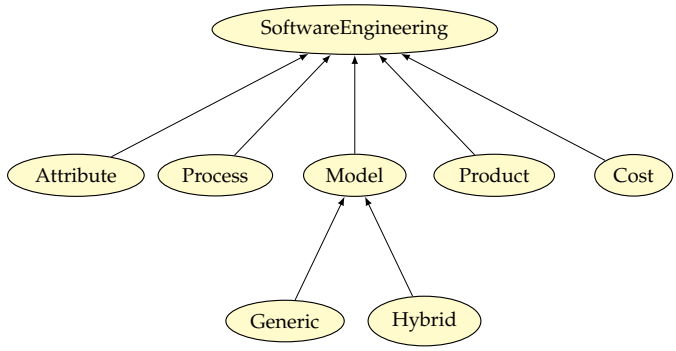


Figure 4.2: Software engineering lightweight ontology

measure in respect of passages.

$$S_{con}(t_{n1}, t_{n2}) = \frac{t_{n1} \cdot t_{n2}}{\| t_{n1} \| \| t_{n2} \|} \tag{4.4}$$

A term square matrix is used to store S_{con} values among all extracted terms t_n. This matrix will later be used in computing an overall correlation between a term extracted from a document and a concept in the ontology.

Further, we extract and use a subset of the terms t_n in order to extend the concept list of ontology. There maybe single label concepts in an ontology as well as compound label concepts. For single label concepts, we use only those terms from the term square matrix for which an exact term exists in the ontology. For example, for concept "Attribute" or "Generic" in the software engineering ontology shown in Figure 4.2, there exists exactly a same term extracted in the term square matrix.

For compound label concepts, we use those terms from the term square matrix which are present as part of a concept in the ontology. For example, consider "SoftwareEngineering" as one of the compound label ontology concept, and the "Software" as one of the extracted terms from the document. Let "Program", "Design", "Development" be the highly correlated terms with the term "Software". In this case, the compound ontology concept "SoftwareEngineering" will be enriched with the correlation terms of the term "Software" namely with "Program", "Design", "Development".

The next step is the computation of the semantic similarity. The semantic similarity score, given in Equation 4.5, is calculated using the Wu&Palmer algorithm [156] implemented in a freely available software package WordNet::Similarity [115].

$$S_{sem}(t_n, c) = \frac{2 * depth(lcs)}{depth(t_n) + depth(c)} \tag{4.5}$$

where t_n, indicates term extracted from document, c denotes term that already exists in ontology as a concept, *depth(lcs)* indicates least common subsumer of term and concept label, *depth(t_n)* indicates the path's depth of term in WordNet::similarity and *depth(c)* indicates path's depth of concept label in WordNet::similarity.

4.2.4 Overall Score

The overall correlation of a term extracted from a document and a concept in the ontology is computed using the contextual and semantic score and it is given in Equation 4.6.

$$S_{overall}(t_n, c) = w * S_{con}(t_n, c) + (1 - w) * S_{sem}(t_n, c) \tag{4.6}$$

where w is a parameter with value set as 0.5 in our case.

Finally, in order to obtain the terms which are more closely related to the ontology concepts, a rank cut-off method is applied to the terms t_n using a specified threshold. Terms which are above the threshold are considered to be the relevant terms for enriching the ontology concepts.

4.3 Experimental Procedures

To evaluate the performance of SEMCON, we have used PowerPoint presentations dataset from 5 different domains: Computer, C++ Programming, Database, Internet, and Software Engineering. We were restricted to a maximum of 5 presentations with a limited number of slides (39 slides), due to subjective nature of the experiment.

The paper uses two approaches to evaluate the performance of SEMCON. The first one is subjective evaluation and the second one is the objective evaluation. The results from software engineering domain are presented in this paper.

4.3.1 Subjective evaluation

To compare term to concept correlation obtained from SEMCON, an online survey based experiment is conducted. The subjective survey was carried out by publishing online a questionnaire to 10 subjects. The subjects were all computer science PhD students and postdocs at the Gjøvik University College. They were asked to select 5 closely related terms from a list of terms for each concept, for 5 different domains, starting from the most relevant term as their first choice, the second relevant term as the second choice and so on.

From the subjective survey, a single score, for each selected term, is calculated using the Borda count method. Borda count, given in Equation 4.7, is an election method used to determine a winner from a voting where voters rank the candidates in order of preference [160].

$$BordaCount(t_n) = \sum_{i=1}^{m}[(m+1-i) * freq_i(t_n)] \tag{4.7}$$

where *BordaCount(t_n)* of a given term t_n is calculated by a total sum of the weights of the frequencies *$freq_i(t_n)$*. *$freq_i(t_n)$* is the frequency of term t_n chosen at Position i, and m is the total number of possible positions, in our case m = 5.

The scores from the Borda count are then sorted to obtain the top 'n' terms, giving us the refined list of the highest scoring terms. For our experiment, we set n = 10, and this gives us the top 10 terms as shown in Table 4.1. The term "Waterfall" has the highest Borda count value cause this term is selected by most of the subjects as the closest term for term "Generic".

Table 4.1: Borda count of subjects' responses for "Generic" concept

Rank	Term	Borda Count
1	Waterfall	36
2	Model	16
3	Generic	10
4	Specification	10
5	System	10
6	Design	8
7	Transformation	7
8	Development	6
9	Phase	5
10	Formal	4

4.3.2 Objective evaluation

The second approach used to evaluate the performance of SEMCON is comparing the results obtained from the SEMCON with results obtained from the *tf*idf* and χ^2.

*tf*idf* is a mathematical algorithm which is used to find key vocabulary that best represents the texts by applying the term frequency and the inverted document frequency together [133].

The traditional *tf*idf* considers only the term to document relation and thus it is not appropriate for comparison as it is. Therefore, in order to take into account the term to term relation, cosine measure is used where the dot product between two vectors of *tf*idf* matrix reflects the extent to which two terms have a similar occurrence pattern in the vector space.

χ^2 is a statistical measurement which computes the degree of interdependency between any two terms [85]. The measurement is carried out by comparing the observation frequency with expected frequency.

We evaluated the performance of objective methods using the top terms scored by these methods. In order to do this, scores for the 10 top terms are taken as the ground truth, and they are compared with the top terms obtained by the objective scores. We used the top 15 terms as the refined terms list, and the effectiveness of objective metrics using the standard information retrieval measures are computed in order to compare with the subjective results. These measures are Precision, Recall and F1. Precision is the number of correctly retrieved terms, while recall is the number of retrieved terms. The F1 is considered as average of precision and recall.

Table 4.2 shows precision, recall and F1 results obtained from the SEMCON for software engineering concepts.

Table 4.2: The performance of SEMCON

Concept	Precision (%)	Recall (%)	F1 (%)
Software	40.0	60.0	48.0
Cost	40.0	60.0	48.0
Product	40.0	60.0	48.0
Attribute	46.7	70.0	56.0
Process	60.0	90.0	72.0
Generic	60.0	90.0	72.0
Hybrid	60.0	90.0	72.0
Average	**49.5**	**74.3**	**59.4**

The performance of SEMCON, in terms of F1 measure, is compared with the performance of *tf*idf* and χ^2. The comparison, depicted in Table 4.3, shows that the SEMCON has achieved an improvement on finding the most related terms to enrich the concepts of an ontology, of 10.64% over the *tf*idf* and 13.04% over the χ^2. This improvement is achieved for all concepts excepts for "Software" and "Product". This may have happened due to the fact that the SEMCON, in contrast to *tf*idf* and χ^2, takes into consideration not only the frequency of occurrences of those terms in the corpus but also the semantics of those terms.

An example of a concept ontology enriched with new terms obtained by SEMCON is shown in Figure 4.3. The "Generic" concept of the software engineering ontology is enriched with new terms such as "System", "Development", "Formal" and "Transformation". These terms are amongst the top 10 terms selected also by subjects in the subjective experiment.

Finally, we evaluated the performance of the objective methods to a larger dataset comprised of lightweight ontologies from domains such as computer, database, internet, and C++ programming (C++). The same experiment, as per Software Engineering domain ontology, was conducted. The obtained results in terms of F1 measure indicated in Table 4.4

Table 4.3: The performance of objective methods using the F1 measure.

Concept	tf*idf (%)	χ^2 (%)	SEMCON (%)
Software	56.0	56.0	48.0
Cost	40.0	40.0	48.0
Product	64.0	56.0	48.0
Attribute	32.0	48.0	56.0
Process	72.0	48.0	72.0
Generic	48.0	64.0	72.0
Hybrid	64.0	56.0	72.0
Average	**53.7**	**52.6**	**59.4**

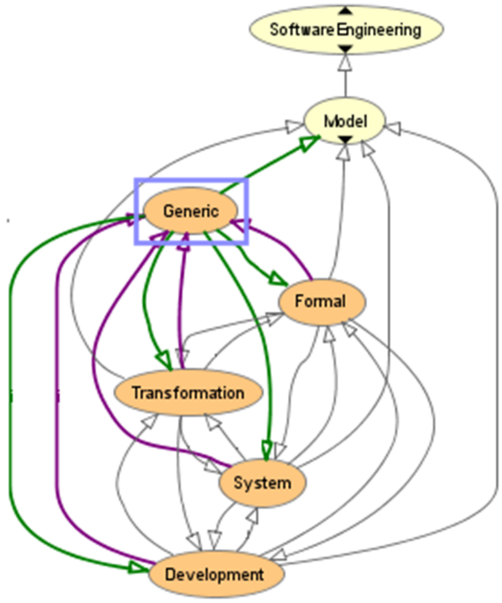


Figure 4.3: Generic concept enriched with new terms

Table 4.4: F1 measure for 5 different domains

Domain name	tf*idf (%)	χ^2 (%)	SEMCON (%)
Computer	44.44	40.89	49.78
C++_Programming	43.20	43.20	44.80
Database	40.00	32.10	41.00
Internet	49.14	41.14	45.71
SoftwareEngineering	53.71	52.57	59.43

show that the SEMCON gives better results then both of the methods for all domains excepts for internet domain ontology. This may have happened due to the fact that subjects are making their selections based on descriptions provided under each concept on the questionnaire, when they were asked to select the 5 more closely related terms. Therefore, this causes the overall score to be mainly affected by the contextual score.

4.4 Conclusion

This paper proposed a new objective metric namely SEMCON to enriching the domain ontology with new concepts by combining contextual as well as semantics of a term. The proposed method can be applied to any existing domain ontology for extending it with new concepts. The SEMCON takes into account the context of a term by first computing an observation matrix which exploits the statistical features. Currently three features - frequency of the occurrence of a term, term's font type and font size are used to compute observation matrix. These features can easily be extended based on the type of the document chosen. The future work may exploits further features for calculating observation matrix, and extracting candidate terms from multiple documents including word documents, audio and video files. We also plan to conduct further research to examine the contribution of the contextual and semantic scores in the overall score.

A2:
SEMCON: A Semantic and Contextual Objective Metric for Enriching Domain Ontology Concepts

Publication details
Kastrati, Z., Imran, A., and Yayilgan, S., "*SEMCON - A Semantic and Contextual Objective Metric for Enriching Domain Ontology Concepts*", International Journal on Semantic Web and Information Systems - IJSWIS (2016), vol. 12, issue 2, pp. 1-24.

Chapter 5

SEMCON: A Semantic and Contextual Objective Metric for Enriching Domain Ontology Concepts

Abstract

This paper presents a novel concept enrichment objective metric combining contextual and semantic information of terms extracted from the domain documents. The proposed metric is called SEMCON which stands for semantic and contextual objective metric. It employs a hybrid learning approach utilizing functionalities from statistical and linguistic ontology learning techniques. The metric also introduced for the first time two statistical features that have shown to improve the overall score ranking of highly relevant terms for concept enrichment. Subjective and objective experiments are conducted in various domains. Experimental results (F1) from computer domain show that SEMCON achieved better performance in contrast to *tf*idf*, χ^2 and *LSA* methods, with 12.2%, 21.8%, and 24.5% improvement over them respectively. Additionally, an investigation into how much each of contextual and semantic components contributes to the overall task of concept enrichment is conducted and the obtained results suggest that a balanced weight gives the best performance.

5.1 Introduction

Domain ontologies are a good starting point to model in a formal way the basic vocabulary of a given domain. They provide a broad coverage of concepts and their relationships within a particular domain. However, in-depth coverage of concepts is often not available, thereby limiting their use in specialized subdomain applications. It is also the business dynamics and changes in the operating environment which requires modification to an ontology [97]. Therefore, the techniques for modifying ontologies, i.e. ontology enrichment, have emerged as an essential prerequisite for ontology-based applications. An ontology can be enriched with lexical data either by populating the ontology with lexical entries or by adding terms to ontology concepts. The former means updating the existing ontology with new concepts along with their ontological relations and types. This increases the size of the existing ontology which requires more computational resources and more time to compute. Thus making it less cost effective. The latter means adding new concepts without taking into account the ontological relations and types between concepts. As a result of this, the ontology structure will remain the same but its concepts will be enriched with their synonyms and homonyms.

Enrichment of ontology concepts aims to improve a given ontology by updating it with similar concepts. It is part of an iterative ontology engineering process [40] and it involves subtasks from only lower part of ontology learning layer cake model [25]. Acquisition of the relevant terminology, identification of synonym terms or linguistic variants and the formation of concepts are subtasks involved. To perform these subtasks, the enrichment process requires an initial ontology which has to be enriched. It then explores available documents and texts from related domain of the given ontology in order to find synonyms

or linguistic variants. Finally, by employing the learning approach, which is the core of an ontology concepts's enrichment process, the concepts are ready for updating the initial given ontology.

There is a variety of learning approaches that are available to enrich concepts of an ontology. These approaches relies on either linguistic, pattern matching, machine learning or statistical techniques [36, 59]. Even though these approaches have been proved useful for enriching ontologies of many domains, they however have some limitations. These approaches use only contextual information without taking into account the semantic information of terms. The contextual information is derived by distributional property of terms such as term frequency or *tf*idf*, and co-occurrence of terms. Therefore, to address this limitation, this paper proposes a new objective metric namely SEMCON to enriching the domain ontology with new concepts by combining contextual as well as semantics of a term.

The new proposed objective metric uses unstructured data as input for ontology learning process and is composed of two parts - contextual and semantic. Context is defined as the part of a text or statement passage that surrounds a given term and it determines term meaning. In our work, it is the cosine distance between the feature vectors of any two terms. The feature vectors are composed of values computed by both the frequency of occurrence of terms in corresponding passages, and the statistical features such as font type and font size. The semantics on the other hand is defined by computing a semantic similarity score using lexical database WordNet.

In addition, we also have investigated into how much each of contextual and semantic components contributes to the overall task of enriching the domain ontology concepts and compared our results with the results obtained by other approaches such as *tf*idf*, χ^2 and *LSA*. We present our results for several domains, namely, Computer, Software Engineering, C++ Programming, Database and Internet.

The rest of the paper is organized as follows. Section 5.2 presents the state of the art in the field of ontology enrichment. Section 5.3 describes our proposed SEMCON model in detail. In Section 5.4 we describe the experiments including subjective and objective evaluation of SEMCON along with measures used to evaluate the effectiveness of objective methods. Results obtained by SEMCON and other objective methods and their comparisons are shown in Section 5.5. Section 5.6 presents some of the application areas of SEMCON and lastly, Section 5.7 concludes the paper and gives some future work directions.

5.2 Related Work

The field of ontology learning from unstructured data has attracted a lot of attention recently, resulting in a wide variety of approaches. There are two main categories of these approaches relevant to the concept enrichment task: 1) Statistical approach, and 2) Linguistic approach.

Statistical approach uses distributional property of terms such as term frequency (*tf*) or term frequency inverse document frequency (*tf*idf*) and term co-occurrence to identify concepts within the textual data. An example of statistical approach as learning technique is DOODLE II [158]. It exploits a machine readable dictionary and domain-specific texts to build domain ontologies with both taxonomic and non-taxonomic conceptual relationships. The non-taxonomic relationships are dependencies between concepts such as synonymy, meronymy, antonymy, attribute-of, possession. These non-taxonomic relationships are exploited using domain specific texts with the analysis of lexical co-occurrence statistics, based on WordSpace, which follows the idea that terms that occur together can have non-taxonomic conceptual relationships. WEB→KB [28] is another ontology learning system which relies on statistical approach to locate and extract concepts from world wide web documents.

Other statistical approaches deal with batches of terms. These approaches follow the idea that the meaning of a term is represented by term co-occurrences and the frequencies of the co-occurrences [89]. The occurrence of two or more terms within a sentence, a passage or a document is known as collocation [63]. Learning by term collocations and co-occurrences is the most addressed technique in statistical concept learning approach. TEXT-TO-ONTO [26, 90] is an example which employs learning by term collocations and co-occurrences technique. It uses textual data as input and computes the frequency of term co-occurrences to identify and extract non-taxonomic relations using background knowledge like a lexicon and a taxonomy. SYNOPSIS [37] is another example of learning by term collocations and co-occurrences. It is a system which automatically builds a lexicon for each specific term called criterion. To identify lexicon terms, the researchers use the partition technique to split the document into several passages. The correlation between terms and the user criterion is computed based on the relative position between each term and the criterion. In other words, the correlation is simply computed by only counting the number of grammatical terms between a given term and the user criterion. This way a lexicon is built for each criterion. An adaptation of SYNOPSIS, namely CoLexIR, is presented in [122]. CoLexIR uses the same methodology as SYNOPSIS but it is used to build the lexicon of an ontology concept automatically.

While statistical approaches depend on term frequencies and co-occurrences, linguistic approaches involve natural language processing techniques, such as syntactic, morpho-syntactic and lexico-syntactic analysis to identify concepts from textual data. An example of linguistic approach as learning technique is HASTI [135]. HASTI is an automatic ontology building method which uses a combination of morpho-syntactic and semantic analysis techniques. Its input is unstructured data in the form of natural language texts in Persian. The ontology in HASTI is a small kernel whose lexicon is nearly empty initially and grows gradually by learning new terms. It learns concepts, taxonomic and non-taxonomic conceptual relations, and axioms, to build ontologies on top of the existing kernel.

KnowItAll [38, 39] is another approach which is dependent on natural language processing technique to extract information. It is a domain-independent system that extracts information from the Web. KnowItAll employs lexico-syntactic patterns approach to identify and extract possible concepts. It selects the concepts by evaluating concept plausibility derived using a version of the pointwise mutual information statistical measure.

SynDiKATe [55] is an ontology learning method based on natural language processing. It uses technical documents in German language taken from test reports from the information technology domain and medical finding reports. The approach to learning new concepts is based on syntactic analysis in both sentence level and text level. The result of the syntactic analysis is captured in a dependency graph, where vertices represent terms and edges represent relations between those terms.

Categorizing SEMCON model in one of the two main categories of approaches of concept enrichment is not an easy task due to differences which exist in many dimensions amongst approaches. Shamsfard and Barforoush [135] identified six main categories of the major distinguishing factors between ontology learning approaches. Even though there are differences amongst approaches, they however have some points in common. From this perspective, SEMCON can be considered as a hybrid approach which to some extent utilizes both approaches, linguistic and statistical. From the linguistic point of view, SEMCON uses morpho-syntactic analysis to identify and extract noun terms, as part of speech, which represent the most meaningful terms in a document. From statistical point of view, SEMCON derives the context using cosine similarity between term vectors whose members are frequencies of terms. SEMCON brings, besides term frequency, two new statistical features to the table, i.e. term font size and font type, to determine the context of term. In addition to contextual information, SEMCON also incorporates the semantic information of terms using the lexical database WordNet and finally aggregates both contextual and semantic information of this particular term.

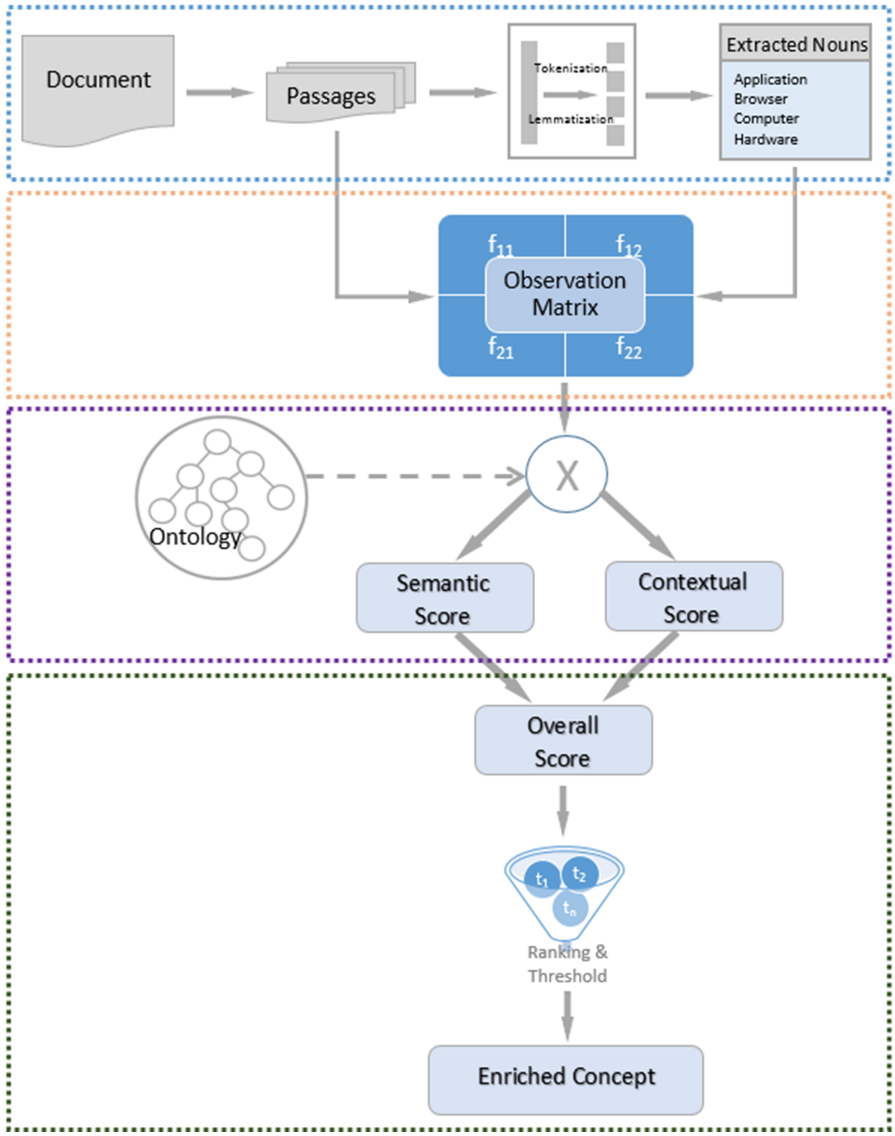


Figure 5.1: Block diagram of SEMCON

5.3 SEMCON

This section describes the proposed SEMCON model to enrich concepts of a domain ontology with new terms which are closely related using the contextual and semantic information. The model, illustrated in Figure 5.1, consists of four modules, which are explained in the following subsections.

5.3.1 Preprocessing

This module initially collects a document and partitions that into subsets of text known as passages. These passages are text portions which have very strong semantic coherence and are clearly disconnected from adjacent parts [129]. The partitioned passages can either be fixed or variable length. They can also be classified into contextual passages if the partitioning takes into account the context of the document or they can be classified as statistical passages.

In this paper, we take into account the context of a document irrespective of the length of partitioned passages. Partitioned passages are treated as independent documents. A morpho-syntactic analysis using TreeTagger [132] is performed on the partitioned passages. Passages are later cleaned by removing all punctuation and capitalization followed by a tokenizer step to separate the text into individual terms. The lemmatization is the last step used to find the normalized form of these terms.

The potential terms that are obtained as a result of this preprocessing step can either be a noun, verb, adverb or adjective. These are different parts-of-speech (POS) of a language. It is a well established fact that nouns represent the most meaningful terms in a document [84], thus our focus is on processing only noun terms for further consideration.

5.3.2 Observation Matrix

Computation of the observation matrix is the next step in the proposed model. Observation matrix is a rectangular matrix where the rows represent the extracted passages from a particular document and columns are the terms extracted from those particular passages. An example of observation matrix is shown in Table 5.1.

Table 5.1: A part of the observation matrix from computer domain

Slide	Computer	Data	Device	Function	Hardware	System	Web
1	6.25	5.25	1.75	0	0	0	0
2	3	0	0	0	0	1.5	8
3	9.25	0	7	1.75	4.75	5.5	0
4	5.5	3.5	8	0	0	0	0
5	5	1.5	1.5	1.5	0	2	0
6	12.25	0	0	1.5	0	6	0
7	2.25	0	0	0	0	6.25	0

Each entry of the observation matrix is calculated by accumulating the sum of term frequency, term font size and term font type in each of the extracted passages, as shown in Equation 5.1. Introducing of term font type and term font size, as very important factors in the information finding process [58], is inspired from the representation of tags in the tag cloud [10]. The effect of these statistical features is discussed in subsection 5.5.1.

$$O_{i,j} = \sum_{i \in t} \sum_{j \in p} (Freq_{i,j} + FT_{i,j} + FS_{i,j}) \tag{5.1}$$

where, t and p show the set of terms and passages, respectively. $Freq_{i,j}$ denotes the frequency of occurrences of a term t_i in passage p_j. $FT_{i,j}$ and $FS_{i,j}$ denote font type and font size of a term t_i in passage p_j, respectively.

We adopt a linear increase model for different font types and font sizes. The linear model assumes that the effect of each variable is the same for all values of the other variables. For example, the model assumes that the effect of bold font type terms is exactly the same as the effect of every underline or italic font type terms. The same way, the effect of underlined font type terms is exactly the same as the effect of every underline bold or italic font type terms, and so on. Font type of a term t is calculated using Equation 5.2, while font

size is calculated using Equation 5.3. Both functions are in normalized form and the results lie between 0 and 1.

$$FT(t_{i,j}) = 0.75 * B + 0.5 * U + 0.25 * I \tag{5.2}$$

$$FS(t_{i,j}) = 1.0 * T + 0.75 * L_1 + 0.50 * L_2 + 0.25 * L_3 \tag{5.3}$$

The font sizes and the font types of terms used to build the observation matrix can be derived for all types of rich text documents using the html tags. In this paper, we used the font sizes from the presentations slides where the level 1 font size is set to 28 pt, level 2 is set to 24 pt and level 3 is set to 20 pt. These parameters can be adjusted for other document types. According to these font size settings, we observed the occurrences of terms among the presentation slides.

The example illustrated in Figure 5.2 shows that term *Web* occurred 4 times in the presentation slides, where 2 times it appeared as level 1 font size and as bold font type and 2 times it appeared as level 2 font size.

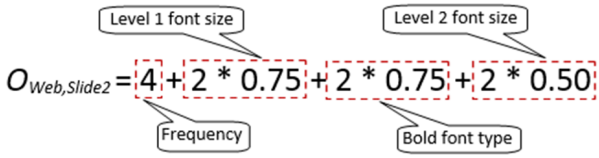


Figure 5.2: Building of observation matrix using statistical features

5.3.3 Computation of Contextual and Semantic Score

The observation matrix is used as an input to compute the term-to-term contextual and semantic score between two terms in order to find a matching term extracted from a passage to a concept in the ontology.

Term to term contextual score (S_{con}) is calculated using the cosine similarity metric with respect to the passages, as given by Equation 5.4.

$$S_{con}(t_i, t_j) = \frac{t_i \cdot t_j}{\| t_i \| \| t_j \|} \tag{5.4}$$

A term square matrix is used to store S_{con} values among all extracted term. This matrix will later be used in computing an overall correlation between a term extracted from a document and a concept in the ontology, as described in subsection 5.3.4.

Further, the proposed model extracts and uses a subset of terms *t* to extend and to enrich ontology concepts. There may be single label concepts in an ontology as well as compound label concepts. For single label concepts, SEMCON uses only those terms from the term square matrix for which an exact match exists in the ontology. For example, for concept in the ontology such as Application or Storage illustrated in Figure 5.3, there exists exactly the same term in the term square matrix.

For compound label concepts, SEMCON uses those terms from the term square matrix which are present as part of a concept in the ontology. For example, consider *InputAndOutputDevices* as one of the compound ontology concepts, and the Device as one of the terms in the term square matrix. Let *Screen*, *Display*, *Input* be the highly correlated terms with the term *Device*, and in that case, the *InputAndOutputDevices* will be enriched with the correlation terms of the term *Device* e.g. with *Screen*, *Display*, *Input*.

Next step is the computation of the semantic score S_{sem}. The semantic score is computed using WordNet database. WordNet [44] is a lexical database for English language

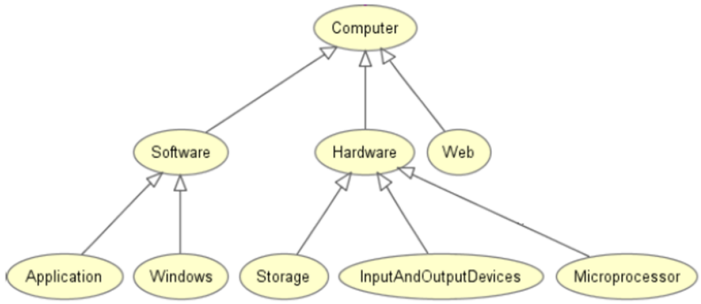


Figure 5.3: Ontology sample of the computer domain

that groups terms into sets of synonyms called synsets and defines the semantic relations between these synsets. SEMCON uses all the synsets to represent specific terms under consideration.

The semantic score, $S_{sem}(t_i, t_j)$, is calculated for all possible pairs t_i and t_j from the observation matrix, where t_i, $t_j \in O$ and O is the observation matrix. As a result, for each term, a hash table is generated where the most similar terms are set as the synonyms for that term. The Wu&Palmer algorithm [156] is used to compute the semantic score. Mathematically, it is computed using Equation 5.5.

$$S_{sem}(t_i, t_j) = \frac{2 * depth(lcs)}{depth(t_i) + depth(t_j)} \tag{5.5}$$

where, *depth(lcs)* indicates the least common subsumer of terms t_i and t_j; *depth(t_i)* and *depth(t_j)* indicate the path's depth of terms t_i and t_j, in the WordNet lexical database.

5.3.4 Overall Score

The overall correlation between two terms, t_i and t_j, is calculated using the contextual and semantic score. Mathematically, the overall score is given in Equation 5.6.

$$S_{ove}(t_i, t_j) = w * S_{con}(t_i, t_j) + (1 - w) * S_{sem}(t_i, t_j) \tag{5.6}$$

where w is a parameter with value set as 0.5 based on the empirical analysis performed on the data set given in Section Experimental Procedure. A thorough analysis about the effect of the weight parameter value on the output of the SEMCON is given in subsection 5.5.2. The overall score is in the range (0,1]. The overall score is 1 if two terms are the same and 0 when there is no relationship between them.

Finally, a rank cut-off method is applied using a threshold to obtain terms which are closely related to a given term in the ontology. Terms that are above the specified threshold (top-N) are considered to be the relevant terms for enriching the concepts.

A simple example of the SEMCON output, given in Table 5.2, shows the top 10 terms obtained as the most relevant terms of *Application* concept. 6 of these terms, namely *Application*, *Program*, *Apps*, *Function*, *Task* and *Software* are amongst the top 10 terms selected by the subjects as the closest terms to concept *Application*.

5.4 Experimental Procedures

The experiment used presentation slides dataset from 5 different domains as shown in Table 5.3. The presentations in the database are from domain of Computer, Database, Internet,

Table 5.2: Top 10 closely related terms of Application concept

Concept	The Top 10 terms obtained by SEMCON model
Application	Apps, Application, Software, Program, Control Task, Part, Master, Operation, Function

C++ Programming and Software Engineering. The dataset was limited to a maximum of 5 presentations with a restricted number of slides due to the subjective nature of the experiment.

Table 5.3: Dataset used for experimenting

No	Domain name	# of slides	# of terms	# of concepts
1	Computer	7	79	9
2	Database	9	105	8
3	Internet	7	73	7
4	C++_Programming	9	70	10
5	Software_Engineering	7	42	7

The paper uses two approaches to evaluate the performance of the SEMCON. The first one is the subjective evaluation and the second one is the objective evaluation.

5.4.1 Subjective Evaluation

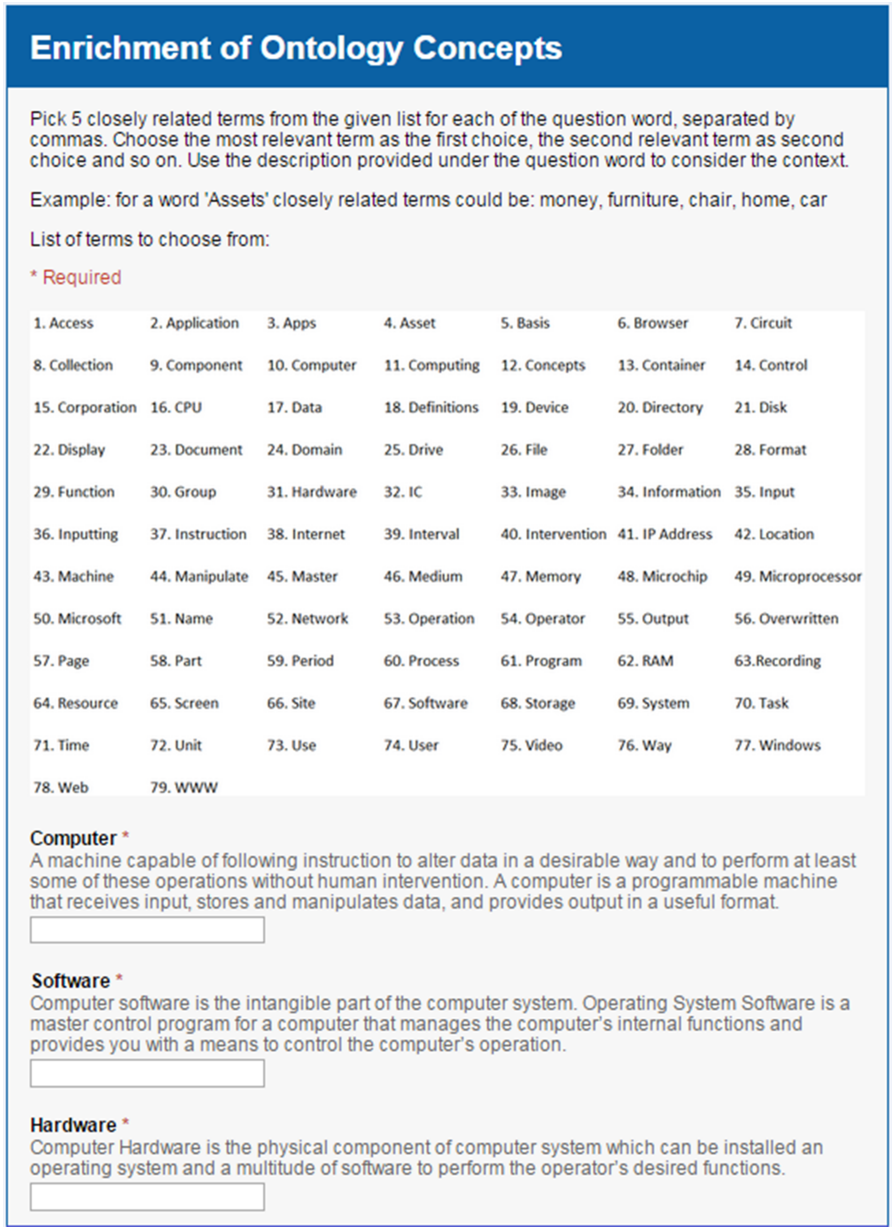
To evaluate the performance of SEMCON, a subjective survey was carried out by publishing an online questionnaire to 15 subjects.

The subjects were all computer science PhD students and Postdocs at the Gjøvik University College. They were asked to select 5 closely related terms from a list of terms for each of the concepts, starting from the most relevant term as their first choice, the second relevant term as the second choice and so on. A screenshot taken from the questionnaire about the computer domain is illustrated in Figure 5.4.

For each of the concepts given in the subjective survey, we obtained the ranking of the corresponding term and its frequency count. An example of ranking terms and calculating the counts of the corresponding term frequencies is given in Table 5.4.

Table 5.4: Terms selected by subjects for the Application concept

Terms	Pos 1	Pos 2	Pos 3	Pos 4	Pos 5	Total
Apps	2	8	2	1	0	13
Software	3	0	4	3	1	11
Program	0	2	3	1	2	8
Application	8	0	0	0	0	8
User	0	1	1	1	3	6
Task	0	0	2	2	1	5
Windows	1	1	0	2	1	5
Browser	0	0	2	2	0	4
Process	0	0	1	1	2	4
Microsoft	1	1	0	0	0	2
System	0	0	0	0	2	2
Computer	0	1	0	0	0	1
Data	0	0	0	0	1	1
Recording	0	0	0	0	1	1



Enrichment of Ontology Concepts

Pick 5 closely related terms from the given list for each of the question word, separated by commas. Choose the most relevant term as the first choice, the second relevant term as second choice and so on. Use the description provided under the question word to consider the context.

Example: for a word 'Assets' closely related terms could be: money, furniture, chair, home, car

List of terms to choose from:

* Required

1. Access	2. Application	3. Apps	4. Asset	5. Basis	6. Browser	7. Circuit
8. Collection	9. Component	10. Computer	11. Computing	12. Concepts	13. Container	14. Control
15. Corporation	16. CPU	17. Data	18. Definitions	19. Device	20. Directory	21. Disk
22. Display	23. Document	24. Domain	25. Drive	26. File	27. Folder	28. Format
29. Function	30. Group	31. Hardware	32. IC	33. Image	34. Information	35. Input
36. Inputting	37. Instruction	38. Internet	39. Interval	40. Intervention	41. IP Address	42. Location
43. Machine	44. Manipulate	45. Master	46. Medium	47. Memory	48. Microchip	49. Microprocessor
50. Microsoft	51. Name	52. Network	53. Operation	54. Operator	55. Output	56. Overwritten
57. Page	58. Part	59. Period	60. Process	61. Program	62. RAM	63.Recording
64. Resource	65. Screen	66. Site	67. Software	68. Storage	69. System	70. Task
71. Time	72. Unit	73. Use	74. User	75. Video	76. Way	77. Windows
78. Web	79. WWW					

Computer *
A machine capable of following instruction to alter data in a desirable way and to perform at least some of these operations without human intervention. A computer is a programmable machine that receives input, stores and manipulates data, and provides output in a useful format.

Software *
Computer software is the intangible part of the computer system. Operating System Software is a master control program for a computer that manages the computer's internal functions and provides you with a means to control the computer's operation.

Hardware *
Computer Hardware is the physical component of computer system which can be installed an operating system and a multitude of software to perform the operator's desired functions.

Figure 5.4: A screenshot taken from the questionnaire

Pos(n) is the position of the selected term from the term list. It shows how many times a particular term is selected at n^{th} position, e.g. the term *Apps* is chosen by 2 subjects as their 1^{st} choice for the *Application* concept, by 8 subjects as their 2^{nd} choice and so on. The total number of times a particular term being selected by subjects for the *Application* concept is

Table 5.5: Borda count of subjects' responses for the Application concept

Rank	Term	Borda Count
1	Apps	50
2	Application	40
3	Software	34
4	Program	21
5	Windows	14
6	Task	11
7	Browser	10
8	Function	9
9	User	9
10	Process	7

computed by aggregating all these frequencies together.

For each selected term, a single score is computed using the Borda Count method. Borda Count method is an election method used to determine a winner from a voting where voters rank the candidates in order of preference [160]. The mathematical formulation of Borda Count is given in Equation 5.7.

$$BC(t) = \sum_{i=1}^{m} [(m + 1 - i) * freq_i(t)] \tag{5.7}$$

where *BC(t)* of a given term *t* is computed by a total sum of the weights of the frequencies *freq$_i$(t)*. *freq$_i$(t)* is the frequency of term *t* chosen at position *i*, and *m* is the total number of possible positions, in our case 5.

The scores from the Borda Count are then sorted to obtain the top 'n' terms, giving us the refined list of the highest scoring terms. For our experiment, we set n = 10, and this gives us the top 10 terms as shown in Table 5.5. This is our ground truth data.

5.4.2 Objective Evaluation

In addition to the subjective experiment, an objective evaluation is carried out where the results obtained from the SEMCON model are compared with the results obtained from the three state-of-the-art methods namely Term Frequency Inverse Document Frequency (*tf*idf*) [133], χ^2 (Chi square) [85] and Latent Semantic Analysis - LSA [80].

*tf*idf* is a mathematical method which is used to find key vocabulary that best represents the texts. Mathematically, it is given in Equation 5.8.

$$tf * idf = tf_{i,j} * \log \frac{N}{df_j} \tag{5.8}$$

where, $tf_{i,j}$ is the term frequency of term *j* that occurs in a passage, *N* is the total number of passages in the corpus and df_j shows the number of passages where the term *j* occurs.

The traditional *tf*idf* considers only the term to document relation and thus it is not appropriate for comparison as it is. Therefore, we modified the existing *tf*idf* in order to take the term to term relation into account. This is achieved using the cosine measure where the dot product between two vectors of *tf*idf* matrix shows the extent to which two terms are similar in the vector space.

χ^2 is a statistical method which computes the relationship between two given terms. Mathematically, it is given in Equation 5.9.

$$\chi^2_{t_a,t_b} = \sum_{i \in \{t_a, \neg t_a\}} \sum_{j \in \{t_b, \neg t_b\}} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} \tag{5.9}$$

where, $O_{i,j}$ and $E_{i,j}$ show the co-occurrence and the expected co-occurrence frequency between two terms t_a and t_b. More formally, the co-occurrence frequency between two terms t_a and t_b is the observed frequency $O_{i,j}$ where $i \in \{t_a, \neg t_a\}$ and $j \in \{t_b, \neg t_b\}$. Thus, O_{t_a,t_b} is the observed frequency of passages which contain term t_a and term t_b. $O_{t_a,\neg t_b}$ is the observed frequency of passages which contain term t_a but do not contain term t_b. $O_{\neg t_a,t_b}$ is the observed frequency of passages which do not contain t_a but contain the term t_b. $O_{\neg t_a,\neg t_b}$ is the observed frequency of passages which contain neither term t_a nor term t_b.

Latent semantic analysis (LSA), sometimes referred as latent semantic indexing, is a method for extracting and representing the content of a text using the relationships between terms that occur in similar context.

The first step of LSA is representing the text document as a matrix in which each row denotes a unique term and each column denotes a passage. Each cell contains the frequency of occurrence of one term from the passage.

The second step of LSA is applying a Singular Value Decomposition (SVD). SVD decomposes the rectangular matrix into the product of three matrices. One matrix is term vectors, another denotes a diagonal matrix and the last one denotes passage vectors. More formally, every rectangular matrix M can be decomposed into three matrices T, Σ and P^T, as shown in Equation 5.10.

$$M = T\Sigma P^T \tag{5.10}$$

where, T is a term vectors matrix, P^T is a matrix of passage vectors and Σ is a diagonal matrix of decreasing singular values.

The singular values represent the semantic space for terms and passages in a corpus of text. When the matrix Σ contains all the singular values of M, then the original matrix M is reconstructed by multiplying the three matrices T, Σ, and P^T.

The dimensionality of the space of semantic representations can be reduced by deleting some of the singular values, starting with the smallest. The matrix M_k, which is the k dimensional approximation to M, can be built by selecting the k largest singular values. In our case, we set the dimensionality parameter k to 2. The reconstruction of matrix M_k is given in Equation 5.11.

$$M_k = T\Sigma_k P^T \tag{5.11}$$

Similarly, the representations of terms and passages by multiplying their corresponding matrix decompositions are obtained. The representations of terms and passages are given in Equation 5.12.

$$T_k = T\Sigma_k \qquad P_k^T = \Sigma_k P^T \tag{5.12}$$

Finally, to calculate the similarity between two terms, we used the cosine measure, where the dot product between two vectors of matrix M_k shows the extent to which two terms are similar in the vector space. Cosine similarity measure is given in Equation 5.13.

$$Similarity_{LSA}(t_i, t_j) = \frac{t_i \cdot t_i}{\parallel t_i \parallel \parallel t_i \parallel} \tag{5.13}$$

where, t_i and t_j are terms, and $\parallel t_i \parallel$ and $\parallel t_j \parallel$ are the corresponding latent term space vectors.

5.4.3 Measures of the Effectiveness for the Objective Methods

We employed the standard information retrieval measures such as Precision, Recall and F1 [133] to evaluate the effectiveness of objective methods. The objective methods are evaluated against the subjective ones. The evaluation is conducted by taking the 10 top subjective terms as the ground truth and the top-N terms obtained by the objective methods as a relevance list.

The definition of precision and recall is adjusted in order to evaluate top-N terms obtained by objective methods. The definitions adopted are as following.

Precision is the ratio of total number of terms which occur simultaneously in the relevance list and in the ground truth list, to the number of terms in the relevance list. Precision is given in Equation 5.14.

$$Precision = \frac{\mid Relevance \cap GroundTruth \mid}{\mid Relevance \mid} * 100 \tag{5.14}$$

Recall is the ratio of total number of terms which occur simultaneously in the relevance list and in the ground truth list, to the number of terms in the ground truth list. Recall is given in Equation 5.15.

$$Recall = \frac{\mid Relevance \cap GroundTruth \mid}{\mid GroundTruth \mid} * 100 \tag{5.15}$$

Precision and recall are often inversely related to each other, such that if the number of relevant terms increases, then the value of recall increases, while at the same time precision decreases. Thus, we used the standard F1 measure, which is defined as the average of precision and recall and it is given in Equation 5.16.

$$F1 = \frac{2 * (Precision * Recall}{Precision + Recall} * 100 \tag{5.16}$$

5.5 Results and Analysis

The performance of objective methods is evaluated on two criteria. First being how well the objective methods score the top subjective terms. In order to do this, scores for the 10 top terms are taken as the ground truth. The score obtained for these terms using the objective methods are then evaluated. An example for the enrichment of the *Application* concept is observed and the comparison is shown in Table 5.6. The final score is in the range of [0,1], where 0 denotes a term with no relatedness and 1 denotes a highly related term for enriching the *Application* concept.

Table 5.6: The overall objective score for the top 10 terms selected by subjects

No	Subjective terms	tf*idf	χ^2	LSA	SEMCON
1	Apps	1.000	1.000	1.000	1.000
2	Application	1.000	1.000	1.000	1.000
3	Software	0.975	0.500	0.981	0.943
4	Program	0.914	0.500	0.894	0.923
5	Windows	0.000	0.028	0.000	0.409
6	Task	1.000	1.000	1.000	0.900
7	Browser	0.000	0.028	0.000	0.479
8	Function	0.569	0.222	0.577	0.701
9	User	0.603	0.417	0.707	0.544
10	Process	0.000	0.067	0.000	0.412

The comparison summarised in Table 5.6, shows that SEMCON generally outperforms the *tf*idf*, χ^2 and LSA. The red highlighted values show cases when one method performs better than the other. It can be seen from the red highlighted values that the SEMCON model gives much better results for the terms *Windows*, *Browser* and *Process* in contrast to the *tf*idf* and LSA which scores 0 to these three terms and χ^2 which scores close to 0. This is most likely because these terms did not occur in document/presentation slides that talk contextually about the *Application* concept but they occurred in the WordNet corpus.

SEMCON also scores higher for the terms *Program* and *Function*. The term *Task* gets a score of 1.0 by the *tf*idf*, χ^2 and LSA, which means that these three methods would rank the term *Task* as its first term to enrich the *Application* concept. The term *Task* however is ranked as the sixth relevant term to enriching the concept *Application* by subjects as shown in Table 5.5.

The second evaluation criteria is to check if the top terms scored by the objective methods are accurate. For this, we compute the precision, recall and F1 measure on the top-15 relevant terms list. Table 5.7 shows the resulting precision and recall of objective methods on retrieving and ranking of terms as the most relevant terms for enriching the *Application* concept in the computer domain. Terms correctly retrieved by the objective methods are highlighted in red in Table 5.7.

Table 5.7: Precision and recall of Application concept

Subjective terms	Objective terms			
	tf*idf	χ^2	LSA	SEMCON
Apps	Apps	Apps	Application	Apps
Application	Application	Application	Control	Application
Software	Control	Control	Apps	Software
Program	Master	Master	Master	Program
Windows	Part	Part	Part	Control
Task	Task	Task	Task	Task
Browser	Software	Program	Web	Part
Function	Program	Software	File	Master
User	Operation	Operation	Page	Operation
Process	Computer	User	Access	Function
	User	Computer	Asset	Computer
	Function	Function	Browser	System
	System	Component	Collection	User
	Component	System	Concept	Browser
	Access	Device	User	Use
Recall	**70.0**	**70.0**	**50.0**	**80.0**
Precision	**46.7**	**46.7**	**33.3**	**53.3**

In the following paragraph, we are giving an example to show how the precision and recall, shown in Table 5.7, are computed. Total number of terms obtained by intersection of ground truth list (column entitled subjective terms) and relevance list (column entitled *tf*idf*) is equal to 7. Number of terms in ground truth list is 10, while number of terms in relevance list is 15. Recall is computed as 7/10*100=70.0% and precision as 7/15*100=46.7%. The example illustrated shows computation of precision and recall for *tf*idf* method but in a similar fashion they are also computed for χ^2, LSA, and SEMCON.

Additionally, Table 5.8 and Table 5.9 shows precision, recall and F1 results obtained by the SEMCON on retrieving and ranking of terms as the most relevant terms for enriching concepts of computer domain and other domains, respectively.

The performance of SEMCON in terms of F1 measure is compared with the performance of *tf*idf*, χ^2 and LSA. The comparison is performed using results of various domains and it shows that SEMCON achieved better results on finding the highly related terms to enrich ontology concepts.

Table 5.10 shows F1 results for computer domain. The results depict that SEMCON achieved the average improvement of 12.2% over the *tf*idf*, 21.8% over the χ^2, and 24.5% over the LSA.

The same comparisons for F1 measure is also conducted for other domains. These results are shown in Tables 5.11 - 5.14.

Table 5.8: The performance of SEMCON on computer domain

Domain	P (%)	R (%)	F1 (%)
Computer	26.7	40.0	32.0
Software	46.7	70.0	56.0
Hardware	33.3	50.0	40.0
Web	46.7	70.0	56.0
Storage	46.7	70.0	56.0
Microprocessor	40.0	60.0	48.0
InputAndOutputDevices	33.3	50.0	40.0
Application	53.3	80.0	64.0
Windows	46.7	70.0	56.0
Average	**41.5**	**62.2**	**49.8**

Table 5.9: The performance of SEMCON on different domains

Domain	P (%)	R (%)	F1 (%)
Computer	41.5	62.2	49.8
Database	34.2	51.3	41.0
Internet	38.1	57.1	45.7
C++_Programming	37.3	56.0	44.8
Software_Engineering	49.5	74.3	59.4

Table 5.10: The F1 of objective methods performed on computer domain

Concept	tf*idf (%)	χ^2 (%)	LSA (%)	SEMCON(%)
Computer	24.0	24.0	32.0	32.0
Software	56.0	48.0	40.0	56.0
Hardware	32.0	40.0	32.0	40.0
Web	32.0	32.0	40.0	56.0
Storage	64.0	56.0	64.0	56.0
Microprocessor	48.0	40.0	56.0	48.0
InputAndOutputDevices	32.0	24.0	8.0	40.0
Application	56.0	56.0	40.0	64.0
Windows	56.0	48.0	48.0	56.0
Average	**44.4**	**40.9**	**40.0**	**49.8**

Table 5.11: The F1 of objective methods performed on SE domain

Concept	tf*idf (%)	χ^2 (%)	LSA (%)	SEMCON(%)
Software	56.0	56.0	40.0	48.0
Cost	40.0	40.0	40.0	48.0
Product	64.0	56.0	48.0	48.0
Attribute	32.0	48.0	32.0	56.0
Process	72.0	48.0	32.0	72.0
Generic	48.0	64.0	64.0	72.0
Hybrid	64.0	56.0	56.0	72.0
Average	**53.7**	**52.6**	**44.6**	**59.4**

Finally, we evaluated the performance of SEMCON and the three other objective methods by comparing the average results of each domain. The obtained results (precision, recall and F1) illustrated in Figure 5.5 - 5.7 show that SEMCON gives better results than the other three methods for all the domains excepts for the internet domain. This may have

Table 5.12: The performance of SEMCON on C++ programming domain

Concept	tf*idf (%)	χ^2 (%)	LSA (%)	SEMCON(%)
C++_Programming	24.0	40.0	40.0	40.0
Syntax	56.0	48.0	48.0	48.0
Technique	24.0	16.0	24.0	24.0
Structure	40.0	40.0	32.0	40.0
Expression	48.0	40.0	40.0	48.0
Operator	24.0	24.0	56.0	24.0
Encapsulation	48.0	64.0	64.0	48.0
Inheritance	64.0	56.0	56.0	56.0
Polymorphism	48.0	48.0	40.0	56.0
Platform	56.0	56.0	48.0	64.0
Average	**43.2**	**43.2**	**44.8**	**44.8**

Table 5.13: The F1 of objective methods performed on database domain

Concept	tf*idf (%)	χ^2 (%)	LSA (%)	SEMCON(%)
Database	24.0	16.0	24.0	16.0
Model	48.0	40.0	16.0	48.0
E-R	48.0	48.0	16.0	48.0
User	40.0	16.0	16.0	16.0
SQL	32.0	40.0	16.0	32.0
DDL	64.0	48.0	40.0	64.0
DML	40.0	24.0	32.0	48.0
Administrator	24.0	24.0	16.0	24.0
Average	**40.0**	**32.0**	**22.0**	**41.0**

Table 5.14: The F1 of objective methods performed on internet domain

Concept	tf*idf (%)	χ^2 (%)	LSA (%)	SEMCON(%)
Internet	40.0	24.0	48.0	40.0
Application	40.0	40.0	40.0	32.0
Web	32.0	32.0	40.0	32.0
Access	56.0	48.0	40.0	48.0
Browser	64.0	48.0	48.0	64.0
ISP	72.0	48.0	56.0	64.0
HTML	40.0	48.0	56.0	40.0
Average	**49.1**	**41.4**	**46.9**	**45.7**

happened due to the fact that subjects are making their selections based on the descriptions provided under each concept in the questionnaire, when they were asked to select 5 closely related terms. In other words, subjects might have used contextual information from the description provided in the questionnaire about each concept rather than their existing prior knowledge. As the ground truth list is composed of terms which carry contextual meaning in a document to describe a particular concept, therefore this might have served better for *tf*idf* for Internet domain where people choose terms based on the context rather than prior domain knowledge. Nevertheless, there is a significant improvement of results for other domains by SEMCON over other methods.

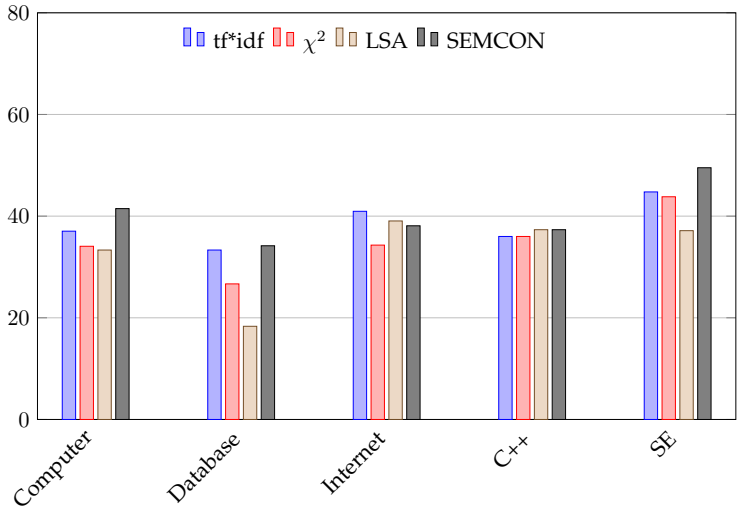


Figure 5.5: Precision for 5 different domains

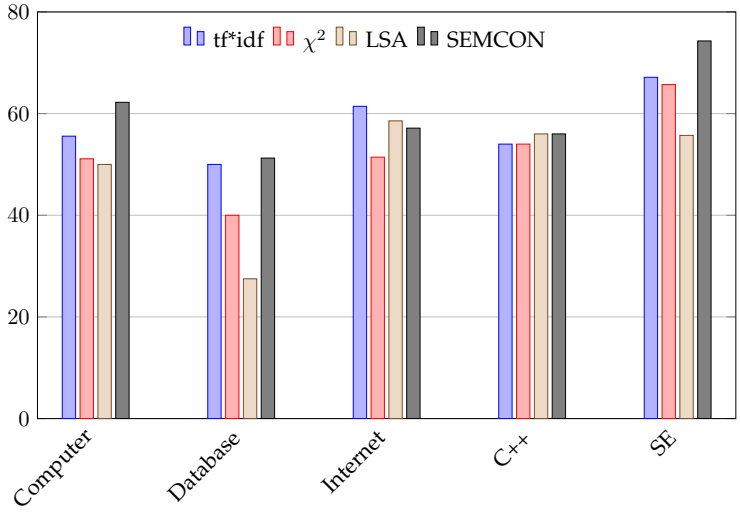


Figure 5.6: Recall for 5 different domains

5.5.1 The Impact of Statistical Features

The SEMCON takes into account the context of a term by computing an observation matrix, which exploits the statistical features such as term font type and term font size besides the frequency of the occurrence of a term. An example of observation matrix, with or without using the statistical features, is shown in Table 5.15.

Table 5.15 depicts two observation matrix scores computed for 5 different terms: *com-*

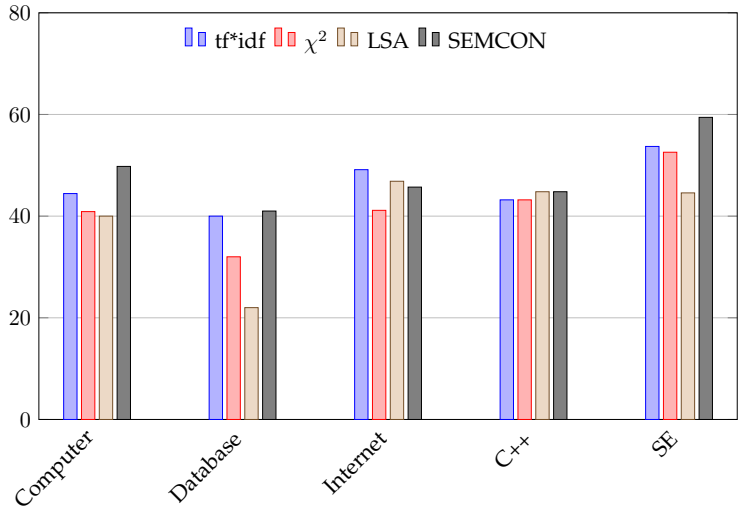


Figure 5.7: F1 for 5 different domains

Table 5.15: An example of observation matrix with/without using statistical features

Slide	Computer		Data		Device		System		Web	
	No	With	No	With	No	With	No	With	No	With
1	3	6.25	3	5.25	1	1.75	0	0	0	0
2	2	3	0	0	0	0	1	1.5	4	8
3	5	9.25	0	0	4	7	3	5.5	0	0
4	3	5.5	2	3.5	5	8	0	0	0	0
5	3	5	1	1.5	1	1.5	1	2	0	0
6	7	12.25	0	0	0	0	3	6	0	0
7	1	2.25	0	0	0	0	3	6.25	0	0

puter, *data*, *device*, *system*, and *web*. For each term, the first column shows the score obtained using only term's frequency (denoted with No), and the second column shows the score obtained using statistical features.

It is evident from Table 5.15 that statistical features do contribute to observation matrix score but there is a need to investigate into how much each of the statistical features i.e. the font size and font type, contribute to the overall performance of SEMCON. The contribution presented for the computer domain dataset is shown in Table 5.16. Furthermore, Table 5.16 gives a comparison of Precision, Recall and F1 measures of SEMCON, when the observation matrix is built using the statistical features and when the observation matrix is built only using the frequency of the occurrence of a term. The average F1 measure is improved by 3.75% when the observation matrix is built using statistical features. The F1 measures of *Web* and *InputAndOutputDevices* concepts are improved by 16.7% and 25.0%, respectively. This happened due to the fact that these terms occurred very often as level 1 font size and bold font type in the passages, hence statistical features have a high contribution in the value of the overall score.

Table 5.16: The performance of SEMCON with/without statistical features

Concept	Precision (%)		Recall (%)		F1(%)	
	No	With	No	With	No	With
Computer	26.7	26.7	40.0	40.0	32.0	32.0
Software	46.7	46.7	70.0	70.0	56.0	56.0
Hardware	33.3	33.3	50.0	50.0	40.0	40.0
Web	40.0	46.7	60.0	70.0	48.0	56.0
Storage	46.7	46.7	70.0	70.0	56.0	56.0
Microprocessor	40.0	40.0	60.0	60.0	48.0	48.0
InputAndOutputDevices	26.7	33.3	40.0	50.0	32.0	40.0
Application	53.3	53.3	80.0	80.0	64.0	64.0
Windows	46.7	46.7	70.0	70.0	56.0	56.0
Average	**40.0**	**41.5**	**60.0**	**62.2**	**48.0**	**49.8**

5.5.2 The Effect of Weight Parameter w

This section investigates into how much each of contextual and semantic components contributes to the overall score. This is achieved by tuning the weight parameter w given in Equation 5.6. We conducted the experiments with various w settings from 0.0 to 1.0 with a step size of 0.1. When the w is set to 0.0 the overall score is computed using only the semantic component, while w=1.0 indicates that the contribution is only from the contextual component. The rest of the values shows that the overall score is composed of both the contextual and semantic information. Figure 5.8 illustrates the precision with respect to the weight parameter w, obtained by experiments carried out on computer domain data set. It can be seen from the chart diagram that the best result in terms of precision is obtained when the value of weighting parameter w is set to 0.5. The precision starts declining with an increase or a decrease in the value of w. This suggests that both semantic and contextual information should contribute equally to computing the overall score as described in subsection 5.3.4.

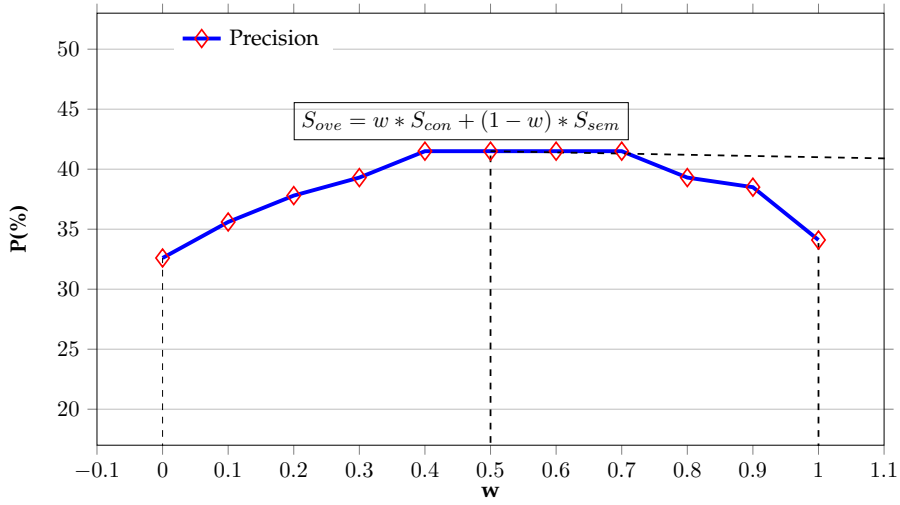


Figure 5.8: Precision as a function of weight parameter w

5.6 The Applications of SEMCON

SEMCON can be used in many application areas including but not limited to information systems, eLearning platforms, open educational resources (OER), online social network (OSN) analysis, etc., for building dictionaries, classifying documents, enriching ontologies - among many others. For instance, it can be applied for document classification in information systems where each record can be grouped into different categories automatically utilizing context and semantics. The two areas where SEMCON has been applied are as follows:

1. The classification of multimedia documents in the web-based eLearning platforms.
2. The analysis of Online Social Networks (OSNs) for identifying criminal activity and possible suspects.

These applications are discussed briefly in this section.

5.6.1 SEMCON for Web-based eLearning Platforms

Today's eLearning platforms consist of multiple media modalities including presentation slides, lecture videos, transcript files, handouts, and additional documents, delivering thousands of learning objects on a daily basis. These media provides a rich source of information that can be utilized for organizing and structuring learning objects.

Structuring and organizing huge amount of learning objects is a labour intensive, prone to errors and a cumbersome task, however. SEMCON on the other hand can prove useful in automatically organizing pedagogical multimedia content using an automatic classification approach based on the ontology described in [67]. Any new unlabeled learning object can be assigned to a predefined category in an eLearning platform using SEMCON. This is plausible by calculating the similarity between the extracted terms from the learning object and the ontology concepts. The learning object can then be assigned to a category having the highest similarity value with respect to that learning object.

An ontology represents semantic aspects of the learning objects through entities defined within a domain ontology. Therefore, each learning object that uses the ontology is represented as a vector, whose elements indicate the importance of concepts in the ontology.

5.6.2 SEMCON for OSN Analysis for Criminal Activity Detection

Analysing users' behaviour in Online Social Networks (OSNs) for investigating criminal activities is an area of great interest these days. The criminal activity analysis provides a useful source of information for law enforcement and intelligence agencies across the globe. Existing methods monitoring criminal activity normally rely on contextual analysis by computing co-occurrences of terms, which is not much effective.

SEMCON on the other hand can provide useful semantic as well as contextual information in identifying criminal activities by analysing users' posts and data, and by maintaining a history of recent user activities in the digital platforms. The proposed model [72] uses web crawlers suited to retrieve users' data such as posts, feeds, comments from Facebook, and exploits them semantically and contextually using the ontology enhancement objective metric SEMCON. The output of the model is a probability value of a user being a suspect which is computed by finding the similarity between the terms obtained from SEMCON and the concepts of criminal ontology.

5.7 Conclusion and Future Work

In this paper, we proposed a new generic approach to enriching the domain ontologies with new concepts by combining contextual and semantic information of terms extracted

from the domain documents. SEMCON employs a hybrid ontology learning approach to identify and extract new concepts. This approach involves functionalities from both linguistic and statistical ontology learning approaches. From the former approach, SEMCON utilizes morpho-syntactic analysis to identify and extract noun terms, as a part of speech, which represent the most meaningful terms in a document. While from the latter approach, SEMCON derives the context using cosine similarity between term vectors whose members are frequencies of terms. SEMCON uses, besides term frequency, two new statistical features, i.e. term font size and font type to determine the context of a term. In addition to contextual information, SEMCON also incorporates the semantic information of terms using the lexical database WordNet and finally aggregates both contextual and semantic information of this particular term.

Several experiments on various small data sets are conducted, where results obtained by SEMCON are compared with results obtained by other objective methods such as *tf*idf*, χ^2 and LSA. Comparison showed that SEMCOM outperforms the three objective methods by 12.2% over the *tf*idf*, 21.8% over the χ^2 and 24.5% over the LSA. We also carried out experiments about the effect of statistical features on the overall performance of the proposed metric and our findings showed an improved performance. Additionally, we investigated into the amount of contribution made by each of the contextual and semantic components to the overall task of concepts enrichment. The obtained results indicated that a balanced weight between the contextual and semantic components gives the best performance.

The future work may further exploit other features for computing observation matrix and nonlinear models, i.e. exponential, for computing the statistical features. Another direction may be extracting candidate terms from multimedia documents including audio and video.

A3:
Analysis of Online Social Networks Posts to Investigate Suspects Using SEMCON

Publication details
Kastrati, Z., Imran, A., and Yayilgan, S., and Dalipi, F., "*Analysis of Online Social Networks Posts to Investigate Suspects Using SEMCON*", in the 17^{th} International Conference on Human-Computer Interaction (HCI'15) (2015), Springer, pp. 148-157.

Chapter 6

Analysis of Online Social Networks Posts to Investigate Suspects Using SEMCON

Abstract

Analysing users' behaviour and social activity for investigating suspects is an area of great interest nowadays, particularly investigating the activities of users on Online Social Networks (OSNs) for crimes. The criminal activity analysis provides a useful source of information for law enforcement and intelligence agencies across the globe. Current approaches dealing with the social criminal activity analysis mainly rely on the contextual analysis of data using only co-occurrence of terms appearing in a document to find the relationship between criminal activities in a network. In this paper, we propose a model for automated social network analysis in order to assist law enforcement and intelligence agencies to predict whether a user is a possible suspect or not. The model uses web crawlers suited to retrieve users' data such as *posts*, *feeds*, *comments*, etc., and exploits them semantically and contextually using an ontology enhancement objective metric SEMCON. The output of the model is a probability value of a user being a suspect which is computed by finding the similarity between the terms obtained from the SEMCON and the concepts of criminal ontology. An experiment on analysing the public information of 20 Facebook users is conducted to evaluate the proposed model.

6.1 Introduction

In recent years, the usage of Online Social Networks (OSNs) has increased rapidly throughout all layers of society. Law enforcement and intelligence agencies analyse traces of digital evidence in order to solve crimes and capture criminals whom are also OSNs users during their investigation activities. Particularly analysing contents shared by users on social networks such as Facebook, Twitter and LinkedIn are of interest. Several approaches of analysis aiming at extracting useful information, modelling users profile, and understanding users behaviour and social activity have been proposed [3].

Analysing users behaviour and social activity for investigating suspects is also an interesting area of research, particularly investigating the activities of users on OSN for crimes. The criminal activity analysis provides a useful source of information for law enforcement and intelligence agencies across the globe. Some agencies are now using social media as a crime-solving tool [77]. Digital traces from social media such as Facebook is gaining fast acceptance for use as evidence in courts [51]. According to a survey conducted by LexisNexis [83] in 2012, there are more than 950 law enforcement professionals with federal, state, and local agencies in United States whom use social media, particularly Facebook and YouTube, to obtain evidence to deepen their criminal investigation. Other similar criminal cases have been reported recently where digital evidence from OSNs is used as support for digital investigation [12, 98].

Criminal activity analysis consists of different stages such as data processing, transformation, analysis, and visualization. Many of these stages are done manually. Thus, it takes much time and human effort to extract the required evidence from the massive amount of information.

Recently some research has been done to automate the social criminal activity analysis to help law enforcement and intelligence agencies discover the criminal networks. In this light, a framework for the forensic analysis of user interaction in OSNs is proposed in [2]. The framework enables searching for actor activities and filtering them further for temporal and geographical analysis. The authors in [157] proposed a framework that consists of major components of a network analysis process: network creation, network partition, structural analysis, and network visualization. Based on this framework, the authors developed a system called *CrimeNet Explorer*. The system has structural analysis functionality to detect subgroups from a network, identifying central members of subgroups, and extracting interaction patterns between subgroups. The authors in [43, 136] used data mining approach for analyzing criminal groups. They used data mining in multiple social networks data to discover criminal networks.

However current approaches to automating the social criminal activity analysis have some limitations. They mainly rely on the contextual analysis using only co-occurrence of terms appearing in a document to find the relationship between criminal activities in a network [157]. Moreover, some of the approaches perform experiment using no real-world datasets [43].

In this paper, we try to fill this gap by proposing a framework for automated social network analysis. This framework will assists law enforcement and intelligence agencies to predict efficiently and effectively whether a user is a possible suspect or not. This is achieved by exploiting users' *posts*, *feeds* and *comments*, semantically and contextually using SEMCON [69]. SEMCON is a context and semantic based ontology enhancement model developed at our lab originally for the purpose of enriching an ontology from posts of multimedia documents.

The rest of the paper is organized as follows. In Section 6.2 we illustrate in detail our proposed model. Section 6.3 describes the setting for experimental procedures whereas Section 6.4 illustrates the experimental results and their analysis. Lastly, in Section 6.5 we sketch conclusions and future work.

6.2 Proposed Model and Methodology

The proposed model, illustrated in Figure 6.1, aims at performing the analysis of social networks profiles. Information such as *posts*, *feeds* and *comments* are extracted and analysed, considering in particular both the context and the semantics of terms used by users. The model is explained in the following sections.

6.2.1 Acquisition Module

The module use web crawlers suited to retrieve and manage data coming from particular social networks such as Facebook, Twitter, LinkedIn, etc. In our case we have used Facebook crawler for managing Facebook posts. Facebook crawler is based on the Facebook Graph APIs and Facebook Query Language (FQL). To fetch Facebook messages and making queries, this paper uses RestFB [124] which is a simple and flexible Facebook Graph API client written in Java. The crawler uses an opaque string called Facebook access token that identifies a user, application, or page and can be used by the application to make graph API calls.

In this work, the Facebook crawler is dedicated to fetch only *posts*, *feeds*, and *comments* of a user. Facebook imposes some limitations on the number of *posts*, *feeds*, and *comments* retrievable through its APIs according to the data access policy of Facebook. It does not allow to retrieve the information of more than 25 *posts*, *feeds* and *comments* per user. The restrictions on the maximum number of retrievable information are overcome by using specific parameters which enable to filter and page through the connection data.

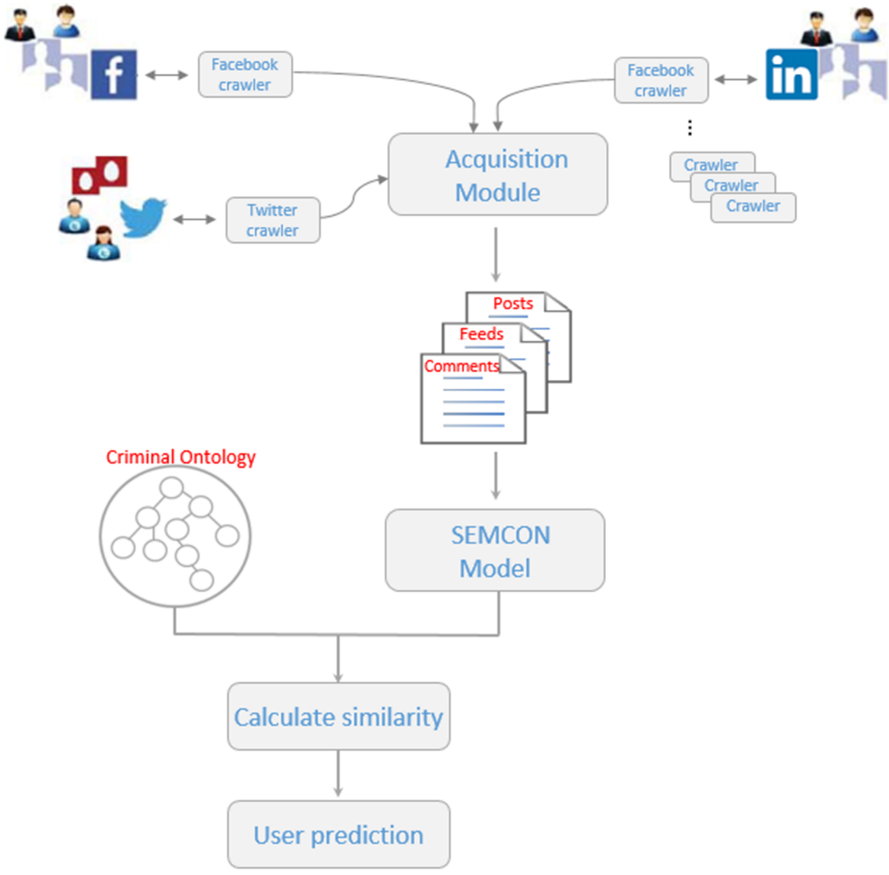


Figure 6.1: Flow chart of the proposed model

6.2.2 SEMCON Module

The information fetched by Acquisition Module is used as input to the SEMCON Module. The SEMCON module treats each *post*, *feed* and *comment* basically as an independent document-passage and it performs the following steps.

Initially a morpho-syntatic analysis using TreeTagger [132] is performed where the partitioned passages are tokenized and lemmatized. The potential terms that are obtained as a result can either be a noun, verb, adverb or adjectives. These are different parts-of-speech (POS) of a language. It is a well-known fact that nouns represent the most meaningful terms in a document [84], thus, our focus is on extracting only common noun terms t for further consideration.

The next step is the calculation of the observation matrix. The observation matrix is formed by calculating the frequency of occurrences of each term t, its font type (*bold*, *underline*, *italic*) and its font size (*title*, *level* 1, *level* 2) as given in Equation 6.1.

$$O_{i,j} = \sum_{i \in t} \sum_{j \in p} (Freq_{i,j} + Type_{i,j} + Size_{i,j}) \tag{6.1}$$

where, t and p indicate the set of terms and passages, respectively. $Freq_{i,j}$ denotes the

frequency of occurrences of term t_i in passage, p_j, $Type_{i,j}$ denotes font type of term t_i in passage p_j, and $Size_{i,j}$ indicates font size of term t_i in passage p_j.

The observation matrix is used as input to compute the contextual and semantic similarity between two terms.

Term to term contextual score (S_{con}) is calculated using the cosine similarity metric with respect to the passages, and it is given in Equation 6.2.

$$S_{con}(t_i, t_j) = \frac{t_i \cdot t_j}{\| t_i \| \| t_j \|} \tag{6.2}$$

A term square matrix is used to store the contextual(S_{con}) values among all extracted terms *t*.

The next step is the computation of the semantic score (S_{sem}). The semantic score is calculated using the Wu&Palmer algorithm [156] and the score is computed using the Equation 6.3.

$$S_{sem}(t_i, t_j) = \frac{2 * depth(lcs)}{depth(t_i) + depth(t_j)} \tag{6.3}$$

where t_i and t_j indicate terms extracted from the passage, *depth(lcs)* indicates least common subsumer of t_i and t_j, *depth*(t_i) and *depth*(t_j) indicate the path's depth of t_i and t_j, respectively.

Go through the all terms, we take all possible pairs and compute the semantic score $S_{sem}(t_i, t_j)$, for each pair t_i and t_j, where t_i, $t_j \in C$ and *C* is the set of terms extracted from the corpus.

The overall correlation between two terms t_i and t_j extracted from the the passage is computed using the contextual and semantic score. Mathematically, the overall score is given in Equation 6.4.

$$S_{overall}(t_i, t_j) = w * S_{con}(t_i, t_j) + (1 - w) * S_{sem}(t_i, t_j) \tag{6.4}$$

where S_{con} is the contextual score, S_{sem} is the semantic score and w is a parameter with value set as 0.5 in our case, based on the empirical analysis from the data set.

The overall score is in the range (0,1]. The overall score is 1 if two extracted terms are the same.

6.2.3 User Prediction Module

The prediction of a user as a suspect or not depends on the similarity score between the terms extracted from the user' *posts*, *feeds* and *comments* via SEMCON module and concepts extracted by the criminal ontology. The higher the score, the closer the user is considered as a suspect user.

The similar calculation is performed using the cosine similarity measurement. More formally, it is given in Equation 6.5.

$$Similarity(O_c, u_i) = \frac{\overrightarrow{O}_c \times \overrightarrow{u}_i}{\| \overrightarrow{O}_c \| \cdot \| \overrightarrow{u}_i \|} \tag{6.5}$$

where, O_c indicates concepts extracted from the criminal ontology and u_i indicates terms extracted by the user postings.

The output of the system is a probability value *P*, of a user being a suspect *s*. If the P_s is greater than a specified threshold *t* then the user is labelled as a suspect.

6.3 Experimental Setting

We have performed the investigation of suspects using the public users' *posts*, *feeds* and *comments*.

The facebook crawler is established to collect the data for the period starting from 1 January till 31 December 2014. The posts are extracted from news and media.

The posts from social networks contain usually noisy text, e.g. null values, therefore we filtered out only the posts which comply with the standard rules of orthography, syntax and semantics. After this process, we created a corpus which consists of 198 posts published by 20 users. The average number of posts per user is 10. The total number of terms used is 8493 with an average of 43 terms for each post. Finally, from these terms we identified and extracted 1042 nouns (singular and plural). The detailed information for each user is shown in Table 6.1.

Table 6.1: The corpus data

User	# of Posts	# of Terms	# of Nouns
1	11	929	55
2	3	121	11
3	12	301	58
4	5	130	25
5	8	376	56
6	4	1550	140
7	11	366	48
8	9	383	51
9	16	600	58
10	11	270	40
11	12	313	46
12	6	117	21
13	21	567	102
14	9	344	46
15	12	336	43
16	8	317	42
17	12	494	58
18	9	307	44
19	10	298	42
20	8	374	56
Total	198	8493	1042

We have also created a criminal ontology shown in Figure 6.2. Basically it is used to predict if a user is a suspect by comparing its concepts with the terms outputted by the SEMCON as described in Section 6.2.3. However, the criminal ontology may also be used for visualization of criminal information by displaying concise overviews of its concepts and their hierarchical relations using treemaps.

6.4 Results and Analysis

In order to evaluate a user being as a suspect or not, we have performed an experiment on 20 Facebook users by analysing their public postings. For each user, we initially computed an overall score by aggregating the semantic and contextual score for each term (noun) extracted. The overall scores of terms are used to find the similarities of the terms with the criminal ontology concepts. Figure 6.3 illustrates the terms score obtained by SEMCON for the *User* #13 as depicted in Table 6.1.

As can be seen from the graph, the contextual score indicated by the blue curve is much lower then the semantic score denoted by the orange curve. This may have happened due to the fact that the user has posted or commented in different topics. However, terms used in these posts have high semantic correlation with each other.

Figure 6.2: A part of criminal ontology

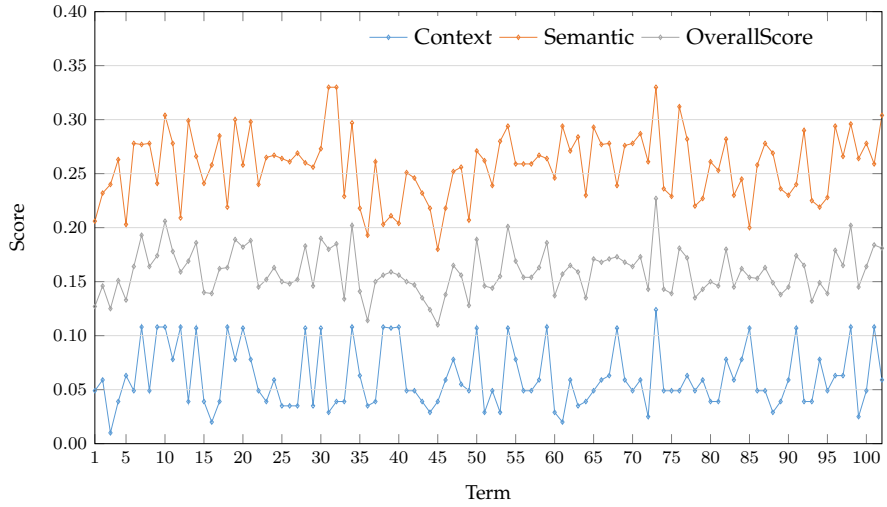


Figure 6.3: Scores obtained by SEMCON for a user in investigation

In the next step we found out how likely that a user is a suspect. This is achieved by comparing the user overall score obtained by SEMCON module and the scores of concepts of criminal ontology. User suspicion is represented by a probability value. The obtained probabilities for users being a suspect are shown in Table 6.2. The probability value 0.000 represents the users whose posts does not contain any of the criminal ontology concepts. Thus, these users are considered as unsuspected users. The probability value 1.000 indi-

cates the users whose posts contain some of the concepts of the criminal ontology, i.e. *gun*, *rifle*, *shooting*, *threat* and *death*. These users are considered to be highly suspected users.

Table 6.2: The probability of users being suspects

User	Probability	User	Probability
1	0.990	11	0.992
2	0.727	12	0.000
3	1.000	13	1.000
4	0.892	14	0.772
5	0.578	15	0.784
6	0.820	16	0.800
7	1.000	17	0.799
8	0.000	18	0.000
9	0.933	19	1.000
10	1.000	20	0.800

Based on the obtained probability results, we can identify three major categories of users; users classified as unsuspected users, moderate suspected users and the highly suspected users. More precisely, if the probability score of user exceeds a given positive threshold value (in our case 0.90) we classify his/her as being a highly suspected user; if his/her probability score is 0.00 we label him/her as unsuspected user, otherwise he/she is considered as being a moderate suspected user. The labelling of users in particular categories is shown in Table 6.3.

Table 6.3: Categorization of user prediction

	Unsuspected	Moderate suspected	Highly suspected
User	8, 12, 18	2, 4, 5, 6, 14 15, 16, 17, 20	1, 3, 7, 9, 10 11, 13, 19

6.5 Conclusion and Future Work

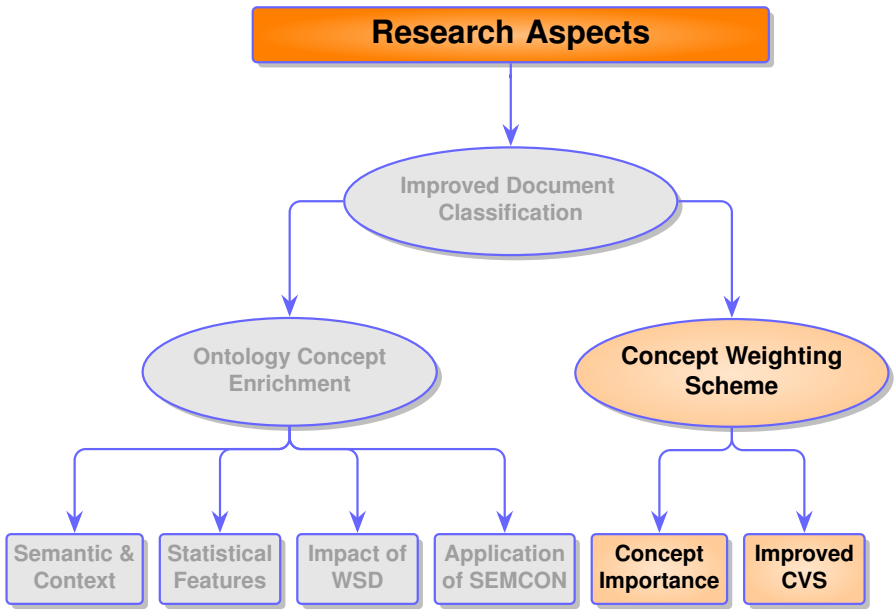
In this paper, we have proposed a new approach to investigating if a user is a suspect by analysing the OSNs data. We used Facebook as a case study of OSNs and Facebook user' *posts*, *feeds* and *comments* have been the object of the study. The approach employs the SEMCON to provide a semantic and contextual data-mining analysis for automatically monitoring users' activity through textual analysis. We initially built a domain ontology called criminal ontology. The prediction of a user as a suspect or not is computed by finding the similarity score between terms extracted from user' *posts*, *feeds*, and *comments* via SEMCON module and the concepts extracted by the criminal ontology.

From the experiment conducted by analysing the postings published within a year by 20 users, we identified three categories of users: unsuspected, moderate suspected and highly suspected users. The categorization of users can assist law enforcement and intelligence agencies to narrow the investigation, identify and focus only on suspected users in order to prevent or solve crimes.

In the future we plan to further extend our proposed approach. Terms obtained from the SEMCON model will be used to build a user ontology. The user ontology can be used to create a history based user activity profile which may actually put light on otherwise invisible relations between a particular social network user and his/her network, and the dependence among a user's various activities. It also can be used by the law enforcement personnels for deeper investigation in order to search for suspicious user activities and filtering them for temporal and geographical analysis.

Part III

Concept Weighting Scheme



This part addresses the second research aspect of this thesis 'Concept weighting scheme' and it aimed to answer the fourth research question listed in Section 1.2.

This part is composed of two chapters constituted by two published research articles. Chapter 7 describes a new Markov-based model to automatically estimate importance of concepts that reflects how important a concept is in an ontology.

Chapter 8 expanded the previous work to improve concept vectors with a new concept weighting scheme composed of two components - concept importance and concept relevance. A thorough discussion and analysis along with an extensive evaluation of the proposed concept weighting approach is also presented in this chapter.

A4: Adaptive Concept Vector Space Representation using Markov Chain Model

Publication details
Kastrati, Z., and Imran, A., "*Adaptive Concept Vector Space Representation Using Markov Chain Model*", in the 19^{th} International Conference on Knowledge Engineering and Knowledge Management (EKAW'14) (2014), Springer, pp. 203-208.

Chapter 7

Adaptive Concept Vector Space Representation Using Markov Chain Model

Abstract

This paper proposes an adaptive document representation (concept vector space model) using Markov Chain model. The vector space representation is one of the most common models for representing documents in classification process. The document classification based on ontology classification approach is represented as a vector, whose components are ontology concepts and their relevance. The relevance is represented by the frequency of concepts' occurrences. These concepts make various contributions in classification process. The contributions depend on the position of concepts where they are depicted in the ontology hierarchy. The hierarchy such as classes, subclasses and instances may have different values to represent the concepts' importance. The weights to define concepts importance are generally selected by empirical analysis and are usually kept fixed. Thus, making it less effective and time consuming. We therefore propose a new model to automatically estimate weights of concepts within the ontology. This model initially maps the ontology to a Markov chain model and then calculates the transition probability matrix for this Markov chain. Further, the transition probability matrix is used to compute the probability of steady states based on left eigenvectors. Finally, the importance is calculated for each ontology concept. And, an enhanced concept vector space representation is created with concepts importance and concepts relevance. The concept vector space representation can be adapted for new ontology concepts.

7.1 Introduction

Today, the web is the main source of information which is consistently increasing. The information is usually kept in unstructured and semi-structured format. More than 80 % of the information of an organization is stored in an unstructured format (reports, email, views, news, etc.), and the rest is stored in structured format [121]. Therefore, discovering and extracting useful information from these resources is difficult without organization and summarization of text documents, and this is an extremely vital and tedious process in today's digital world [4]. An automatic classification in this regard plays a key role in organizing these massive sources of unstructured text information into an organized format.

Automatic text document classification is a process of automatically assigning a text document in a given domain to one or more class labels from a finite set of predefined categories. The first step in classification process is the preprocessing of the text documents and storing the information in a data structure, which is more appropriate for further processing. This can be achieved by a vector space model. The vector space model is one of the most common models for representing text documents and it is widely used in text document classification [75].

There are two major approaches for a text document representation into vector space model - machine learning approach and ontology based approach. The machine learning approach, which is based on the idea that a text document can be represented by a set

of words known as bag-of-words representation, and the ontology approach which follows the idea that text document can be represented by a set of concepts known as bag-of-concepts representation. Ontology based approach represents semantic aspects of the text documents through entities defined within the domain ontology. A text document using the domain ontology is represented as a vector, whose components are concepts and their relevance. Concepts are extracted from a domain ontology and the relevance is calculated using frequency of concepts' occurrences in the corpus which makes this domain.

It is argued in [54, 120] that contribution of ontology concepts in classification process depends on the position of concepts where they are depicted in the hierarchy and this contribution is indicated by a weight. The hierarchy consists of classes, subclasses and instances that may have different weights to represent the concepts importance. These weights are usually calculated either manually or empirically through trial and error by conducting experiments. Researchers in [120] calculate weights of ontology concepts by performing experiments. They experimented many times to adjust the parameters which denote the importance of the concepts in ontology. After the experiment was conducted several times, they proposed to set the parameter value 0.2 when the concepts were classes, 0.5 when concepts were subclasses and 0.8 when concepts were instances in ontology. The approach implemented in [54, 42] proposed to use layers of ontology tree to indicate the abstract degree of concepts. Researchers used layers to represent the position of concepts in ontology and the weight of a concept is calculated by counting the length of path from the root node. The same approach of using layers for calculating concepts' weight values was used in [116]. They proposed the idea to consider only the leaf concepts of an ontology, in contrast of using all concepts, presuming that leaf concept are the most important elements in the ontology. The leaf concepts can be any subset of ontology that forms a set of mutually independent concepts. They assume that more general concepts, such as super-classes, are implicitly taken into account through the leaf concepts by distributing their importance to all of their sub-classes down to the leaf concepts in equal proportion. The drawback of these approaches is that they do not calculate the weights of concepts in ontology automatically. In fact, they tune it empirically through trial and error by conducting experiments thus keeping these weights fixed. We therefore address these issues in this paper by proposing a new approach for automatically calculating weights of concepts in an ontology and then using these weights to enhance the concept vector space representation model.

The rest of the paper is organized as follows. Section 7.2 describes our proposed method in detail while section 7.3 concludes the paper.

7.2 Proposed model and methodology

The following section describes the proposed model which is inspired from [48]. The proposed model consists of three subtasks; mapping the domain ontology into a Markov chain model, calculation of the transition probability matrix for Markov chain model and calculation of the importance for each concept in ontology. The final step is building a concept vector space model.

7.2.1 Modelling of Markov Model by a Domain Ontology

Following the formal definition of the domain ontology, we will adopt a model where the ontology will be presented as a directed acyclic graph in which classes and their instances are structured in a hierarchy. This definition will be represented by the tuple O= (C, H, I, *type(i)*, *rel(i)*) [48], where:

- C is a non-empty set class identifiers;
- H is a set of taxonomy relationship of C;
- I is a potentially empty set I of instance identifiers;

- *type(i)* is an instance to class relation;
- *rel(i)* is an instance to instance relation.

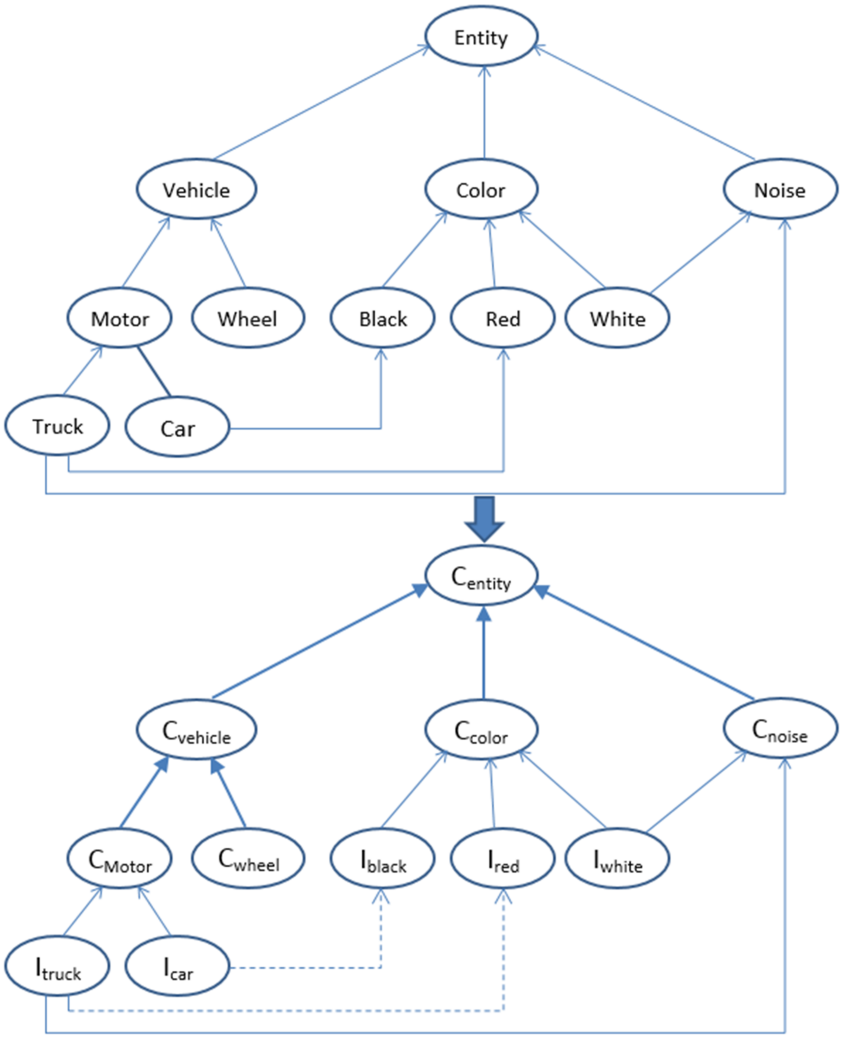


Figure 7.1: Mapping of entity ontology to Markov chain model

The graphical representation of the domain ontology will be implemented using the Markov chain model. The Markov chain model is adopted because of its ability to deal with flexible relationships, such as inter-instance relationships and non-hierarchical relationships between classes.

To be consistent with the ontology definition we partitioned the set of Markov chain states into two disjoint subsets; S_C, which contains the states corresponding to ontology classes, and S_I, which contains the states corresponding to ontology instances.

The Markov chain modelling is an equivalent mapping which means that classes (*entity*, *vehicle*, *motor*) in the ontology and the instances of those classes (*truck*, *car*, *red*) are mapped to states (C_{entity}, $C_{vehicle}$, C_{motor}, I_{truck}, I_{car}, I_{red}) in the Markov chain. Whereas, all instance-to-instance relations, instance-to-class relations and non-hierarchical relations between instances, and classes are mapped to state transitions. As can be seen from the Figure 7.1, three types of state transitions are identified as a result of mapping of ontology to Markov chain: concept-to-concept state transitions (C_{entity}, $C_{vehicle}$) indicated with bold line arrows, concept-to-instance state transitions (C_{motor}, I_{truck}) indicated with lines with open arrows and instance-to-instance state transitions (I_{truck}, I_{red}) indicated with dash line arrows.

7.2.2 Calculation of transition probability matrix and calculation of concepts importance

The transition probability matrix for Markov chain model will be calculated based on the Page Rank algorithm [19]. We can employ this algorithm since our Markov chain model meets the so called irreducible property. This means that graph is finite and from every state it is possible to go to every other state, and the probability of transition from a state *i* to a state *j* is only dependent on the state *i* and not on the path to arrive at state *j*.

The irreducible property is very important because it guarantees the convergence of the algorithm.

The page rank algorithm will be adjusted with a new parameter called the probability distribution weight (ω)[48]. This parameter determines how probabilities are distributed between states representing classes (S_C), and states representing instances (S_I), following each random jump. If ω = 0, random jump probability is distributed only among instance states, and if ω = 1, random jump probability is distributed only among class states.

Once we get the transition probability matrix, then we can calculate the importance of each concept in a given ontology. The importance (*Imp*) is calculated using Equation 7.1.

$$Imp(c) = -\log_2 \left(\frac{\overrightarrow{e}_{state(c)}}{\sum_{s \in S_c} \overrightarrow{e}_s} \right) \tag{7.1}$$

where $\overrightarrow{e}_{state(c)}$ indicates the principal left eigenvector component calculated from transition probability matrix for the Markov chain state S_c.

7.2.3 Building the concept vector space representation model

The final step of the proposed model is building a concept vector space representation model. The concept vector space model is created using the relevance of ontology concepts (frequency of occurrence of concepts), and the importance of ontology concepts calculated as described in section 7.2.2.

The concept vector space representation model will be employed as a tool in the classification process in order to organize text documents in a structured form. As a result, every new unlabelled text document will be assigned to a predefined category. This will be done by calculating the similarity between the concept vector space representation created by the ontology and the concept vector space representation created by the unlabelled text document. Then, the text document will be assigned to a category having the highest similarity value with respect to that text document.

To evaluate the performance of the proposed model an experiment will be conducted. The aim is to evaluate and compare the classification results, in terms of accuracy (precision/recall), obtained using the enhanced concept vector space representation with results obtained using the traditional concept vector space representational model.

7.3 Conclusion

In this paper, an adaptive concept vector space representation model using Markov Chain model is proposed. The vector space model is one of the most common models for representing text documents in classification process and it can be represented by terms or concepts. The concept representation uses the domain ontology where the document is represented by ontology concepts and their relevance. These concepts make various contributions in classification process and this depends on the position of concepts where they are depicted in the ontology hierarchy. The existing techniques build the concept vector space model calculating the actual position of the concepts in an ontology hierarchy, either manually or empirically through trial and error. We proposed a new approach to automatically estimate the importance indicated by weights of ontology concepts, and to enhance the concept vector space model using automatically estimated weights.

Further research is required on implementation of the proposed model on real domain ontology in order to have a reliable comparison and evaluation of performance with the existing approaches. We also plan to conduct further studies to examine how the proposed model can improve the performance of text document classification process.

A5:
An Improved Concept Vector Space Model for Ontology Based Classification

Publication details
Kastrati, Z., Imran, A., and Yayilgan, S., "*An Improved Concept Vector Space Model for Ontology Based Classification*", in the 11^{th} International Conference on Signal Image Technology & Internet Systems (SITIS'15) (2015), IEEE, pp. 240-245.

Chapter 8

An Improved Concept Vector Space Model for Ontology Based Classification

Abstract

This paper proposes an improved concept vector space (iCVS) model which takes into account the importance of ontology concepts. Concept importance shows how important a concept is in an ontology. This is reflected by the number of relations a concept has to other concepts. Concept importance is computed automatically by converting the ontology into a graph initially and then employing one of the Markov based algorithms. Concept importance is then aggregated with concept relevance which is computed using the frequency of concept occurrences in the dataset.

In order to demonstrate the applicability of our proposed model and to validate its efficacy, we conducted experiments on document classification using concept based vector space model. The dataset used in this paper consists of 348 documents from the funding domain. The results show that the proposed model yields higher classification accuracy comparing to the traditional concept vector space (CVS) model, ultimately giving better document classification performance. We also used different classifiers in order to check for the classification accuracy. We tested CVS and iCVS on Naive Bayes and Decision Tree classifiers and the results show that the classification performance in terms of F1 measure is improved when iCVS is used on both classifiers.

8.1 Introduction

The amount of data produced nowadays is tremendous. According to the computer giant IBM, 2.5 quintillion bytes of data is produced everyday, and this huge volume of data is expected to grow at a massive rate. Before the penetration of Internet and digital devices to household users traditional data was organized and structured neatly into relational databases. Today 80% of the information coming from various sources ranging from transmission sensors to electronic gadgets is unstructured [121]. Organizing and structuring gigantic amount of data is not a trivial task and without it, finding and extracting useful information from massive Internet resources is a challenge [4]. Ontologies play a vital role in this regard.

Ontologies are one of the data representation techniques that not only help better organize data but also help categorize and classify data objects for easy search and retrieval. For instance, text document classification widely employs ontologies to classify and organize text based documents. The text documents are represented using a vector space model [75]. Vector space model consists of concepts extracted from a domain ontology and of concepts relevance which is calculated using frequency of concept occurrences in the dataset of this particular domain. Researchers in [17, 24, 31], have widely used concept vector space model for document classification using only concepts relevance as the classification criteria. Even though this approach has proven useful for document classification of many domains, it however has some limitations. One of the limitations of this approach is that it considers all concepts equally important regardless of where the concepts are depicted in the hierarchy of ontology. The importance is not equal for all concepts and it depends on relations of concepts with other concepts in the ontology hierarchy. Concepts

which have more relations with other concepts are more important than the concepts which have less relations [155].

Therefore, we address this issue in this paper by proposing an improved concept vector space model which takes into account the importance of ontology concepts. The concept importance is computed automatically. To achieve this, we initially convert the ontology into a graph and then implement one of the Markov based algorithms to compute the concept importance. The obtained importance is then aggregated with the concept relevance. Aggregating both concept importance and concept relevance into vector space affects the classification quality yielding an improved classification accuracy.

The remainder of this paper is organised as follows. Section 8.2 describes related work while Section 8.3 presents a detailed description of our proposed concept vector space model. In section 8.4, we describe the implementation and validation of the proposed model. Lastly, section 8.5 presents some conclusions and gives some directions for future work.

8.2 Related Work

The field of document classification has attracted a lot of attention in recent years, thereby resulting in a wide variety of approaches. Depending on the document representation model employed there are two main categories of these approaches relevant to the classification task: 1) Keyword based vector space approach, and 2) Concept (ontology) based vector space approach.

The first approach relies on a set of terms (words) extracted from the documents in the dataset. This approach has some limitations as it does not consider the dependency between the terms and it also ignores the order and the syntactic structure of the terms in the documents. To overcome these limitations, concept based vector space approach comes into effect. This approach relies on a set of concepts taken from a domain ontology to derive the semantic representation of documents. A drawback of this approach is that it considers all concepts equally regardless of where in the hierarchy the concepts occur. There have been some efforts to find concepts importance depending on the position of concepts where they are depicted in the hierarchy. For instance, researchers in [120] used three different weights for concepts depending on the position where they occur in the ontology hierarchy. The first weight was assigned to concepts which are occurring as classes, second weight for concepts occurring as subclasses and the third weight for concepts occurring as instances. The value of these weights is set empirically through trial and error by conducting experiments. The value of 0.2 is set for the concepts which occur as classes, 0.5 for concepts occurring as subclasses and 0.8 when concepts occur as instances.

A slightly different approach of computing weights was implemented in [42, 54] where layers of ontology tree are used to represent the position of concepts in the ontology. The weight of each concept is then computed by counting the length of path from the root node to the given concept. The same approach of using layer for calculating weight values for concepts were used in [116]. Path length is also used to compute the weight of concepts but rather than considering all ontology concepts, only the leaf concepts were used. The idea behind this approach was that more general concepts, such as superclasses, are implicitly taken into account through the use of leaf concepts by distributing their weights to all of their subclasses down to the leaf concepts in equal proportion.

The drawback of these approaches is that they compute the concepts weight either empirically through trial and error by conducting experiments thus keeping these weights fixed or using the path length. Furthermore, the approach presented in [116] uses only the top-level ontology for computing weights. Our approach uses a Markov based PageRank algorithm to compute the concept importance. The algorithm uses all concepts of ontology and the importance of a concept is computed relative to all other concepts in the ontology.

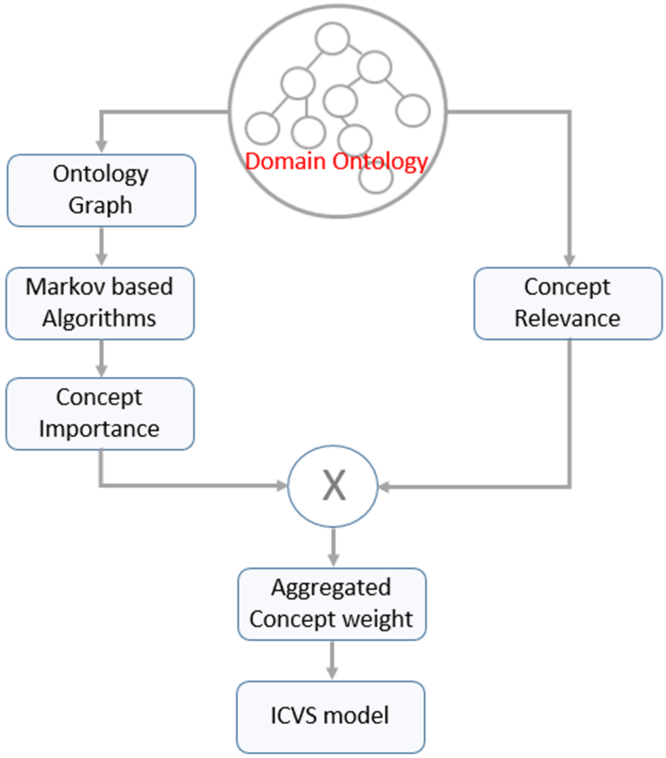


Figure 8.1: The proposed model for an improved concept vector space (iCVS) model

8.3 Proposed model and methodology

The main goal of the proposed model is to improve the concept vector space representation model (refer to subsection 8.3.3 for details) for classifying text documents with higher accuracy and for calculating concepts weight automatically. The concept vector space model is enriched with a new parameter, namely concept importance (*Imp*). Concept importance aggregated with concept relevance (*Rel*) forms the concept weight. Our proposed model is given in Figure 8.1. From the domain ontology, shown in the model, concept importance and concept relevance are derived. These two are then combined into an aggregated concept weight.

The proposed model involves three steps; 1) mapping the domain ontology into an ontology graph, 2) applying Markov based algorithms to compute the importance of each concept in the ontology graph, and 3) the final step is building an iCVS model using both concept importance and concept relevance. The detailed explanation of each of these steps is given in the following subsections.

8.3.1 Mapping Domain Ontology into Ontology Graph

The first and the foremost step is to convert the domain ontology into an ontology graph for calculating concept importance.

A domain ontology is a data model which represents the concepts and the relations

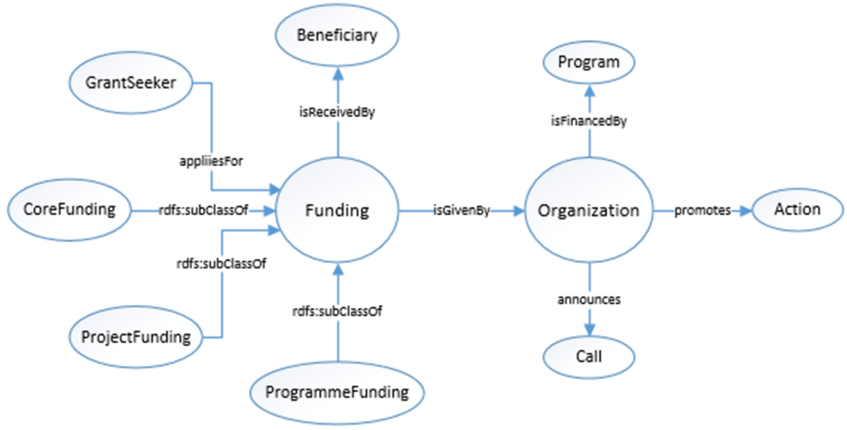


Figure 8.2: A part of INFUSE ontology graph

between them in a given domain. To represent this data structure, we adopt a model where the ontology is represented as a directed acyclic graph. The modelling is an equivalent mapping which means that an ontology concept is mapped into a graph vertex and an ontology relation into a graph edge which connects two vertices. The formal definition of this graph, known as ontology graph, is given as follows.

Definition 8.1
Given a domain ontology O, the ontology graph $G = \{V, E, f\}$ of O is a directed acyclic graph, where V is a finite set of vertices mapped from concepts in O, E is a finite set of edge labels mapped from relations in O, and f is a function from E to $V \times V$.

In Figure 8.2, we present part of the INFUSE ontology graph which consists of a subset of concepts and relations from the funding domain. The details of the INFUSE ontology are given in Section 8.4.

In the semantic web, a formal syntax for defining ontologies is Web Ontology Language (OWL) and Resource Description Framework (RDF) Schema. These languages represent the ontology as a set of Subject-Predicate-Object (*SPO*) expressions known as RDF triples. The set of RDF triples is known as RDF graph where subject is the source vertex and object is the destination vertex, and predicate is a directed edge label which links those two vertices. The formal definition of RDF graph is given as following.

Definition 8.2
Given a set of RDF triples T, the RDF graph $G = \{V, E, f\}$ of T is a directed acyclic graph, where V is a finite set of vertices (subjects and objects) in G defined as $V = \{v_u : u \in (S(T) \cup P(T))\}$, E is a finite set of edge labels (predicates) in G defined as $E = \{e_{SPO} : SPO \in T\}$, f is a function linking subject S to an object O by an edge E defined as $f = \{f_P : f_P = V_S \rightarrow V_O, V_S, V_O \in T\}$

The ontology graph and RDF graph are not the same for a given ontology. The difference is that a relation in an ontology graph is defined as a vertex in the RDF graph. For example, relation *isReceived* in ontology graph shown in Figure 8.2 is represented as a vertex in RDF graph, as shown in Figure 8.3. In other words, a relation in RDF graph is a link between a subject denoted by *rdfs:domain* property and an object denoted by *rdfs:range* property as given in Definition 8.2.

Figure 8.3: An example RDF graph representation

8.3.2 Markov Based Algorithms

In order to compute the importance of vertices of the graph, we adopt the Markov based algorithms. The graph can be either ontology graph or RDF graph as defined in subsection 8.3.1. The idea behind Markov based algorithms is representing the graph as a stochastic process, more concretely as a first-order Markov chain where the importance for a given vertex is defined as the fraction of time spent traversing that vertex for an infinitely long time in a random walk over the vertices. The probability of transitioning from a vertex *i* to a vertex *j* is only dependent on the vertex *i* and not on the path to arrive at vertex *j*. This property, known as the Markov property, enables the transition probabilities to be represented as a stochastic matrix with non-negative entries and the maximum probability of 1.

In this paper, we use PageRank [19] algorithm as one of the most well known and successful example of Markov based algorithms [153].

A simplified principle of work of PageRank algorithm is as follows. It initially defines the importance of a vertex *i* as given in Equation 8.1.

$$PR(i) = \sum_{j \in V_i} \frac{PR(j)}{Outdegree(j)} \tag{8.1}$$

where, *PR(j)* is the importance of vertex *j*, V_i is the set of vertices that links to vertex *i*, and *Outdegree(j)* is the number of vertices that have outlinks from vertex *j*.

As we can see from the Equation 8.1, the PageRank is an iterative algorithm. It assigns an initial importance to a vertex *i* as shown in Equation 8.2.

$$PR^{(0)}(i) = \frac{1}{N} \tag{8.2}$$

where, *N* is the total number of vertices in the graph. Then PageRank iterates as per Equation 8.3 and continues to iterate until a convergence criterion is satisfied.

$$PR^{(k+1)}(i) = \sum_{j \in V_i} \frac{PR^{(k)}(j)}{Outdegree(j)} \tag{8.3}$$

The process can also be defined using the matrix notation. Let *M* be the square, stochastic transition probabilities matrix corresponding to the directed graph *G*, and *Imp(k)* is the Importance vector at the k^{th} iteration. Then the computation of one iteration corresponds to the matrix-vector multiplication as shown in Equation 8.4.

$$PR^{(k+1)} = M * PR^{(k)} \tag{8.4}$$

The entry of transition probability matrix *M*, for a vertex *j* which links to vertex *i*, is defined using Equation 8.5.

$$p_{i,j} = \begin{cases} \frac{1}{Outdegree(j)}, & \textit{if there is a link from j to i} \\ 0, & \textit{otherwise} \end{cases} \tag{8.5}$$

There are two properties which are necessary to be satisfied in order for a Markov based algorithm to converge; It should be aperiodic and irreducible [112]. The transition probability matrix *M* is a stochastic matrix with probability 1 and this makes the PageRank

algorithm aperiodic. The PageRank algorithm is not irreducible due to the definition given in Equation 8.5, where some of the transition probabilities in matrix *M* may be 0. This does not meet the criteria of irreducibility property which requires the transition probabilities to be greater than 0.

To make the PageRank algorithm irreducible in order to converge, a damp factor $1 - \alpha$ is introduced. As a result of this, a new transition probability matrix M^* is defined where a complete set of outgoing edges with probability α/N are added to all vertices in graph. The definition of matrix M^* is given in Equation 8.6.

$$M^* = (1 - \alpha)M + \alpha \left[\frac{1}{N}\right]_{N \times N} \tag{8.6}$$

The damp factor besides enabling the PageRank algorithm to converge also overcomes the problem of rank sinks [112].

Finally, replacing M^* with M in Equation 8.4, the PageRank algorithm is defined as given in Equation 8.7.

$$PR^{(k+1)} = (1 - \alpha)M \times Pr^{(k)} + \alpha \left[\frac{1}{N}\right]_{N \times N} \tag{8.7}$$

8.3.3 Building the Concept Vector Space Model

The final step of the proposed model is building a concept vector space model. This model consists of two components: concepts and their weights.

Concepts are taken from the domain ontology using the matching method **[?]**. The idea behind this method is to search for concepts in the ontology that have labels matching either partially or exactly/fully with a dataset term. The obtained concepts are then represented as a concept vector space model. Exact matches represent cases where a concept label is identical with the term extracted from the documents in the dataset. Partial matches represent cases when concept label contains terms extracted from the document in the dataset. The formal definition of exact and partial matches is given as the following.

Definition 8.3
Let O be the domain ontology and D the dataset composed of documents of this given domain. Let $d \in D$ be a document defined as a finite set of terms $d = \{t_1, t_2, ..., t_n\}$.

The mapping of term $t_i \in d$ into concept $c_j \in O$ is defined as exact match $EM(t_i, c_j)$, where

$$EM(t_i, c_j) = \begin{cases} 1, & \text{if label } (c_j) = t_i \\ 0, & \text{if label } (c_j) \neq t_i \end{cases} \tag{8.8}$$

The mapping of term $t_i \in d$ into concept $c_j \in O$ is defined as partial match $PM(t_i, c_j)$, where

$$PM(t_i, c_j) = \begin{cases} 1, & \text{if label } (c_j) \text{ contains } t_i \\ 0, & \text{if label } (c_j) \text{ does not contain } t_i \end{cases} \tag{8.9}$$

If $EM(t_i, c_j) = 1$, it means that term t_i and concept label c_j are identical, then term t_i is replaced with concept c_j. For example, for a concept in the ontology such as *Organization* or *Call* as shown in Figure 8.2, there exists an identical term extracted from the document.

If $PM(t_i, c_j) = 1$, it means that term t_i is part of concept label c_j, then term t_i is replaced with concept c_j. For example, the *ProjectFunding* compound ontology concept shown in Figure 8.2, contains terms extracted from the document such as *Project* and/or *Funding*.

Weight of concepts, as the second component of concept vector space model represented by the tuple shown in Equation 8.12, is computed using concept importance and concept relevance. The value of a concept weight is in the range of [0,1] because both concept importance and concept relevance are normalized.

$$w(c_i) = Imp(c_i) \times Rel(c_i) \tag{8.10}$$

Concept importance *Imp* is computed using the PageRank algorithm as described in Subsection 8.3.2, while concept relevance *Rel* is computed using Equation 8.11.

$$Rel(c_i) = \sum_{i=1}^{m} Freq(c_i) \tag{8.11}$$

where *Freq*(c_i) is the frequency of occurrences of a concept c_i in the dataset.

Finally, a document is represented using concept vector space representation model by the following tuple:

$$d_i = \{(c_1, w_1), (c_2, w_2), (c_3, w_3), ..., (c_i, w_i)\} \tag{8.12}$$

where c_i is the i^{th} concept of the domain ontology and w_i is the weight of the concept c_i computed using Equation 10.1.

Table 8.1 illustrates an example of building new proposed concept vector space model and its implementation for representing the documents to be classified.

Table 8.1: An example of building concept vector space

Doc	GeographicalArea			Applicant		
	Imp	Rel	w	Imp	Rel	w
d1	0.130	0.797	0.104	0.020	0.797	0.016
d2	0.130	0.624	0.081	0.020	0.624	0.012
d3	0.130	0.000	0.000	0.020	0.860	0.017

8.4 Results and Analysis

The dataset used in this paper consists of 348 grant documents that had been collected and classified into 5 categories by field experts as part of the INFUSE [1] project.

The dataset is split randomly, in which 70% of the documents (244) are used to build the classifier and the remaining 30% (104) to test the performance of the model. The number of documents in each category varied widely, ranging from the Society category which contains 125 documents to the Music category which contains only 10 documents. Table 8.2 shows 5 categories along with the number of training and testing documents in each category.

Table 8.2: Dataset size

No	Category	# Train	# Test	Total
1	Culture	75	32	107
2	Health	52	28	80
3	Music	8	2	10
4	Society	90	35	125
5	Sportssociety	19	7	26
6	Total	244	104	348

The ontology used for experimenting in this paper is from the same domain as the dataset and it consists of 85 concepts and 18 object properties which connect these concepts. To compute concepts importance, we have used the RDF rank algorithm. This algorithm is part of the extensions module of GraphDB [111] and it computes the importance for every

[1] http://infuse.scan4news.com/?cat=4

Table 8.3: Concept importance for the top ten concepts of the INFUSE ontology

No	Concept	Concept Importance
1	Coverage	0.20
2	GeographicalArea	0.13
3	Topic	0.11
4	County	0.07
5	Participant	0.06
6	Programme	0.05
7	Organisation	0.05
8	Funding	0.05
9	Applicant	0.04
10	Candidate	0.04

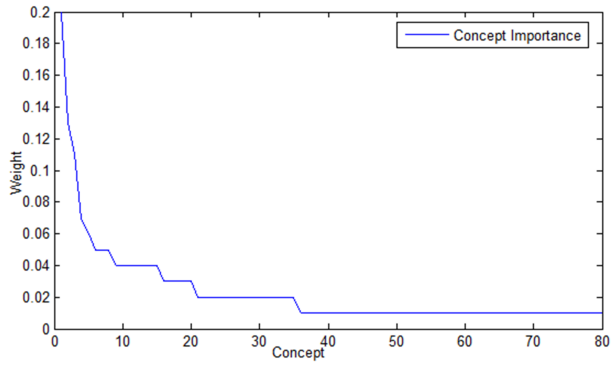


Figure 8.4: Concept importance for all concepts of the INFUSE ontology

vertex in the entire RDF graph. Table 8.3 shows the concept importance values of the top ten concepts of the INFUSE ontology computed using the proposed method described in subsection 8.3.2. The concept importance is a floating point number with values varying between 0 and 1.

Figure 8.4 shows the concept importance values in ranking order after having computed them for all the concepts of the INFUSE ontology. As can be seen from the chart diagram, the concept importance is different for different concepts, varying from 0.2 - 0.02 for almost half of the concepts set, while for the rest of the concepts it is 0.01.

In order to demonstrate the general applicability of our proposed model and to validate its effectiveness, we conducted experiments on documents classification using concept based vector space model. To achieve this, we initially performed the document classification using the traditional CVS. This CVS consists of ontology concepts and their relevances computed using Equation 8.11. Decision Tree (J48) implemented in the open source platform Weka [57] is used as a classifier and the INFUSE dataset is used for training and testing the classifier. The standard information retrieval measures such as precision, recall and F1-measure are used to evaluate the performance of the document classification. The obtained results are shown in Table 8.4.

In the second experiment, we performed the document classification using the same classifier (Decision Tree) and the same dataset but employing the iCVS model as proposed in subsection 8.3.3. The proposed iCVS, in addition to the concepts relevance, also takes into account the concepts importance computed using the PageRank algorithm. Precision, recall and F1 results of each category and the weighted average precision, recall and F1

Table 8.4: The performance of Decision tree classifier using CVS

Category	Precision (%)	Recall (%)	F1 (%)
Culture	76.5	81.3	78.8
Health	77.3	60.7	68.0
Music	16.7	50.0	25.0
Society	77.8	80.0	78.9
Sportssociety	33.3	28.6	30.8
Weighted avg.	**73.1**	**71.2**	**71.6**

Table 8.5: The performance of Decision tree classifier using iCVS

Category	Precision (%)	Recall (%)	F1 (%)
Culture	76.5	81.3	78.8
Health	82.6	67.9	74.5
Music	100.0	50.0	66.7
Society	80.0	91.4	85.3
Sportssociety	33.3	28.6	30.8
Weighted avg.	**76.9**	**76.9**	**76.4**

over all tests set are shown in Table 8.5.

As can be seen from the results shown in Table 8.4 and Table 8.5, the proposed iCVS model yields higher weighted average classification accuracy compared to the traditional CVS model. The high accuracy is achieved as a result of using both the concept relevance and concept importance. The concept relevance reflects the contribution of a concept to a document/category by the frequency of the occurrences of a concept in that document/category alone. In other words the higher the frequency of occurrences of a concept the more relevant it is. The concept importance reflects the contribution of a concept to an ontology as a combination of incoming and outgoing edges of the concept. This combination represents important concepts in an ontology. Therefore, multiplying the frequency of occurrences of concepts and their importance yields concepts with weight having better discriminative power, ultimately giving better classification performance.

The classification performance is improved in almost all categories. For example, the Music category has achieved a 100.0% precision using iCVS, compared to a precision value of 16.7% with CVS. A considerable improvement is observed for each of Health and Society categories as well.

Another criteria to evaluate the performance of CVS and iCVS is to check for the classification accuracy using different classifiers. We employed CVS and iCVS for Naive Bayes and Decision Tree classifiers. The obtained results, illustrated in Figure 8.5, show that the classification performance in terms of F1 measure improves when iCVS is used on both classifiers.

8.5 Conclusion and Future Work

In this paper, we have proposed an improved concept vector space model which takes into account the importance of ontology concepts. Concept importance is computed automatically and this is done by converting the ontology into a graph and then employing the PageRank algorithm. Importance of ontology concept is then incorporated in the CVS in addition to the concept relevance which is computed using the frequency of appearances of a concept in the dataset.

The experiments conducted on document classification showed that the proposed model yields higher weighted average classification accuracy comparing to the traditional concept vector space model. Employing CVS and iCVS on Naive Bayes and Decision Tree classi-

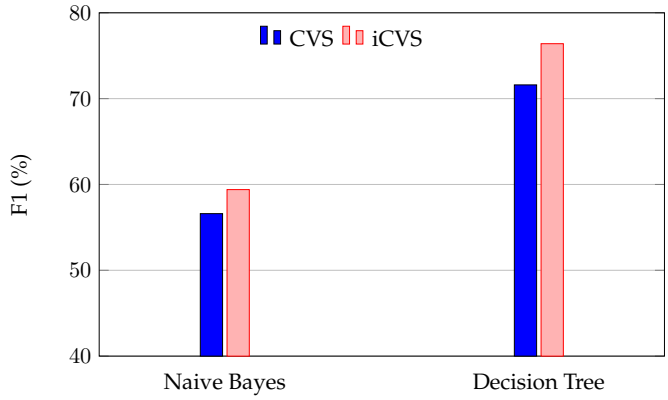


Figure 8.5: F1 measure of two different classifiers using CVS and iCVS on the INFUSE dataset

fiers demonstrates that the classification performance in terms of F1 measure is improved when iCVS is used on both classifiers. More concretely, the F1 measure is improved from 56.6% to 59.4% for the Naive Bayes and from 71.6% to 76.4% for the Decision Tree when iCVS is used. Those results validate the applicability of the proposed method and making it a generic model which can be applied to classify documents efficiently.

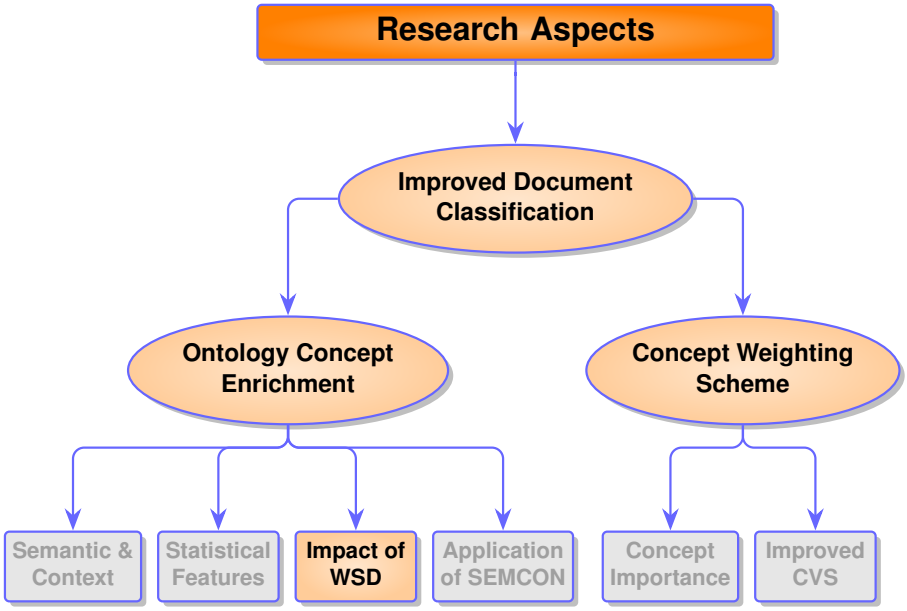
The future work includes implementing and testing other Markov based algorithms for computing concept importance and compare those with the PageRank algorithm for document classification.

Acknowledgment

The authors would like to thank Cristina Marco from the INFUSE project for providing the domain ontology and the dataset used in this paper.

Part IV

Improving Document Classification



This part focuses on the third research aspect of this thesis 'Improving document classification' which is on top of the two previous research aspects. The work presented in this part intended to answer the first and the third (second part) research question listed in Section 1.2.

This part is comprised of two chapters constituted by two published research articles. Chapter 9 presents a classification approach that utilizes an ontology for labelling text documents. The ontology is primarily enriched with new concepts and then it is used to exploit the background knowledge for document representation.

Chapter 10 gives an ontology-based document classification approach which involves enriching of a baseline ontology with new concepts, and the new scheme for evaluating weights of concepts. The effect of disambiguation issue on the quality of ontology enrichment in general and on the classification performance in particular is also investigated in this chapter. Extensive experiments using a real dataset and ontology are conducted to validate our approach.

A6: Automatically Enriching Domain Ontologies for Document Classification

Publication details
Kastrati, Z., Yayilgan, S., and Hjelsvold, R., "*Automatically Enriching Domain Ontologies for Document Classification*", in the 6^{th} International Conference on Web Intelligence, Mining and Semantics (WIMS'16) (2016), ACM, pp. 1-4.

Chapter 9

Automatically Enriching Domain Ontologies for Document Classification

Abstract

The ontology-based document classification approach relies on the content meanings of a given domain exploited and captured using ontologies of this particular domain. Domain ontologies consist of a set of concepts and relations which link these concepts. However, they often do not provide an in-depth coverage of concepts thereby limiting their use in some subdomain applications. Therefore, the techniques for enhancing ontologies, particularly ontology enrichment, have emerged as an essential prerequisite for ontology-based applications. In this paper, we propose a new objective metric called SEMCON to enrich the domain ontology with new terms. To achieve this, SEMCON combines semantic as well as contextual information of terms within the text documents. Experiments are conducted to demonstrate the applicability of the proposed model and the obtained results from the funding domain show that document classification achieved better performance using the enriched ontology in contrast to using the baseline ontology.

9.1 Introduction

The ontology-based document classification approach relies on the content meanings rather than on literal strings (keyword). The content meaning is exploited and captured using domain ontologies. Domain ontologies represent the basic vocabulary of a given domain. They provide a broad coverage of concepts and relations connecting these concepts for a particular domain but an in-depth coverage of concepts is often not available. Therefore, the techniques for modifying ontologies have emerged as an essential prerequisite for ontology-based applications. In this regard, ontology enrichment is one of these techniques that plays an important role in the ontology-based document classification process.

Enrichment of ontology concepts is an automatic process aiming at improving a given ontology by enriching it with new terms and it is a part of the iterative ontology engineering process [40]. The enrichment process departs from an existing ontology. It then exploits available textual data from the domain of the given ontology in order to find synonyms or linguistic variants of the existing concepts in the ontology. Finally, concepts are formed by employing the learning approach, which is the most important step in the process of ontology enrichment.

There are different learning approaches available to enrich concepts of an ontology. Although these approaches have proved useful for enriching ontologies of many domains, they however have some drawbacks. The major limitation of these approaches is that they are dependent on only statistical (context) or semantic information of terms. Therefore, this paper addresses this limitation by proposing a learning approach called SEMCON [69] to enriching the domain ontology with new terms.

The SEMCON uses text documents as input for ontology learning process and it is composed of two parts - contextual and semantic. Context is derived using the cosine distance between the feature vectors of any two terms. The feature vectors are composed of values computed by both the frequency of occurrences of terms in corresponding documents, and

the statistical features such as font types and font sizes. The semantic on the other hand is defined by computing a semantic similarity score using lexical database WordNet.

The proposed SEMCON model is tested on classifying documents from the funding domain.

The structure of the paper is as the following. Section 9.2 shows the state of the art in the field of ontology enrichment. Section 9.3 describes the proposed method to enrich concepts of a domain ontology. In Section 9.4 we present the results and analyze them while Section 9.5 concludes the paper.

9.2 Related Work

There are to main categories of approaches relevant to the ontology concepts enrichment task: statistical approach and semantic approach.

Statistical approach exploits domain specific textual data to enrich a domain ontology. It involves term frequency (*tf*, *tf*idf*) technique and term co-occurrence (collocation) feature to identify and extract relevant terms from the textual data. An example of employing collocation feature to enriching ontologies is used in [92]. More concretely, researchers introduced the notion of higher order co-occurrences to find semantically related words automatically from a given corpus in order to enrich an ontology. Higher order co-occurrences represent the N highest ranked co-occurrences of a term computed through an iterative process. The co-occurrence, at each iteration, is found using the frequency of appearances of two words together in similar context (sentence). Another research, which involves using co-occurrence feature to discover related terms for extending ontologies, is conducted also in [86]. This research proposes a system to semi-automatically extend ontologies by mining textual data from the Web sites of international online media. They assume that two semantically related terms regularly co-occur in the same text segments. The Log Likelihood Algorithm is used to analyze the significance of co-occurrence of the target term, both at the sentence level and the document level. Parekh et al in [113] use a similar approach to enriching domain ontologies. Domain specific texts are exploited in order to automatically generate sets of terms which are related to each other based on their lexical co-occurrence within similar contexts.

Semantic approaches involve semantic similarity and relatedness measure based on WordNet to identify and extract terms from textual data for enriching ontologies. [142] is an example of employing semantic approach for enriching ontologies. Three WordNet based similarity measure namely, domain path between the synsets of two words, Resnik, Jiang and Conrath, are used to enrich the ontology from the music domain. Warin et al in [152] perform the same approach to enrich an ontology but by employing five semantic similarity measures based on WordNet.

Our proposed approach utilizes to some extent both approaches, statistical and semantic. By aggregating both contextual and semantic information we expect to obtain the most relevant terms for enriching the ontology concepts and thus to achieve better classification performance.

9.3 Proposed Model

The proposed model, shown in Figure 9.1, is composed of 4 modules, which are described in the following subsections.

9.3.1 Preprocessing Module

This module initially collects all documents that are in a particular dataset. A morpho-syntactic analysis is then performed on these documents. Documents are cleaned to remove all punctuation and capitalization. This is followed by a tokenizing step to separate the

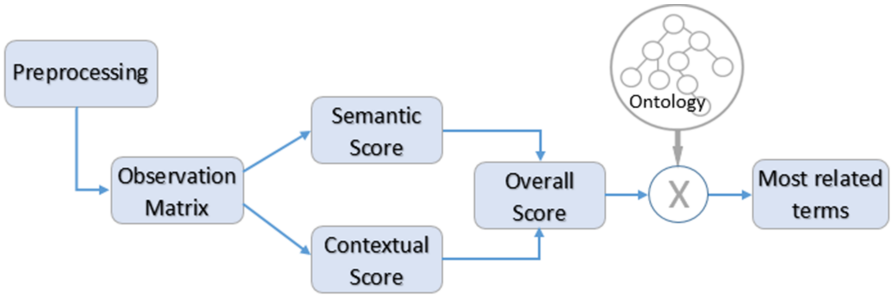


Figure 9.1: Block diagram of SEMCON model

textual data into individual terms. Finally, the normalized form of these terms is computed using the lemmatization step. The final output of the preprocessing step is a list of noun terms.

9.3.2 Observation Matrix Module

The next phase of SEMCON is computation of the observation matrix. Observation matrix is a rectangular matrix where columns represent documents of particular domain and rows are the terms extracted from those documents.

Each entry of the observation matrix is computed by aggregating the frequency of term occurrences, font sizes and font types of this term appearing in a document. An entry of observation matrix is calculated using Equation 9.1.

$$E_{i,j} = tf_{i,j} + \sum_{k \in tf} (ft_{i,j,k} + fs_{i,j,k}) \tag{9.1}$$

where, $tf_{i,j}$ shows the frequency of occurrences of a term i in document j. $ft_{i,j,k}$ and $fs_{i,j,k}$ show the added value of font types and font sizes computed over all occurrences k of a term i in a document j.

Algorithm 9.1 shows the way how the observation matrix is built. The input of the algorithm is a collection of documents. In this paper, the collection of documents from which font sizes and font types of terms used to build the observation matrix are derived are saved in pdf format. However, the input of algorithm can be a collection of documents other than pdfs because font sizes and font types of terms can be computed for all types of rich texts using the html tags.

Algorithm 9.1 describes the computation of observation matrix using bold font type of a term and 4 different font sizes. More concretely, line 3-13 in the algorithm shows entries of the observation matrix computed using the number of times a term appears in bold (α) and the number of times with font sizes (β) as, either level 3 (line 4), level 2 (line 7), level 1 (line 10), or title (line 13). In the same way, we computed entries of the observation matrix for the terms which appear in a document as either italic, underline, and regular and with font sizes as either level 3, level 2, level 1, or as title.

9.3.3 Contextual and Semantic Module

The observation matrix is used as an input to compute contextual and semantic score for all pairs of terms t_i, t_j.

Algorithm 9.1: Calculation of Observation Matrix

Input : A collection of pdf documents
Output: Entries of the observation matrix

```
for each Doc ∈ D do
    for each t ∈ Doc do
        if t ∈ Doc is bold then
            if t_size < 10pt then
                Compute E as E + tf + 0.75 * α + 0.25 * β
            end
            if 10pt ⩽ t_size < 14pt then
                Compute E as E + tf + 0.75 * α + 0.50 * β
            end
            if 14pt ⩽ t_size < 18pt then
                Compute E as E + tf + 0.75 * α + 0.75 * β
            end
            if t_size ⩾ 18pt then
                Compute E as E + tf + 0.75 * α + 1.00 * β
            end
        end
    end
end
return E;
```

Contextual information score, $S_{con}(t_i, t_j)$, for a pair of terms t_i and t_j is computed using the cosine similarity metric with respect to the documents, as given in Equation 9.2.

$$S_{con}(t_i, t_j) = \frac{t_i \cdot t_j}{\| t_i \| \| t_j \|} \tag{9.2}$$

where, t_i and t_j show the term vectors of the observation matrix. The dot product between two term vectors reflects the similarity between these two terms in the vector space.

Next step is computation of the semantic score. The semantic score is computed using the information found in the lexical database WordNet by employing the Wu&Palmer similarity measure [156]. The semantic score, $S_{sem}(t_i, t_j)$, is computed for all possible pairs of terms t_i and t_j of the observation matrix, where $t_i, t_j \in O$ and O is the observation matrix. The first sense heuristic technique is used as a baseline to find the correct sense of terms t_i and t_j. The semantic score mathematically is given in Equation 9.3.

$$S_{sem}(t_i, t_j) = \frac{2 * depth(lcs)}{depth(t_i) + depth(t_j)} \tag{9.3}$$

where, *depth(lcs)* denotes the least common subsumer of terms t_i and t_j; *depth(t_i)* and *depth(t_j)* denote the depth of the path of term t_i and t_j, in the WordNet database.

9.3.4 Overall Score Module

The overall score between two terms t_i, t_j, shown in Equation 9.4, is computed using contextual and semantic score.

$$S_{overall}(t_i, t_j) = w * S_{con}(t_i, t_j) + (1 - w) * S_{sem}(t_i, t_j) \tag{9.4}$$

where w is a parameter whose value is set to 0.5 based on the empirical analysis.

Next we search for concepts in the ontology that have labels matching either partially or fully with a term. As a result of this, a hash table with concepts and the relevant terms for enriching these concepts with their ranked overall score is built. Lastly, we apply a rank cut-off method considering only terms that are above the specified threshold (Top-N) as the most relevant terms to enrich given concepts of the baseline ontology.

9.4 Results and Analysis

The dataset used in this paper is a real dataset consists of 467 grant documents that had been collected and classified into 5 categories by the field experts as part of the INFUSE [1] project. The dataset is split randomly, in which 70% (327) of the documents are used to train the classifier and the remaining 30% (140) to test the performance of the classifier.

The ontology used in this paper is from the funding domain. It is composed of 85 concepts and 18 object properties which link these concepts. Figure 9.2 presents part of the INFUSE ontology which consists of a subset of concepts and their relationships from the funding domain.

We choose to use a real ontology for experimenting due to the fact that the ontology to be enriched and the dataset exploited for enriching the given ontology have to be from the same domain.

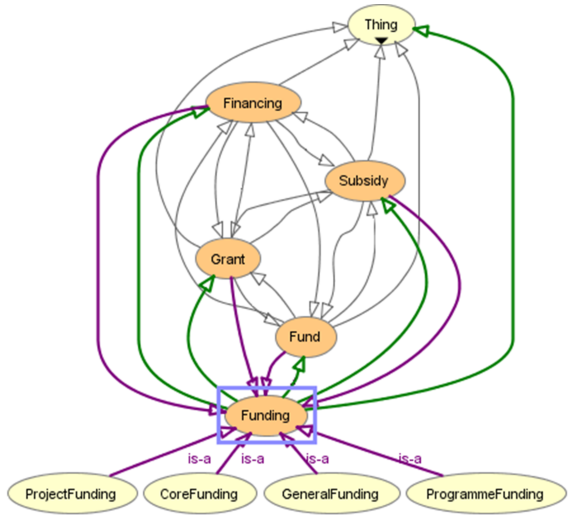


Figure 9.2: A part of the INFUSE ontology

A preprocessing step is done on all pdf documents to extract only noun terms and their font types and font sizes in order to built the observation matrix. To achieve this, we used Apache PDFBox library. This is an open source Java library which allows creation of new pdf documents, manipulation of existing documents and the extraction of content from documents. The observation matrix is used then as input to compute the contextual and semantic score. First sense heuristic is employed to find the correct sense of terms as part of the semantic score computation. First sense heuristic also known as the predominant sense is a technique used in word sense disambiguation to find the correct sense of a term. It assumes that the most common sense of a word represent the correct sense of this given

[1] http://infuse.scan4news.com/?cat=4

Table 9.1: The Top-5 terms obtained by SEMCON using first sense heuristic disambiguation technique

Terms	Financing	Fund	Funding	Grant	Programme	Subsidy
Top-1	Finance	Budget	Fund	Welfare	Rule	Subvention
Top-2	Investment	Amount	Amount	Support	Relevance	Grant
Top-3	Participation	Funding	Part	Partner	Framework	Welfare
Top-4	Implementation	Finance	Subsistence	Cost	Participation	Loss
Top-5	Compliance	Share	Grant	Commission	Implementation	Scholarship

word. Finally, an aggregated contextual and semantic score is found through which the proposed model gives a list of ranked terms which are the most relevant terms for enriching a given concept. The Top-5 terms obtained by the proposed model as the most relevant terms for enriching the INFUSE ontology concepts shown in Figure 9.2, is given in Table 9.1.

In order to demonstrate the applicability of our proposed model on enriching the ontology with new terms and to validate its effectiveness, we conducted experiments on document classification using the baseline ontology (existing INFUSE ontology) and the ontology after it has been enriched. Decision Tree is used as a classifier and the INFUSE dataset is used for training and testing the classifier.

The standard information retrieval measures such as precision, recall and F1-measure are used to evaluate the performance of the document classification.

The first experiment conducted on document classification uses the baseline ontology and the obtained result for each category and the weighted average result of all categories is shown in Table 9.2.

Table 9.2: The performance of the Decision tree classifier using the baseline ontology

Category	Precision (%)	Recall (%)	F1 (%)
Culture	63.8	66.7	65.2
Health	81.8	75.0	78.3
Music	50.0	16.7	25.0
Society	62.0	75.6	68.1
Sportssociety	50.0	33.3	40.0
Weighted avg.	**66.1**	**66.4**	**65.5**

The second experiment performed on document classification uses the same conditions, in terms of the classifier (Decision Tree) and the dataset (INFUSE) used as in the previous experiment but now it employs the ontology after it has been enriched. The Top-5 terms achieved by the proposed model are used for enriching the baseline ontology. The result for each category and the weighted average results of the dataset are shown in Table 9.3.

Table 9.3: The performance of the Decision tree classifier using the enriched ontology

Category	Precision (%)	Recall (%)	F1 (%)
Culture	72.0	80.0	75.8
Health	87.5	77.8	82.4
Music	33.3	16.7	22.2
Society	83.7	87.8	85.7
Sportssociety	58.3	58.3	58.3
Weighted avg.	**76.6**	**77.1**	**76.6**

As can be seen from the results shown in Table 9.2 and Table 9.3, the higher weighted average classification performance is achieved when the classification is performed using the

ontology whose concepts are being enriched with new terms comparing to the classification using the baseline ontology. An improvement of document classification performance is observed for almost all categories. For example, the Society category has achieved an 85.7% F1 measure using the enriched ontology, compared to a F1 measure value of 68.1% using the baseline ontology. On the contrary, only the Music category achieves slightly worse performance when the classification is performed using the enriched ontology comparing to the classification performed using the baseline ontology. This happened due to the fact that some terms, e.g. *Amount*, used to enriching the baseline ontology concepts are not homogeneous terms. In other words, these terms are likely to occur in documents belonging to the categories rather than music category and thus they are non-discriminative in terms of classification.

Another evaluation criteria is to investigate into how much the number of terms used to enrich concepts of an ontology affects the classification performance. This is achieved using Top-N terms, where N=1-5. Top-1 means the very top first term obtained by the proposed model is used to enrich a particular concept of the baseline ontology, Top-2 means the top 2 terms and so on. Performance of the Decision tree achieved on the funding domain is shown in Table 9.4.

Table 9.4: The performance of the Decision tree classifier using the Top-N terms

Ontology	Precision (%)	Recall (%)	F1 (%)
Baseline	66.1	66.4	65.5
Baseline + Top-1	75.1	74.3	72.7
Baseline + Top-2	71.0	73.6	70.9
Baseline + Top-3	74.4	77.9	75.8
Baseline + Top-4	77.1	77.9	77.1
Baseline + Top-5	76.6	77.1	76.6

Furthermore, Table 9.4 shows a comparison of precision, recall and F1 measures achieved by Decision tree classifier, when the document classification is performed using the baseline ontology and when the classification is done using the ontology enriched with new terms, starting from 1 term to 5 terms for a particular concept. The average F1 measure is improved from 65.5% to 77.1% when the classification is performed by employing the ontology whose concepts have been enriched with 4 new terms.

9.5 Conclusion and Future Work

In this paper, we presented an ontology based classification approach for classifying text documents from the funding domain. To achieve this, a baseline ontology is primarily enriched with new terms. For enriching the baseline ontology, we used a new learning technique called SEMCON which combines the semantic and contextual information.

The proposed approach is tested on the document classification using the baseline ontology and the enriched ontology. Results of the experiments showed that the performance of the document classification conducted using the enriched ontology is improved comparing to the performance of the classification accomplished using the baseline ontology.

In future work we plan to employ other word sense disambiguation techniques, i.e machine learning, in order to improve the ontology enrichment.

A7:
Supervised Ontology-Based Document Classification Model

Publication details
Kastrati, Z., and Yayilgan, S., "*Supervised Ontology-Based Document Classification Model*", in the International Conference on Compute and Data Analysis (ICCDA'17) (2017), ACM, pp. 1-7.

Chapter 10

Supervised Ontology-Based Document Classification Model

Abstract

Ontology-based document classification relies on background knowledge exploited by ontologies to represent documents. Background knowledge is embedded in a document using the exact matching technique. The basic idea of this technique is to map a term to a concept by searching only the concept labels that explicitly occur in a document. Searching only the presence of concept labels limits the capabilities to capture and exploit the whole conceptualization involved in user information and content meanings. Therefore, to address this limitation, we propose a new document classification model based on ontologies. The proposed model uses background knowledge derived by ontologies for document representation. It associates a document with a set of concepts by not only using the exact matching technique but also by identifying and extracting new terms which can be semantically related to the concepts of ontologies. Additionally, the proposed model employs a new concept weighting technique which computes the weight of a concept using the relevance and the importance of the concept.

We conducted several experiments using a real ontology and a dataset to test our proposed model. The results obtained by experiments run on 3 different classification algorithms using the baseline ontology, the improved concept vector space model by using the new concept weighting technique, and the enriched ontology, show that our proposed model achieved a considerable improvement of classification performance.

10.1 Introduction

Document classification also known as document categorization is the process of labelling a text document to one or more class labels from a finite set of predefined categories. It has been tackled in literature as either text-based, or ontology-based (semantic-based). Text-based classification relies simply on using tokens and keywords, whereas ontology-based classification relies on content meaning exploited through domain ontologies.

The very first step of classification process is the representation of a document from a full text version to a document as a vector of features using statistical vector space model. In the ontology-based classification, the feature vectors are composed of concepts (background knowledge) gathered from a domain ontology and concepts relevance represented by the frequency of concepts occurrences. The background knowledge gathered from ontology is incorporated in a document using the exact matching technique. This technique searches only concept labels that explicitly occur in a document [150].

Although the existing ontology-based approaches proved useful in classifying text documents, they however are limited due to two important issues. The first one is the limited capability of capturing and exploiting the whole conceptualization involved in user information and content meaning due to searching only the presence of concept labels. The second limitation is computing of the concept weight using only the concept relevance.

To address these limitation issues, we propose in this paper a new document classification model supervised by domain ontologies. The input to the proposed model is a set of

documents pre-classified by a domain expert. Next step is the representation of these documents as feature vectors. Background knowledge derived by domain ontologies is used to built feature vectors. The proposed model associates a document with a set of concepts by not only using the exact matching technique but also by identifying and extracting new terms which can be semantically related to the concepts of ontologies. To achieve this, we use SEMCON [69, 70], which deals with both, exact matching, and identification of new terms that are associated semantically with these concepts. SEMCON is an objective metric developed for enriching domain ontologies with new concepts by combining context and semantic of terms occurring in a corpus. To compute weight of concepts, the proposed model uses the concept weighting technique described in [68]. This weighting technique computes the weight of a concept using the relevance and the importance of the concept. Next, a classifier is trained and a predictive model is built by passing these feature vectors into one of the machine learning algorithms. Finally, an unlabelled document is classified into an appropriate category by using the predictive model built by the classifier.

The model is tested on a real ontology and a data set. We conducted several experiments using the baseline ontology, the improved concept vector space model using the new concept weighting technique (hereafter iCVS model), and the enriched ontology. We also run some experiments to test the model performance on different classification algorithms, namely Decision Tree, Naive Bayes, and Support Vector Machine. The results obtained by these experiments show that the classification performance is improved using our proposed model.

The rest of the paper is structured as the following. Section 10.2 presents state of the art in the field of document classification. Section 10.3 describes in detail the proposed ontology-based model for classification of text documents. The dataset and the domain ontology used to demonstrate the applicability of the proposed model are shown in Section 10.4, while the results and the analysis are given in Section 10.5. Conclusions of the paper and future work are given in Section 10.6

10.2 Related Work

Ontology-based document classification has become increasingly an important and an attractive research topic in many areas, especially in enrichment of document and category representation achieved by exploiting ontologies. An example of document classification which uses ontologies and relations between documents as a background knowledge to enrich the document representation is presented in [109]. A set of binary T×T matrices that contain all relations between terms such as hyponyms, hypernyms, hyponyms of hyponyms, etc., extracted from General Finish Ontology YSO is defined as the background knowledge. This way, the traditional bag of words classifier is extended with new relations utilizing the background knowledge.

Background knowledge derived by ontologies is extensively used for enriching documents with semantics from the biomedical domain. Such an example is presented by Camous et al. [21]. The authors use Medical Subject Headings (MeSH) ontology to enrich the existing MeSH based representation of documents with semantics. To identify and extract new concepts which are semantically close to the initial representation (concepts), a semantic similarity measure derived by simply counting edges (relations) between concepts in the MeSH hierarchy is used. A similar approach to Camous et al. is presented by Dinh and Tamine in [32]. The authors also rely on the MeSH ontology for enriching document representation but they use a different similarity measure to identify and extract the domain concepts. They use a content-based cosine similarity measure. Another similar approach where the MeSh ontology is used to enrich document representation is also the subject of the research in [143]. The authors developed an ontology-based system called OBIRS which is used for enriching documents with semantics from the domain of biomedicine.

Rather than using all ontology concepts, Fang et al. [41] proposed an ontology-based automatic classification method that uses only a small number of concepts. These concepts, which are primarily the lowest level concepts of an ontology, and the instances of the ontology obtained by ontology reasoning, are used to represent the document.

Our classification approach also uses background knowledge exploited by ontologies but it distinguishes from the approaches presented above in two aspects: 1) rather than using different semantic similarity measures to identify and extract new terms which can be semantically close to the existing ontology concepts, our approach uses SEMCON which integrates the semantic and contextual similarity measure, and 2) we provide a new concept weighting technique which besides concept relevance takes also into account the concept importance.

10.3 Proposed Model

In this section, we present the proposed ontology-based document classification model. The proposed model, illustrated in Figure 10.1, is a supervised learning model and its details are described in the following paragraphs.

The very first step of the model concerns with the predefined categories and the documents within these categories which are organized manually by an expert from that certain domain. The documents are simply represented as plain text and there is no semantics associated with them at this point.

The next step is to represent categories using a vector space representation model [75], where a concept vector is created per each category. This is a representation which is a step away from the keyword-based representation towards the semantic-based document representation. The semantic-based document representation is achieved by using the background knowledge constructed from domain ontologies. More precisely, the semantic of each document is embedded using the matching technique which relies on matching the terms in the document with the relevant concepts in the domain ontology. Term to concept mapping can be achieved by using the exact matching which searches only the concepts (concept labels) that explicitly occur in a document, and through identification of semantically associated terms. To accomplish term to concept mapping we have used the SEMCON model which deals with both, exact matching, and identification of new terms that are associated semantically with these concepts.

The exact matching performed by SEMCON is a straightforward process. There maybe single label concepts (*grant*, *funding*, etc.) in a domain ontology as well as compound label concepts ($project\_funding$), as shown in Figure 10.3. To acquire single label concepts, we use only those terms from the document for which an exact term exists in the domain ontology. For example, for concepts in the domain ontology such as *grant*, *funding*, etc., there exists the term that explicitly occurs in the document. While to identify and acquire compound label concepts, we use those terms from the document which are present as part of a concept in the domain ontology. For example, consider $project\_funding$ as one of the compound ontology concept. In this case, we map $project\_funding$ if either term, *project* or *funding* occurs in the document.

Identifying and extracting new semantically associated terms is a more complex process accomplished by SEMCON. Rather than simply searching for an ontology concept, it searches for new terms in documents that are associated semantically with the concepts of the ontology. To achieve this, SEMCON as an objective metric combines the contextual and the semantic information of the given terms through its learning process. Consequently, SEMCON primarily computes an observation matrix which is composed of three statistical features, namely term frequency, term font type, and term font size. Observation matrix serves as input to SEMCON to derive the context computed by using the cosine measure. In addition to the contextual information, SEMCON embeds the semantics by computing

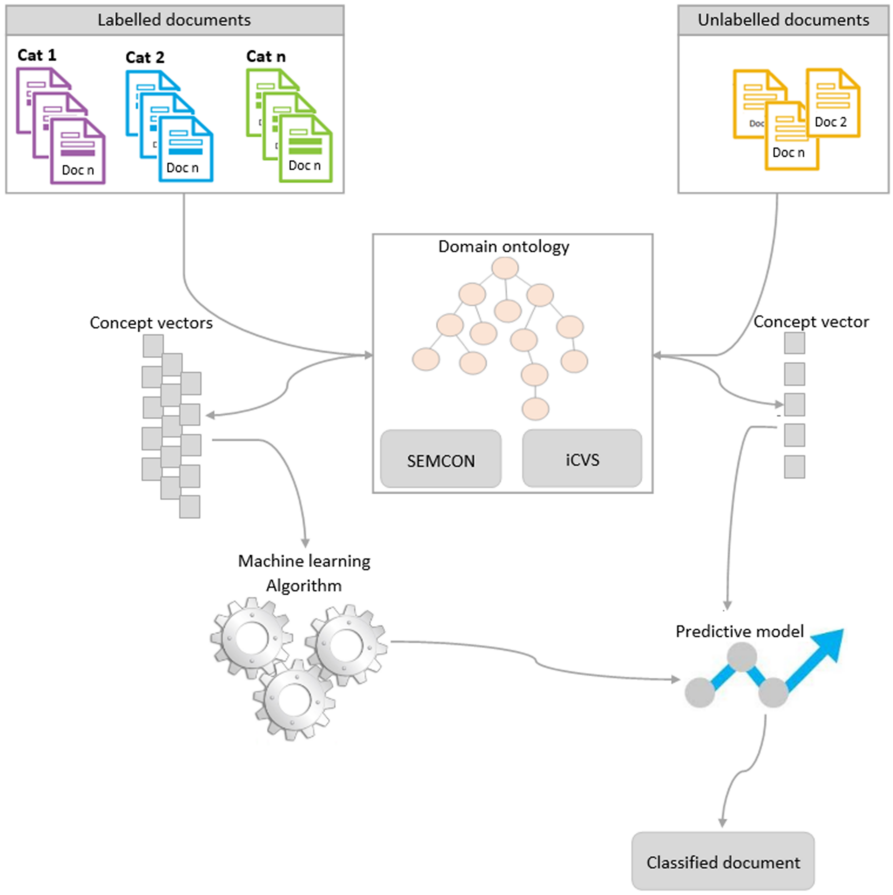


Figure 10.1: Proposed classification model

a WordNet based semantic similarity score between two terms - term that appears in the document and term that already exists in the ontology as a concept.

Once the semantics of documents are built, we then incorporate the category semantics built by aggregating the semantics of all documents which belong to the same category. The process of associating semantics to documents and to the category is shown in Figure 10.2. By doing this, the classification system can replicate the way an expert categorizes the documents into appropriate categories. This way, each category is represented as a vector composed of two components: concepts of the domain ontology and concepts weight. Concepts weight are typically computed using only concepts relevance as the classification criteria [31, 24] but this way of calculating weight of concepts is limited as it considers all concepts equally important regardless where the concepts are depicted in the hierarchy of ontology. However, the importance is not equal for all concepts and it depends on relations of concepts with other concepts in the ontology hierarchy. Concepts which have more relations with other concepts are more important than the concepts which have less relations. In order to take into account relationship between ontology concepts reflected by concept importance, we will use the weighting technique described in [68] to compute

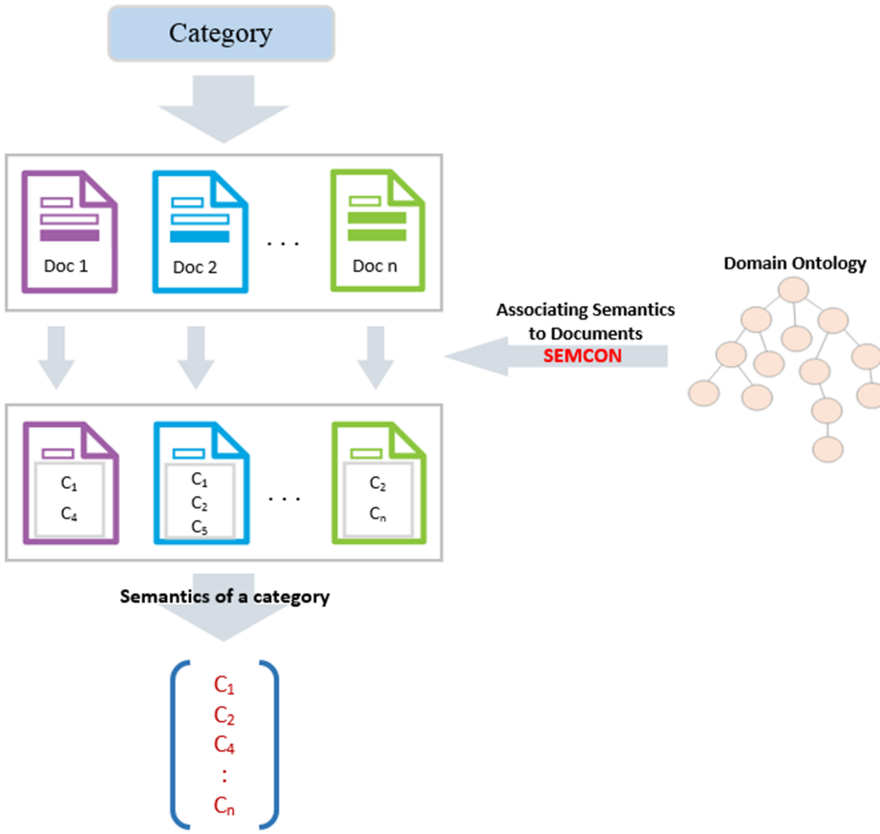


Figure 10.2: Associating semantics to documents and to the category

concepts weight. This technique defines the weight of a concept as the quantity of information given by the presence of that concept in a document and in an ontology hierarchy, and it is computed by the relevance of concept and the importance of concept. Relevance of a concept is simply defined by the frequency of occurrences of that concept in the document, whereas concept importance is defined by the number of relations a concept has to other concepts and it is computed by employing one of the Markov-based algorithms into the ontology graph. The formal definition of concept weight is given in Equation 10.1.

$$w(c) = Imp(c) \times Rel(c) \tag{10.1}$$

where, *Imp(c)* and *Rel(c)* denote the importance and relevance of a concept *c*, respectively.

After all these steps are performed, a category is finally represented by the tuple shown in Equation 10.2.

$$Cat = \{(c_1, w_1), (c_2, w_2), (c_3, w_3), ..., (c_i, w_i)\} \tag{10.2}$$

where c_i is the concept exploited by the domain ontology and w_i is its weight computed using Equation 10.1.

Next, different machine learning algorithms can be employed to train the classifier and to create a predictive model which can be used for classifying a new unlabelled document into an appropriate category.

The final step of the model deals with the corpus of new unlabelled documents. Each document from this corpus, which has to be classified is primarily represented as a concept vector. Such representation of the document is done by using the steps described above in this Section and it is defined by the same tuple given in Equation 10.2.

Lastly, the unlabelled document goes through the predictive model built by the machine learning algorithms and finally it is classified into the appropriate category.

10.4 Experimenting Procedures

This section describes the dataset and the domain ontology used to conduct the experiments in order to demonstrate the applicability of our proposed model and to validate its efficacy in terms of classification effectiveness.

10.4.1 Dataset

The dataset used in this paper is a real dataset consisting of 467 grant documents. These documents are classified into 5 categories by the field experts as part of the INFUSE [1] project.

Table 10.1: Dataset size

No	Category	# Train	# Test	Total
1	Culture	101	45	146
2	Health	69	36	105
3	Music	8	6	14
4	Society	124	41	165
5	Sportssociety	25	12	37
6	Total	327	140	467

The dataset is divided randomly, where 70% (327) of the documents are used to train the classifier and the remaining 30% (140) to test the performance of the classifier. The number of documents in each category varied widely, i.e Society category consists of 165 documents while Music category contains only 14 documents. Table 10.1 illustrates the 5 categories along with the number of training and testing documents in each category.

10.4.2 Domain Ontology

The ontology used for experimenting in this paper also comes from the funding domain. It is composed of 85 concepts and 18 object properties which link these concepts. Figure 10.3 presents part of the INFUSE ontology which consists of a subset of concepts and their relationships (*is-a*, *appliesFor*, *isReceivedBy*, etc.) from the funding domain.

We have to use a real ontology due to the fact that the ontology to be enriched and the dataset exploited for enriching the ontology with new terms have to be from the same domain.

10.5 Results and Analysis

Several experiments on document classification are conducted to demonstrate the applicability of our proposed model and to validate its effectiveness. INFUSE domain ontology is used as a baseline ontology and the standard information retrieval measures such as precision, recall and F1-measure are used to evaluate the performance of the document classification.

[1] http://infuse.scan4news.com/?cat=4

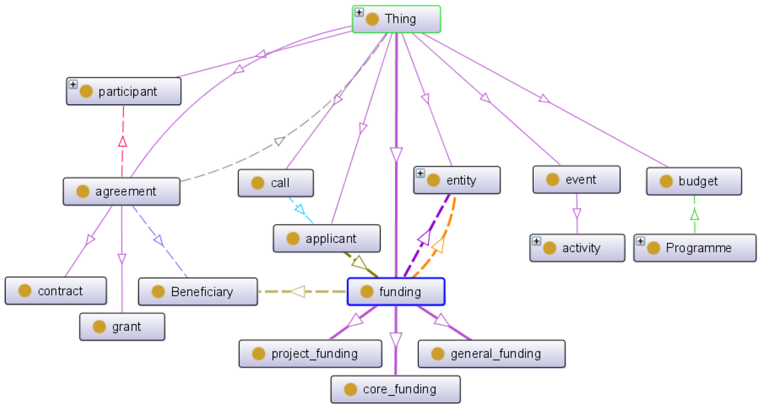


Figure 10.3: A part of the INFUSE ontology

The initial experiment conducted investigates the accuracy of classification by employing the improved concept vector space model (iCVS). Decision Tree is used as a classifier and the INFUSE dataset is used for training and testing the classifier. The classification result for each category and the weighted average results obtained using the baseline ontology, and the iCVS are shown in Table 10.2.

Table 10.2: The performance of classification using the baseline ontology and the iCVS

Concept	Baseline (%)			iCVS (%)		
	P	R	F1	P	R	F1
Culture	63.8	66.7	65.2	64.8	77.8	70.7
Health	81.8	75.0	78.3	80.0	77.8	78.9
Music	50.0	16.7	25.0	50.0	16.7	25.0
Society	62.0	75.6	68.1	69.8	73.2	71.4
Sportssociety	50.0	33.3	40.0	66.7	33.3	44.4
Weighted avg.	**66.1**	**66.4**	**65.5**	**69.1**	**70.0**	**68.8**

The next experiment performed under the same conditions, in terms of dataset (INFUSE) and classifier (Decision tree), relies on the baseline ontology whose concepts are primarily being enriched with new terms. In this case, we have used the Top-5 terms to enrich ontology concepts and classification result for each category and the weighted average results obtained using the baseline ontology, and the enriched ontology are given in Table 10.3.

Table 10.3: The performance of classification using the baseline and the enriched ontology

Concept	Baseline (%)			Enriched (%)		
	P	R	F1	P	R	F1
Culture	63.8	66.7	65.2	82.5	73.3	77.6
Health	81.8	75.0	78.3	83.3	83.3	83.3
Music	50.0	16.7	25.0	100	16.7	28.6
Society	62.0	75.6	68.1	71.1	92.7	80.9
Sportssociety	50.0	33.3	40.0	70.0	58.3	63.6
Weighted avg.	**66.1**	**66.4**	**65.5**	**79.2**	**77.9**	**76.7**

Table 10.4: The Top-5 terms obtained by model using the First sense heuristic technique

Concept	Top-5 terms: overall score
financing	finance:0.88, investment:0.78, participation:0.74, implementation:0.73, compliance:0.73
fund	budget:0.81, amount:0.81, funding:0.80, finance:0.76, share:0.74
funding	fund:0.80, amount:0.78, part:0.74, subsistence:0.72, grant:0.71
grant	provide:0.79, welfare:0.77, partner:0.75, cost:0.73, commission:0.72
programme	rule:0.44, relevance:0.44, framework:0.44, aspect:0.44, implementation:0.43
subsidy	subvention:0.61, grant:0.57, welfare:0.45, loss:0.45, scholarship:0.43

Table 10.5: The Top-5 terms obtained by model using the Maximizing semantic similarity technique

Concept	Top-5 terms: overall score
financing	finance:0.89, funding:0.88, field:0.82, contribution:0.82, investment:0.79
fund	provision:0.84, issue:0.84, protection:0.83, part:0.83, budget:0.81
funding	support:0.95, financing:0.88, part:0.84, field:0.84, contribution:0.83
grant	development:0.95, verification:0.95, section:0.94, article:0.93, agreement:0.92
programme	framework:0.44, aspect:0.44, rule:0.43, organisation:0.43, implementation:0.43
subsidy	subvention:0.61, grant:0.57, aid:0.52, present:0.47, award:0.47

The results given in Table 10.2 and Table 10.3 show an improvement of weighted average classification accuracy when classification is performed using the iCVS, and the enriched ontology, respectively. The improvement can be reflected by the F1 measure which is increased from 65.5% to 68.8% when the iCVS is being employed. This improvement of classification effectiveness is achieved due to the fact that iCVS through its concept weighting technique of aggregating concept importance and concept relevance provides concepts with weights that have better discriminative power in terms of classification. Moreover, a substantial improvement of classification performance is achieved using the enriched ontology comparing to the classification using the baseline ontology. This improvement is reflected with an increase of F1 measure from 65.5% using the baseline, to 76.7% using the enriched ontology. Consequently, the improvement is observed for almost all categories. For example, the Music category has achieved a 100.0% precision using the enriched ontology, compared to a precision of 50.0% achieved using the baseline ontology, and the iCVS, respectively.

Next, we have investigated the impact of the word sense disambiguation technique on the quality of ontology enrichment. To achieve this we experimented with two techniques, namely First sense heuristic and Maximizing semantic similarity. First sense heuristic also called the predominant sense is a technique used to find the correct meanings of a term. It relies on distribution property of word senses and assumes that the correct meaning of a word is represented by its most commonly used sense. Maximizing semantic similarity is also a word sense disambiguation technique which assumes that the correct meaning of a term is the one which maximizes the relatedness between the term and a sense among all possible senses (meanings). We employed these techniques in word sense disambiguation in order to observe the terms which can be obtained as relevant by the model for enriching concepts of the ontology. A part of this observation is summarised in Table 10.4 and Table

10.5 which show the Top-5 terms obtained by the model along with their overall scores using First sense heuristic and Maximizing semantic similarity technique, respectively.

As can be seen from Table 10.4 and Table 10.5, First sense heuristic and Maximizing semantic similarity techniques yield different results in terms of finding the relevant terms for enriching a particular concept. For example, *fund*, *amount*, *part*, *subsistence*, and *grant*, are the top five terms retrieved by the model using the First sense heuristic for enriching the concept *funding*, whereas, *support*, *financing*, *part*, *field*, and *contribution*, are the top five terms retrieved by the model using the Maximizing semantic similarity technique. Besides the difference on the terms retrieved, these techniques also differ in the overall score assigned to the obtained terms, i.e., *finance:0.88*, and *finance:0.89*. These differences resulted due to these two word sense disambiguation techniques ultimately yield different classification performances. An observation of accuracy of Decision tree classifier using First sense heuristic and Maximizing semantic similarity technique is shown in Table 10.6.

Table 10.6: The performance of Decision tree classifier using First sense heuristic and Maximizing semantic similarity technique

Concept	First Sense (%)			Max Similarity (%)		
	P	R	F1	P	R	F1
Culture	72.0	80.0	75.8	82.5	73.3	77.6
Health	87.5	77.8	82.4	83.3	83.3	83.3
Music	33.3	16.7	22.2	100	16.7	28.6
Society	83.7	87.8	85.7	71.1	92.7	80.9
Sportssociety	58.3	58.3	58.3	70.0	58.3	63.6
Weighted avg.	**76.6**	**77.1**	**76.6**	**79.2**	**77.9**	**76.7**

Furthermore, we used another evaluation criteria where we investigated into how much the number of terms used to enrich concepts of an ontology affects the classification accuracy. We achieved this using Top-N terms obtained by the model using both the two word sense disambiguation techniques. In our case we set the values of N from 1 to 5, where Top-1 means the very top first term, Top-2 means the top 2 terms and so on. The observations, in terms of F1 measures, of the classification experiment run for 5 different values of N and on 3 different classifiers using First sense heuristic technique and Maximizing semantic similarity technique are illustrated in Figure 10.4 and Figure 10.5, respectively.

It can be seen from the chart diagram shown in Figure 10.4 that the best result in terms of F1 measure is obtained when 2 terms are used for enriching one particular concept of the baseline ontology (Naive Bayes and SVM). The performance of these two classifiers start declining by increasing the number of new terms. This is in contrast to the Decision tree classifier's performance, which improves by increasing the number of terms used to enriching given concepts. These results suggest that up to 2 terms can be used to enrich one particular concept of an ontology in order to achieve the best performance on classification process using Naive Bayes and SVM classifiers. Increasing the number of terms to enrich a given concept above this threshold (2 terms) yields the same performance or even lower. On the contrary, the Decision tree classifier achieves the best performance when 4 terms are used to enrich a given concept.

On the contrary, Figure 10.5 shows that the classification accuracy increases by increasing the number of terms used for enriching ontology concepts and this is shown by all the three classifiers which have achieved the best performance when the model have used the top five terms. This may happen due to the fact that the overall score of top terms retrieved by model using the First sense heuristic drops quickly after the top 2 terms while it drops slowly when the model uses the Maximizing semantic similarity technique. This suggests that the top 2 terms are relevant for enriching the ontology concepts and give contribution on the classification accuracy of the model which is based on the First sense heuristic

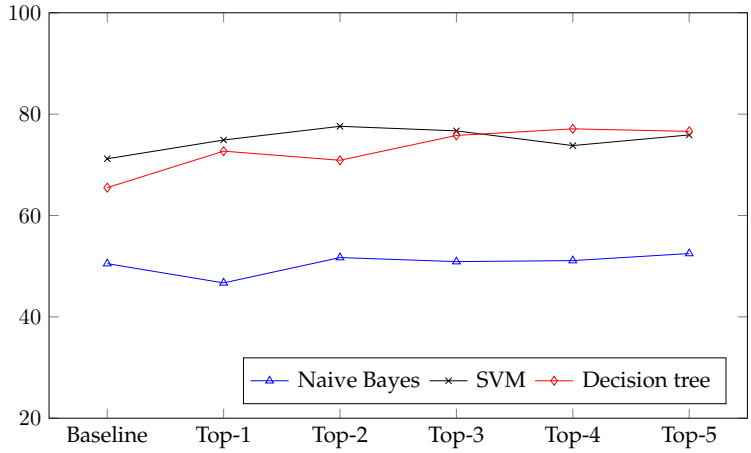


Figure 10.4: F1 measures obtained by 3 different classifiers using First sense heuristic technique

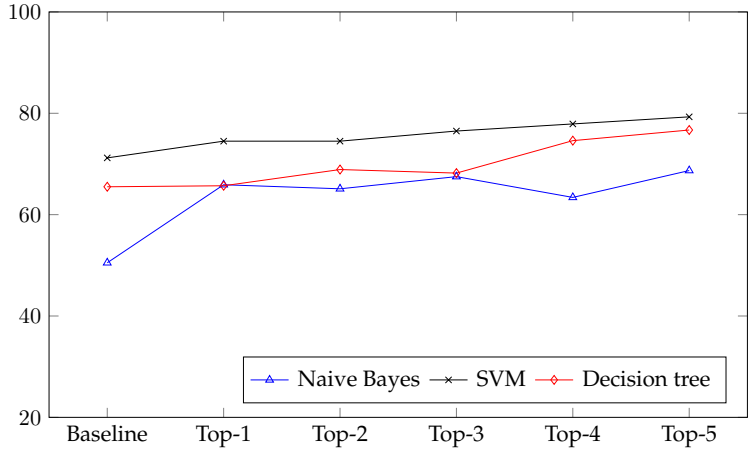


Figure 10.5: F1 measures obtained by 3 different classifiers using Maximizing semantic similarity technique

and the top 5 terms are relevant for the model which relies on the Maximizing semantic similarity technique in word sense disambiguation.

It is also interesting to note from the graph shown in Figure 10.4 and Figure 10.5 that the Decision tree and the Naive Bayes classifier achieve completely opposite results when the number of terms used to enrich the baseline ontology concepts is increased. This happened due to the fact that the Decision tree through information gain figure out attributes which have the highest information gain values. These attributes are the most homogeneous terms and they have great impact on the classification performance. Naive Bayes classifier assumes that the value of a particular term is independent of the value of any other term which means that it considers each of these new terms to contribute indepen-

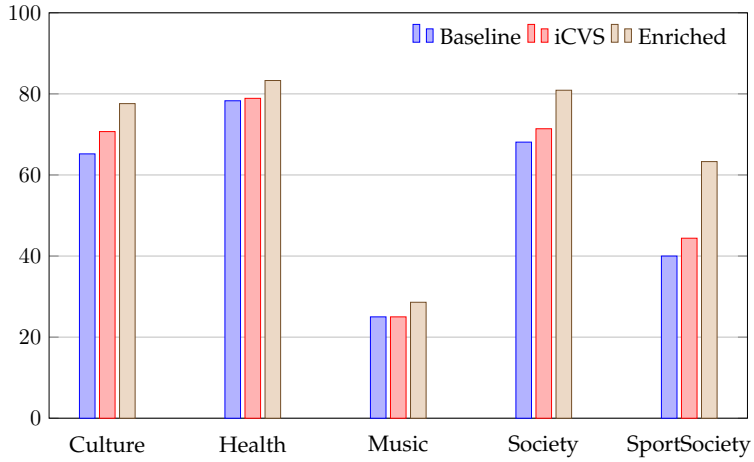


Figure 10.6: F1 measure obtained by Decision tree classifier using the baseline ontology, the iCVS, and the enriched ontology

dently to classification regardless of any possible correlations between these terms.

Lastly, a comparison of results (F1 measure) for each category of the INFUSE dataset achieved by the Decision tree classifier using the baseline ontology, the iCVS, and the enriched ontology, is illustrated in Figure 10.6. It can be seen from the chart diagram that a considerable improvement, from 65.5% to 68.8%, is achieved by employing iCVS, and a substantial one, from 65.5% to 76.7%, is achieved by employing the enriched ontology.

10.6 Conclusion and Future Work

In this paper, we presented an ontology-based classification model for classifying text documents from the funding domain. The model relies on background knowledge exploited by ontologies for document representation. A document is represented by a set of ontology concepts derived by using the exact matching technique, and by identifying new terms which can be semantically related to these concepts. This representation is achieved using the SEMCON. The weight of ontology concepts is computed using a new concept weighting technique which is composed of two components, the relevance and the importance of the concept.

The proposed classification model is tested on different classifiers using a real ontology and the results obtained by the experiments showed that the accuracy of the document classification is improved. In addition to this, we investigated the impact of the word sense disambiguation techniques on the accuracy of classifiers and we found that Naive Byes classifier is the most affected one performing differently in different techniques, while SVM and Decsion tree classifiers are less affected. The F1 measure of 52.5% is achieved by Naive Bayes classifier using First sense heuristic, comparing to a 68.7% of F1 measure achieved using Maximizing semantic similarity technique.

In future work we plan to test our model on larger datasets and ontologies from different domains and to evaluate and compare the obtained results of classification.

Part V

Conclusions

Chapter 11

Conclusions

This chapter concludes our work. Section 11.1 provides a summary of results achieved in this thesis, while in Section 11.2, we discuss the future work.

11.1 Summary of Findings

As the volume of information available on the Internet is growing rapidly, we believe that the document classification systems based on ontologies will continue to play a vital role on organizing and classifying in a semantic way this huge amount of information accordingly. In this thesis, we examined several aspects about ontology-based document classification with an emphasis on ontology enrichment and concept weighting scheme, and we developed some approaches that would contribute to improving the performance of these ontology-based systems. A summary of our findings is given below.

- We explored and developed an automatic ontology enrichment model which relies on semantic and contextual information of terms appearing within a document. Contextual information of a term is defined as the part of text, i.e. passage, which surrounds that term and it is determined using the cosine measure between feature vectors. The semantic information on the other hand is determined through a semantic similarity score based on the English lexical database WordNet. The output of the model is a list of relevant terms to enriching ontology concepts along with their final scores obtained by aggregating contextual and semantic information. We conducted experiments to examine the contribution of each of the components, contextual and semantic. Based on the empirical analysis, it is suggested to use a balanced weight between these two components.

- The definition of contextual information is expanded by introducing for the first time in the research two new statistical features of a term, namely, term font type, and term font size. Various values set to these features for a particular term are used in addition to the frequency of appearances of that term in the corresponding document, in order to constitute the feature vectors. Experimental results showed that the performance of our proposed ontology enrichment model is improved when statistical features are used to derive the context.

- We have validated our ontology enrichment SEMCON model by using subjective and objective experiments. Results obtained from the subjective experiments conducted through publishing an online questionnaire to subjects were used as a ground truth for validating the results achieved by our proposed SEMCON model. Our results were also validated against three objective models, namely, *tf*idf*, χ^2, and *LSA*, showing a considerable improvement achieved by our SEMCON model.

- A new concept weighting scheme constituted by concept relevance and concept importance is proposed and developed in order to improve the concept vectors. We also proposed an automatic approach based on Markov model to compute the importance of concepts of an ontology. To achieve this, the approach converts an existing ontology into a direct acyclic graph in which a concept is mapped into a vertex and a rela-

tion into an edge and it employs then the PageRank from the family of Markov-based algorithms to compute the importance of concepts.

- We demonstrated that the effectiveness of the document classification systems can be improved by proposing an approach that uses background knowledge derived by an ontology. We initially enriched an existing ontology with new concepts through the ontology enrichment model described above, and a new weighting scheme composed of concept relevance and importance is then used to assess the weight of concepts of this ontology. According to the results of the experiments conducted on three different classifiers, the proposed classification approach has achieved a considerable improvement on the tested datasets. In addition, we also examined the impact of the disambiguation techniques on the performance of classifiers and we found that some classifiers are more affected in terms of precision then the others by performing differently in different techniques.

11.2 Future Work

Ontology-based document classification has become increasingly important and attractive research topic in many areas where continuously new challenges emerge and which require further study in order to achieve the desirable performance. The future work of the ontology-based document classification approaches is discussed in the following.

- **Use multiple media modalities to widen the coverage of knowledge resources**

  Ontology enrichment is a process constituted of three steps and identification and extraction of the relevant terminology such as synonym terms or linguistic variants is one among these steps which can be achieved by exploring the input data resources. Our ontology enrichment model explores only textual data that basically can be either in structured, semi-structured, or unstructured format (see Section A.2.1 in Appendix). However, there is more than textual data resources which can be explored by our proposed model for identification and extraction of the relevant terminology providing more advanced enrichment capabilities to the model. In this respect one possible direction to work on in the future is extending our proposed enrichment SEMCON model to support the identification and acquisition of the relevant terminology for enriching an ontology from multiple and diverse modalities including text, images, video, etc.

- **Exploring other statistical features and models for deriving the context**

  Introduction of two new statistical features for deriving the context of a term proved to be useful in terms of improving the performance of the ontology enrichment model. A linear model is adopted for different font sizes and font types and various values are set to these features accordingly. It might be worth to exploit other statistical features of terms which would contribute on computation of the contextual information in particular and improving the effectiveness of the enrichment model in general. In addition, other automatic models and techniques for assessing the weight of font types and font sizes can be studied and we believe that machine learning techniques would play an important role in this regard by providing empirically evaluated weights to various features.

- **Disambiguation techniques for untagged corpus**

  The disambiguation techniques used to find the correct sense (meaning) of terms yielded better performance in terms of ontology enrichment. Two disambiguation techniques, namely, First sense heuristic, and Maximizing semantic similarity, were employed and we observed that these techniques produced different results in terms

of finding the relevant terminology for enriching a particular concept. These techniques rely on the lexical database WordNet, in which, senses are ordered based on the frequency distribution in the manually tagged resource such as SemCor. However, the frequency distribution of the senses of terms, especially topical terms, is more related to the genre and the domain of the discourse [96]. In this regard, one possible direction for future work would be exploring of word sense disambiguation techniques that are based on untagged corpus data to find the correct sense of terms. By analysing the untagged corpus, the disambiguation techniques can be adjusted to the domain and the genre of the discourse and we believe that this analysis can improve the performance of the proposed ontology enrichment model.

- **Exploring other algorithms to compute concept importance**

  In chapter 7 we came up with the idea of considering the concept importance as a very important part for evaluation of the concept's weight and we implemented this idea by developing a model that was capable to compute automatically the concept importance. The developed model employs a Markov-based algorithm on a converted ontology graph to compute the importance. Particularly, we employed PageRank from the family of Markov-based algorithms. It would be interesting to explore in the future other algorithms which apart from the hyperlink relations considered by the PageRank, take into account other edge (relation) types, i.e. *property-of, subclass, etc* that are present in an ontology graph. The edge types play an important role on determining the importance of concepts and we think that using the algorithms which address this issue (link-analysis ranking algorithms [155], e.g. ObjectRank) will improve the model with the capabilities to compute the importance of concepts in a more effective way.

- **Further evaluation of the document classification approach**

  The proposed ontology-based document classification approach demonstrated to be useful on classifying documents into appropriate categories but still requires a further evaluation. It was evaluated on a real dataset composed of 467 documents, and an ontology consisting of 85 concepts and 18 objects, both coming from the funding domain. We were restricted to use the real dataset and real ontology due to the fact that the existing ontology that we want to enrich, and the dataset explored for enriching it, have to be from the same domain. To the best of our knowledge there is no public dataset containing the documents and the ontology from the same domain. In order to thoroughly evaluate and validate the efficacy of the approach in terms of classification effectiveness, we consider that the test dataset needs to be expanded to cover more documents and a bigger ontology. In addition, the proposed approach is tested on three machine learning based classifiers, namely, Decision Tree, Support Vector Machine, and Naive Bayes, but it might be worth to study in the future the employment of deep learning techniques such as word embedding to evaluate the approach and compare the obtained results with the results achieved from the classifiers used in this research work.

Part VI

Appendix

A8: A Hybrid Concept Learning Approach to Ontology Enrichment

Publication details
Kastrati, Z., Imran, A., and Yayilgan, S., "*A Hybrid Concept Learning Approach to Ontology Enrichment*", IGI Global 2017, ch. Innovations, Developments, and Applications of Semantic Web and Information Systems.

Appendix A

A Hybrid Concept Learning Approach to Ontology Enrichment

Abstract

The wide use of ontology in different applications has resulted in a plethora of automatic approaches for population and enrichment of an ontology. Ontology enrichment is an iterative process where the existing ontology is continuously updated with new concepts. A key aspect in ontology enrichment process is the concept learning approach. A learning approach can be a linguistic-based, statistical-based, or hybrid-based that employs both linguistic as well as statistical-based learning approaches.

This chapter presents a concept enrichment model that combines contextual and semantic information of terms. The proposed model called SEMCON employs a hybrid concept learning approach utilizing functionalities from statistical and linguistic ontology learning techniques. The model introduced for the first time two statistical features that have shown to improve the overall score ranking of highly relevant terms for concept enrichment.

The chapter also gives some recommendations and possible future research directions based on the discussion in following sections.

A.1 Introduction

Domain ontologies are a good starting point to model in a formal way the basic vocabulary of a given domain. They provide a broad coverage of concepts and their relationships within a domain. However, in-depth coverage of concepts is often not available, thereby limiting their use in specialized subdomain applications. It is also the business dynamics and changes in the operating environment which require modification to an ontology [97]. Therefore, the techniques for modifying ontologies, i.e. ontology enrichment, have emerged as an essential prerequisite for ontology-based applications.

An ontology can be enriched with lexical data either by populating the ontology with lexical entries or by adding terms to ontology concepts. The former means updating the existing ontology with new concepts along with their ontological relations and types. This increases the size of the existing ontology which requires more computational resources and more time to compute. Thus, making it less cost effective. The latter means adding new concepts without taking into account the ontological relations and types between concepts. Because of this, the ontology structure will remain the same but its concepts will be enriched with their synonym terms or linguistic variants.

Enrichment of ontology concepts is aiming at improving an existing ontology with new concepts. It is part of the iterative ontology engineering process [40]. The core of this process is the learning approach which constitute tasks such as identification and acquisition of the relevant terminology through exploring various knowledge resources, and the creation of the concepts.

There is a variety of concept learning approaches that are available to enrich concepts of an ontology. These approaches rely on either linguistic, statistical, or hybrid techniques [36, 59]. Although, these approaches proved useful for enriching ontologies of many domains, they do have some limitations, especially when it comes to semantic information

of terms. The existing approaches use only contextual information without considering the semantic information of terms. Moreover, the contextual information is simply derived by distributional property of terms such as term frequency *tf* or term frequency inverse document frequency *tf*idf*, and co-occurrences of terms.

The focus of this chapter is to enlighten the reader with the ontology concept enrichment process, explore state-of-the-art methods and techniques in this regard, review input data resources, learning approaches and systems build upon them, discuss their limitations and to propose solutions and to give some recommendations accordingly. It also describes the SEMCON model to enriching the domain ontology with new concepts by combining contextual as well as the semantics of terms.

SEMCON uses unstructured data as input for ontology learning process and is composed of two parts - contextual and semantic. Context is defined as the part of a text or statement passage that surrounds a given term and it determines term meaning. In this work, it is the cosine distance between the feature vectors of any two terms. The feature vectors are composed of values computed by both the frequency of occurrence of terms in corresponding passages, and the statistical features such as font type and font size. The semantics on the other hand is defined by computing a semantic similarity score using lexical database WordNet.

Additionally, this chapter investigates into how much each of contextual and semantic components contributes to the overall task of enriching the domain ontology concepts. Obtained results are compared with *tf*idf*, χ^2 and *LSA*. Results for several domains including Computer, Software Engineering, C++ Programming, Database, and the Internet are presented in this chapter.

The rest of the chapter is organized as follows. Section A.2 presents ontology enrichment pipeline, describes various input data modalities, discusses text-based resources followed by concept learning techniques and applications using them. This section also presents the state-of-the-art systems in the field of ontology enrichment. Section A.3 describes the proposed SEMCON model in detail. In section A.4, we describe the experiments including subjective and objective evaluation of SEMCON along with measures used to evaluate the effectiveness of objective methods. Results obtained by SEMCON and other objective methods and their comparisons are shown in section A.5. Section A.6 highlights some key points for ontology enrichment followed by a future research directions given in section A.7. Section A.8 presents two important fields where SEMCON has successfully been employed. Lastly, section A.9 concludes the paper.

A.2 Background

This section describes the fundamentals of building an ontology concept enrichment model as shown in Figure A.1. Enrichment of ontology concepts aims to improve a given ontology by populating it with new concepts. As part of an ontology engineering process, it involves subtasks from only lower part of ontology learning layer cake model [25]. Acquisition of the relevant terminology, identification of synonym terms or linguistic variants, and the creation of concepts are the subtasks involved. To accomplish these subtasks, the enrichment process departs from an initial ontology that will be enriched with new concepts. In a simplified view, this initial ontology is constituted by a set of concepts and relations that link these concepts. The next step is the identification and acquiring of the relevant terminology such as synonym terms or linguistic variants. This is achieved by exploring the knowledge input data resources which can be in structured, semi-structured, or unstructured format. Finally, a concept learning approach, which is the core of the entire enrichment process, is employed to the extracted terminology in order to create new concepts for populating the initial ontology.

A vast number of ontology enrichment models are available which rely on a variety of knowledge resources. These resources are primarily used to identify and extract rele-

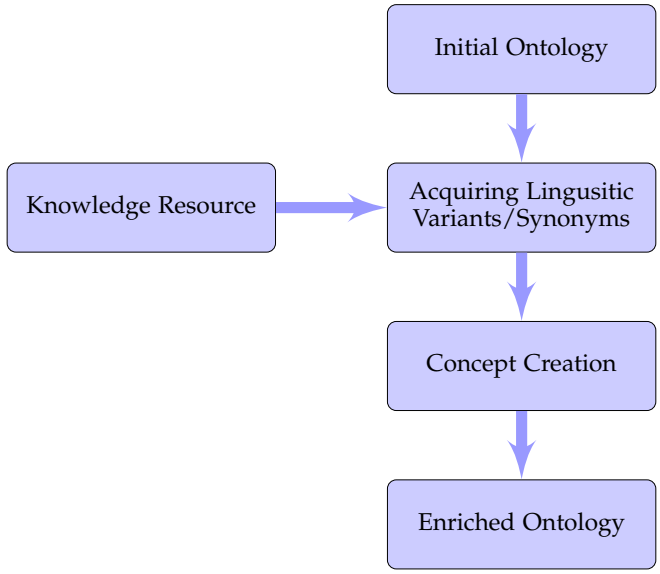


Figure A.1: The ontology concept enrichment process

vant terminology, and their linguistic variants and synonyms. Various concept learning approaches have also been used in literature to enrich an ontology by populating it with new concepts. Therefore, one way of categorizing the enrichment models is by looking at the approach it takes for ontology enrichment and the kind of data resources it uses. The main aim of this section is to provide an overview of the existing models in light of the learning approach and knowledge resources (input data) it uses, and shed some light on similarities and differences between these approaches.

A.2.1 Modality of Input Resources

Ontology enrichment systems identify and extract their knowledge of interest from the input resources. The input resource encompasses multimedia modalities including but not limited to textual data, audio, images, and videos. The primary focus of this chapter is the text-based resources which can broadly be categorized into structured resources, semi-structured resources, and unstructured resources. These are explained in following subsections.

A.2.1.1 Structured Resources

The ontology enrichment approaches based on structured input data uses database schemata, existing ontologies, linked data, and lexical semantic databases (e.g. WordNet) to acquire the relevant concepts to enriching an ontology.

Acquiring ontological knowledge using databases is done through the conversion of relational elements into ontological ones. This conversion is achieved using the corresponding E/R model and a set of basic translation rules. These rules enable to identify and extract concepts of the ontology, particularly, they describe what entities and relationships of the E/R model can be modeled as a concept in the ontology. Another resource of structured information to be exploited is linked data. Linked data, in contrast to databases where the data schema is defined formally, have no explicit schema for their dataset due to the fact the

publishers of the linked data are more focused on publishing data first rather than creating the schema. Therefore, exploiting of linked data for ontology enriching refers to the process of detecting meaningful patterns in RDF graphs. This can be achieved using statistical analysis where frequent patterns and correlations between these patterns in large data sets are evaluated.

Other approaches utilizing structured data for ontology enrichment rely on the adaptation of existing ontologies to new domains. This adaptation concerns with re-using existing schematic structures of the ontology.

A.2.1.2 Semi-structured Resources

The semi-structured resources consist of some structured and unstructured information given by the markups. The category of approaches which uses semi-structured input data utilizes existing markup structures within the textual data to perform the process of enriching ontologies. Web document structures such as Hypertext Markup Language (HTML) or its extension, Extensible Hypertext Markup Language (XHTML), Extensible Markup Language (XML), or Document Type Definition (DTD's) have been exploited.

The semi-structured goes beyond plain text. It represents an added value created by many Web authors which are worth to be exploited. The steps are as following:

1. The first and foremost step of using semi-structured data is the conversion of web document collection into XHTML web document collection.

2. Next step is the use of web document markups such as text spans. Text occurring within the text span pairs are used for this purpose.

3. Next step is the cleaning process. Whitespaces appear in the text span list therefore some cleaning step are needed to remove these whitespaces. Additionally, other cleaning steps such as eliminating punctuation and numbers, and converting all characters to lower case, can be performed.

4. The final step is the frequency analysis where the frequency of occurrences of the text span is computed. This results in a list of candidate terms ranked by their frequency within the Web document collection. Terms above a specified threshold are considered relevant terms for ontology enrichment.

Another approach which exploits semi-structured data to acquire the terminology for ontology enrichment is presented by Kruschwitz [79]. Kruschwitz initially pre-processes the web document to extract only the text associated to a set of markups such as <*meta*>, <*head*>, <*title*> or emphasizing tags such as <*b*> or <*i*>. Next, the importance of the exploited terms is computed using frequency analysis such as term frequency or term frequency using context where context is defined using co-occurrences of terms within the same unit or block structure, i. e. *title*, *keywords*, *meta*, *headers*, etc. [95].

A.2.1.3 Unstructured Resources

Unstructured data, known as free text, is the most difficult input resource to extract the relevant knowledge for enriching ontologies. Approaches that utilize this input data are dependent on natural language processing. They use the interacting constraints on the various language levels to discover and extract concepts and their relationships. Moreover, Hazman et al. [59] showed from the survey performed that Natural Language Processing - NLP is the most common among all the techniques. Hence, they classified all these approaches based on the technique used in addition to NLP. Additionally, they identified three major classes of approaches. The first group of approaches integrates NLP with the statistical techniques. These approaches extract concepts using a shallow parser for identification of noun and noun phrases and frequency of occurrences of these noun and

noun phrases. The second category employs pure NLP technique using syntactical dependency and parsers to discover concepts and their relationships. The third category of approaches integrates techniques from different disciplines such as information retrieval, lexical databases, and machine learning, in addition to computational linguistics.

A.2.2 Concept Learning Technique

The next processing step of ontology enrichment is the acquisition of the terminology and their linguistic variants and synonyms from the knowledge resources. This is carried out via concept learning techniques. There are various concept learning techniques employed by different ontology enrichment approaches which generally can be classified into three major categories: 1) linguistic, 2) statistical, and 3) hybrid.

The linguistic approach also known as symbolic relies on linguistic components, e.g. noun phrases, to identify and acquire relevant concepts for enriching the ontology. The most common linguistic approach is the one which uses NLP technique of lexico-syntactic pattern analysis. Hearst [60] was the first who introduced and explored some lexico-syntactic patterns in the form of regular expressions to extract ontological knowledge from English texts. The list of Hearst's lexico-syntactic patterns is shown in Table A.1.

Table A.1: Hearst's lexico-syntactic patterns

No	Lexico-syntactic patterns
1	NP_H such as {NP, }* {(or|and)} NP
2	such NP_H as {NP, }* {(or|and)} NP
3	NP{, NP}* {,} (and|or) or other NP_H
4	NP{, NP}* {,} (and|or) and other NP_H
5	NP_H {,} including {NP,}* {(or|and)} NP
6	NP_H {,} especially {NP,}* {(or|and)} NP

The Hearst's patterns proved to be successful at identifying and extracting a set of relationships, i.e. hypernym, but this technique of ontology learning is tedious and limited to a small number of patterns. To address this limitation, a machine learning technique has emerged. It tends to replace manually-created patterns with an automatic one and to achieve this it primarily uses a set of known hypernym pairs to automatically identify large numbers of useful lexico-syntactic patterns. More concretely, noun pairs from corpora are collected and a set of hypernym pairs using WordNet is obtained. Next step is collection of sentences in which nouns pairs occur. These sentences are parsed and patterns are extracted automatically from the parsed tree. Finally, a classifier is trained based on these patterns.

Other linguistic approaches rely on the syntactic dependencies analysis. Such approaches follow the idea that syntactic dependencies provide information on the semantic relations between the concepts. Dependencies are found out via a process composed of two phases. In the first phase, the corpus is tagged by a part of speech tagger, while in the second phase the tagged corpus is analyzed in sequences of basic chunks where two consecutive chunks represent a syntactic dependency.

There is another linguistic approach that uses syntactic analysis but with the focus being placed on the syntactic structure of component terms. This approach assumes that a compound/multi-world term, such as *prostate cancer*, is more specific than a single compositional term, i.e. *cancer*, and therefore, it is very likely that a compound term to be a hyponym of a single term.

While linguistic approaches rely on NLP analysis techniques to extract concepts from input data, the statistical approaches rely on the frequency analysis of terms. To identify and extract the relevant knowledge for ontology enrichment, these approaches utilize large

corpus of textual data for calculating a distributional property of terms such as term frequency - *tf* or/and term frequency inverse document frequency - *tf*idf*.

Other statistical approaches are concerned with batches of terms. These approaches are based on the assumption that identification and extraction of ontological terminology relies not only on the meaning of terms, but also on the basis of their co-occurrences with other terms and the frequencies of the co-occurrences [89]. Term co-occurrences, also referred to as collocation, defines the context within a discourse which can be either a sentence, paragraph, or an entire document [63]. A major advantage of these approaches is that they require no prior knowledge of the dataset and their ability to be generalized to other domains. This advantage makes these approaches the most addressed techniques among the statistical concept learning approaches. However, a disadvantage of these techniques is the need of a large corpus of textual data in order to be able to identify and obtain the relevant terminology to enrich ontologies.

Even though both symbolic and statistical approaches have proved useful as concept learning technique for ontology enrichment, they however have some limitations. For example, statistical approaches provide better coverage than symbolic approaches but their results are only probabilities without a conceptual explanation. As a result, a hybrid approach which combines the statistical and the symbolic approaches is introduced. The hybrid approach employs the benefits of both approaches and eliminates their limitations.

A.2.3 Ontology Enrichment Application

This subsection present systems based on the concept learning approaches described in the previous subsection for enriching ontologies. It starts by listing the systems which use linguistic approaches employed as concept learning, continuing with the statistical one, and finalizing with the systems which employ hybrid approaches as concept learning.

A.2.3.1 Linguistic Approaches

SynDiKATe [55] is an ontology enrichment application which relies on natural language processing analysis. Technical documents in the German language taken from test reports from the information technology domain and medical finding reports are exploited and modelled into a directed graph. The syntactic dependency (sentence level and text level) is then computed using the graph dependency of nodes and edges. The nodes represent terms occurring in documents and edges denote relations between these terms.

medSynDiKATe [56] is an extension of SynDiKATe application. It is designed to automatically acquire medical knowledge from medical finding reports. Emphasis was put on the role of various input textual resources required for text understanding with a focus being placed on grammar and domain knowledge. Additionally, a focus is put on alternative ways to support knowledge acquisition to foster the scalability of the system. Two concept learning approaches, automatic and semi-automatic, are employed and fully embedded in the text understanding process.

HASTI [135] is an ontology enrichment application which uses Persian free text as an input. It utilizes a combination of morpho-syntactic and semantic analysis. The enrichment process departs from a seed ontology whose lexicon is nearly empty at the beginning. The new obtained concepts are then inserted on top of the existing ontology.

KnowItAll [38, 38] is another system which utilizes natural language processing to identify and acquire the information. It is a domain-independent system. It explores the Web by employing lexico-syntactic patterns analysis to discover relevant information for enriching ontologies. Relevant concepts are selected by computing a version of pointwise mutual information measure called concept plausibility.

A.2.3.2 Statistical Approaches

DOODLE II [158] is an example which uses the statistical approach as a learning technique. A machine-readable dictionary and domain specific texts are used as input to the system to build domain ontologies with both taxonomic (vertical) and non-taxonomic (horizontal) relationships between concepts. The non-taxonomic relationships composed of dependencies between concepts such as synonymy, meronym, antonymy, attribute-of, and possession, are exploited using domain specific texts with the analysis of lexical co-occurrence statistics based on WordSpace. The idea behind the lexical co-occurrence statistics is that terms that appear together may have non-taxanomic relationships between concepts.

EXTREEM-T [20] is a system which exploits the semi-structured resources to acquire the relevant terminology to enrich an ontology. It stands for Xhtml TREE Mining and it utilizes statistical technique such as frequency of occurrences of markups.

DL-Learner [81] is an ontology enrichment system which uses structured input data and relies on Inductive Logic Programming technique. This technique aims to extract concept via logic learned from examples and prior knowledge.

SYNOPSIS [37] is another system which uses the technique of learning by term collocations and co-occurrences. It automatically builds a lexicon for each specific term called criterion by splitting a document into several passages. The correlation between terms and the user criterion is computed using the relative position of these terms and the given criterion. Relative position refers to the number of terms between a term and the user criterion. For each criterion, a lexicon is built in this way.

CoLexIR [122] is and adaptation of SYNOPSIS. It implements the same learning technique as SYNOPSIS but rather than building lexicon of a term, it builds automatically the lexicon of ontology concepts.

A.2.3.3 Hybrid Approaches

WEB $\rightarrow$ KB [28] is a system which combines statistical (Bayesian learning) and logical techniques to identify and extract concepts. The system is primarily trained to acquire the relevant terminology and is then allowed to explore semi-structured web documents to locate and extract these concepts. Two inputs are required to train the system; the first is a set of concepts and relations of interest when creating the knowledge base, and the second is a set of training data consisting of labelled regions of hypertext that represent instances of these concepts and relations.

TEXT-TO-ONTO [26, 89] is an ontology learning system which employs learning by term collocations and co-occurrences technique with a basic linguistic processing technique. The input to the system can be a structured, semi-structured, or an unstructured resource. The frequency of term collocations is computed to locate and acquire non-taxonomic relations using background knowledge i.e. a lexicon and a taxonomy.

BOEMIE [118] is an ontology enrichment system which utilizes both symbolic such as shallow syntactic analysis and statistical concept learning technique to identify and extract concept from the input data. It uses large corpora which can be either a text, image, or a video.

A.2.4 Consolidated Overview

A consolidated overview of the approaches for enriching ontologies presented in this chapter is shown in Table A.2. It constitutes some of the characteristics of the approaches that are presented in this chapter.

The first column of the table contains the reviewed approaches while the following columns denote the characteristics considered and evaluated in this chapter. The entries of the table are values that show which of the evaluated characteristics are supported (denoted with ✓) or not supported (denoted with ✗) by the approaches. As can be seen from

Table A.2: A consolidated overview of the evaluated approaches

Approach	Input Resource			Concept Learning Technique		
	Struct	Semi	Unstruct	Linguistic	Statistical	Hybrid
DOODLE II	✗	✓	✓	✗	✓	✗
CoLexIR	✗	✓	✓	✗	✓	✗
HASTI	✗	✗	✓	✗	✓	✗
KnowItAll	✗	✓	✓	✓	✗	✗
EXTREEM-T	✗	✓	✓	✗	✓	✗
SynDiKATe	✗	✗	✓	✓	✗	✗
MedSynDiKATe	✗	✗	✓	✓	✗	✗
SYNOPSIS	✗	✗	✓	✗	✓	✗
TEXT-to-ONTO	✓	✓	✓	✗	✗	✓
WEB $\rightarrow$ KB	✗	✓	✓	✗	✗	✓
DL-Learner	✗	✓	✓	✗	✓	✗
BOEMIE	✓	✓	✓	✗	✗	✓

the Table A.2, unstructured data are used among all the approaches as input resources to extract the concepts; the structured data is the one supported by only a few approaches (TEXT-to-ONTO and BOEMIE) and semi-structured data are used as input resources by 8 out of 12 approaches presented in this chapter. We also observed that there exists almost an equal use of concept learning techniques among all the approaches shown in this chapter.

Categorizing SEMCON model in one of the categories of approaches of concept enrichment is not an easy task due to differences which exist in many dimensions amongst approaches. Shamsfard and Barforoush [135] identified six main categories of the major distinguishing factors between ontology learning approaches. Although there exist differences amongst approaches, they however have some dimensions in common. From this perspective, SEMCON can be considered as a hybrid approach that to some extent utilizes both approaches, linguistic and statistical. From the linguistic point of view, SEMCON uses morpho-syntactic analysis to identify and extract noun terms, as part of speech, which represents the most meaningful terms in a document. From the statistical point of view, SEMCON derives the context using cosine similarity between term vectors whose members are frequencies of terms. SEMCON employs, besides term frequency, two new statistical features, i.e. term font size and term font type, to determine the context. In addition to context, SEMCON also incorporates the semantic information of terms using the lexical database WordNet and finally aggregates both contextual and semantic information of this term.

A.3 SEMCON

This section describes the proposed SEMCON model to enrich concepts of a domain ontology with new terms which are closely related using the contextual and semantic information. The model, illustrated in Figure A.2, consists of four modules, which are explained in the following subsections.

A.3.1 Preprocessing

This module initially collects a document and partitions that into subsets of text known as passages. These passages are text portions which have very strong semantic coherence and are clearly disconnected from adjacent parts [129]. The partitioned passages can either be

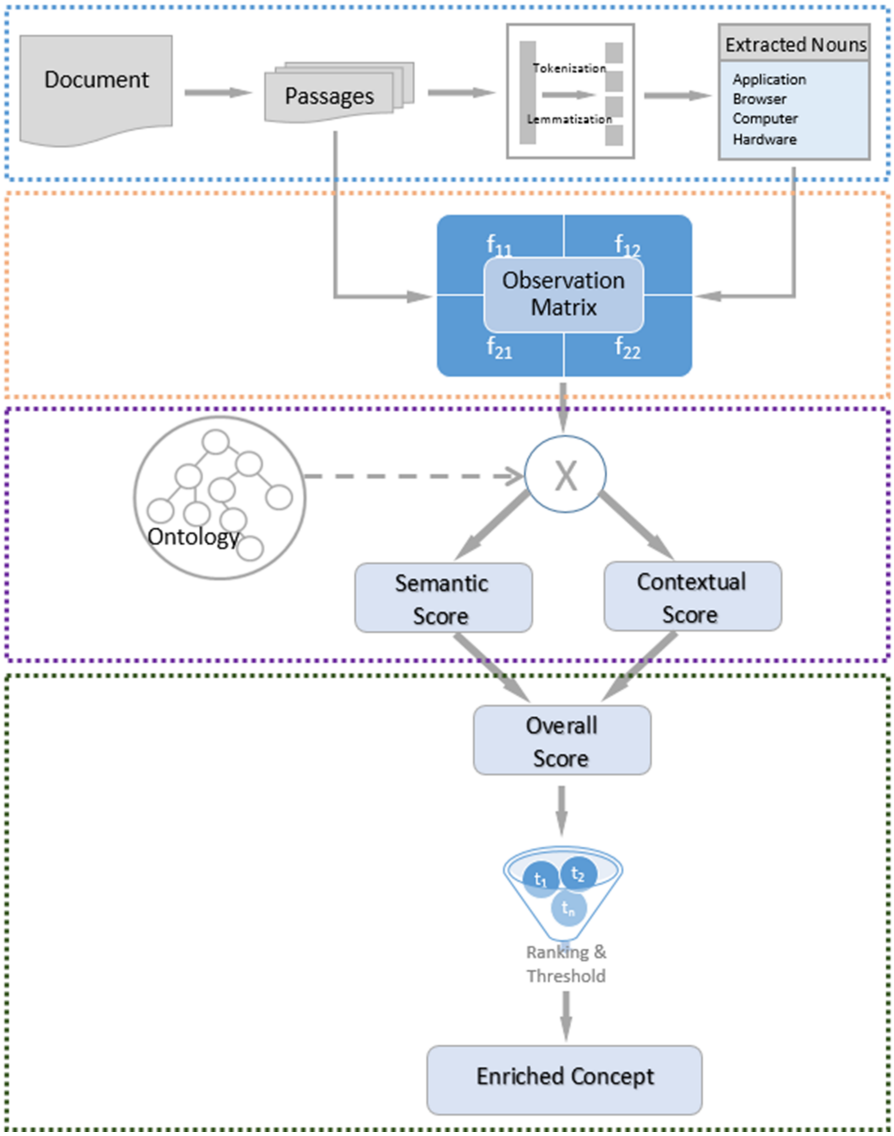


Figure A.2: Block diagram of SEMCON

fixed or variable length. They can also be classified into contextual passages if the partitioning takes into account the context of the document or they can be classified as statistical passages.

In this paper, we take into account the context of a document irrespective of the length of partitioned passages. Partitioned passages are treated as independent documents. A morpho-syntactic analysis using TreeTagger [132] is performed on the partitioned passages. Passages are later cleaned by removing all punctuation and capitalization followed by a

tokenizer step to separate the text into individual terms. The lemmatization is the last step used to find the normalized form of these terms.

The potential terms that are obtained as a result of this preprocessing step can either be a noun, verb, adverb or adjective. These are different parts-of-speech (POS) of a language. It is a well established fact that nouns represent the most meaningful terms in a document [84], thus our focus is on processing only noun terms for further consideration.

A.3.2 Observation Matrix

Computation of the observation matrix is the next step in the proposed model. Observation matrix is a rectangular matrix where the rows represent the extracted passages from a particular document and columns are the terms extracted from those particular passages. An example of observation matrix is shown in Table A.3.

Table A.3: A part of the observation matrix from computer domain

Slide	Computer	Data	Device	Function	Hardware	System	Web
1	6.25	5.25	1.75	0	0	0	0
2	3	0	0	0	0	1.5	8
3	9.25	0	7	1.75	4.75	5.5	0
4	5.5	3.5	8	0	0	0	0
5	5	1.5	1.5	1.5	0	2	0
6	12.25	0	0	1.5	0	6	0
7	2.25	0	0	0	0	6.25	0

Each entry of the observation matrix is calculated by accumulating the sum of term frequency, term font size and term font type in each of the extracted passages, as shown in Equation A.1. Introducing of term font type and term font size, as very important factors in the information finding process [58], is inspired from the representation of tags in the tag cloud [10]. The effect of these statistical features is discussed in subsection A.5.1.

$$E_{i,j} = tf_{i,j} + \sum_{k \in tf} (ft_{i,j,k} + fs_{i,j,k}) \tag{A.1}$$

where, $tf_{i,j}$ denotes the frequency of occurrences of a term *i* in document *j*. $ft_{i,j,k}$ and $fs_{i,j,k}$ indicate the aggregated values of font types and font sizes computed over all occurrences *k* of a term *i* in a document *j*.

We adopt a linear increase model for different font types and font sizes. The linear model assumes that the effect of each variable is the same for all values of the other variables. For example, the model assumes that the effect of bold font type terms is the same for every value of underline or italic font type terms. The same way, the effect of underlined font type terms is the same for every value of underline bold or italic font type terms, and so on.

Algorithm 1 in Figure A.1 describes the computation of observation matrix using three statistical parameters: frequency of term occurrences, bold font type and four different font sizes. More precisely, lines 3-13 of the algorithm show entries of the observation matrix computed using the frequency of occurrences of terms that appear in bold (α) and the frequency of occurrences of these terms with font sizes (β) as, either level 3 (line 4), level 2 (line 6), level 1 (line 8), or title (line 10). In the same fashion, we computed entries of the observation matrix using the terms that appear in a document as either italic, underline, and regular and with font sizes as either level 3, level 2, level 1 or as a title.

The input of the algorithm is a collection of rich documents from which font sizes and font types of terms used to build the observation matrix are derived. In this work, we used the font sizes from the presentations slides where the level 1 font size is set to 28 pt, level 2 is

Algorithm A.1: The algorithm for computation of the observation matrix

Input : A collection of pdf documents
Output: Entries of the observation matrix

for *each* $Doc \in D$ **do**
  for *each* $t \in Doc$ **do**
    if $t \in Doc$ *is bold* **then**
      if $t_{size} < 10pt$ **then**
        $Compute\ E\ as\ E + tf + 0.75 * \alpha + 0.25 * \beta$
      end
      if $10pt \leqslant t_{size} < 14pt$ **then**
        $Compute\ E\ as\ E + tf + 0.75 * \alpha + 0.50 * \beta$
      end
      if $14pt \leqslant t_{size} < 18pt$ **then**
        $Compute\ E\ as\ E + tf + 0.75 * \alpha + 0.75 * \beta$
      end
      if $t_{size} \geqslant 18pt$ **then**
        $Compute\ E\ as\ E + tf + 0.75 * \alpha + 1.00 * \beta$
      end
    end
  end
end
return E;

set to 24 pt, and level 3 is set to 20 pt. These parameters can be adjusted for other document types. According to these font size settings, we observed the occurrences of terms among the presentation slides. However, the input of algorithm can be a collection of documents other than ppt as long as font sizes and font types of terms can be computed for all types of rich texts using the HTML tags.

The example illustrated in Figure A.3 shows that term *Web* occurred 4 times in the presentation slides, where 2 times it appeared as level 1 font size and as bold font type and 2 times it appeared as level 2 font size.

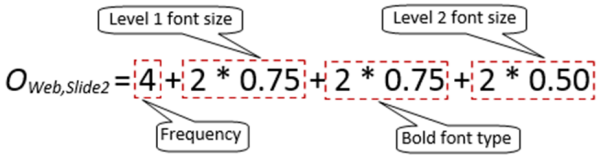


Figure A.3: Building of observation matrix using statistical features

A.3.3 Computation of Contextual and Semantic Score

The observation matrix is used as an input to compute the term-to-term contextual and semantic score between two terms in order to find a matching term extracted from a passage to a concept in the ontology.

Contextual information score (S_{con}) for a pair of terms t_i and t_j is computed using the cosine similarity metric with respect to the passages, as given by Equation A.2.

$$S_{con}(t_i, t_j) = \frac{t_i \cdot t_j}{\| t_i \| \| t_j \|} \tag{A.2}$$

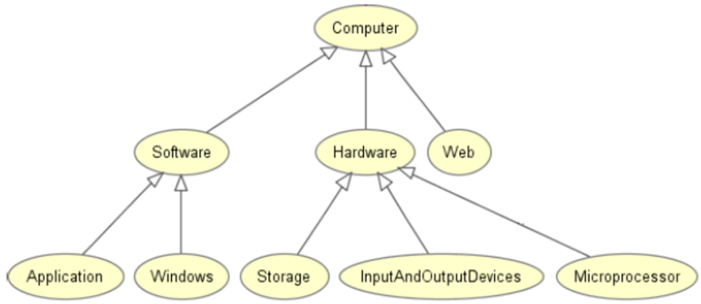


Figure A.4: Ontology sample of the computer domain

where, t_i and t_j represent the term vectors of the observation matrix. The dot product between two term vectors reflects the extent to which two terms are similar in the vector space.

A term square matrix is used to store S_{con} values among all extracted term. This matrix will later be used in computing an overall correlation between a term extracted from a document and a concept in the ontology, as described in subsection A.3.4.

Further, the proposed model maps a term to a concept of ontology via the matching technique. The basic idea behind this technique is to search for the concept labels that occur exactly and/or partially in the observation matrix. The exact and partial matching is defined as the following.

Definition A.1
Let O be the domain ontology and M the observation matrix constituted of a finite set of terms, $M = \{t_1, t_2, ..., t_i\}$.

The mapping of term $t_i \in M$ into concept $c_j \in O$ is defined as the exact matching $EM(t_i, c_j)$, where,

$$EM(t_i, c_j) = \begin{cases} 1, & \text{if label } (c_j) = t_i \\ 0, & \text{if label } (c_j) \neq t_i \end{cases} \tag{A.3}$$

The mapping of term $t_i \in M$ into concept $c_j \in O$ is defined as the partial matching $PM(t_i, c_j)$, where

$$PM(t_i, c_j) = \begin{cases} 1, & \text{if label } (c_j) \text{ contains } t_i \\ 0, & \text{if label } (c_j) \text{ does not contain } t_i \end{cases} \tag{A.4}$$

If $EM(t_i, c_j) = 1$, it means that term t_i and single concept label c_j are exactly the same, then term t_i is replaced by SEMCON with concept c_j. For example, for the concept in the ontology such as *Application* or *Storage* illustrated in Figure A.4, there exists the same term in the term square matrix.

If $PM(t_i, c_j) = 1$, it means that term t_i is part of compound concept label c_j, then term t_i is replaced by SEMCON with the highly-correlated terms of concept c_j. For example, consider *InputAndOutputDevices* as one of the compound ontology concepts, and the *Device* as one of the terms in the term square matrix. Let *Screen*, *Display*, *Input*, be the highly-correlated terms with the term *Device*, and in that case, the *InputAndOutputDevices* will be enriched with the correlation terms of the term *Device* e.g. with *Screen*, *Display*, *Input*.

The next step is the computation of the semantic information score. The semantic score is computed using the information found in WordNet database by employing Wu& Palmer similarity measure [156]. WordNet [44] is a lexical database for the English language that groups terms into sets of synonyms called synsets and defines the semantic relations between these synsets. To find the correct meaning of terms textitt$_i$ and t_j under consideration, we have tested with two Word Sense Disambiguation techniques, namely, the Pre-

dominant sense heuristic, and the Maximizing semantic similarity. The Predominant sense heuristic also known as the First sense heuristic technique relies on the distribution of the senses and it assumes that the most common sense of a word represents the correct meaning of this given word. Maximizing semantic similarity is also a technique used to disambiguate word senses. It follows the idea that the right sense (correct meaning) of a term is the one which maximizes the relatedness between the term and a sense among all possible senses. The empirical analysis shows that both these disambiguation techniques yield almost the same performance in terms of precision but the predominant sense heuristic technique is often used as a baseline [96]. Therefore, SEMCON employs the predominant sense heuristic disambiguation technique for finding the correct sense and all the results presented in this chapter are computed based on this technique.

The semantic score, $S_{sem}(t_i, t_j)$, is calculated for all possible pairs t_i and t_j from the observation matrix, where t_i, $t_j \in O$ and O is the observation matrix. As a result, for each term, a hash table is generated where the most similar terms are set as the synonyms for that term. Mathematically, the semantic score is computed using Equation A.5.

$$S_{sem}(t_i, t_j) = \frac{2 * depth(lcs)}{depth(t_i) + depth(t_j)} \tag{A.5}$$

where, *depth(lcs)* indicates the least common subsumer of terms t_i and t_j; *depth(t_i)* and *depth(t_j)* indicate the path's depth of terms t_i and t_j, in the WordNet lexical database.

A.3.4 Overall Score

The overall correlation between two terms, t_i and t_j, is calculated using the contextual and semantic score. Mathematically, the overall score is given in Equation A.6.

$$S_{ove}(t_i, t_j) = w * S_{con}(t_i, t_j) + (1 - w) * S_{sem}(t_i, t_j) \tag{A.6}$$

where w is a parameter with value set as 0.5 based on the empirical analysis performed on the data set given in Section Experimental Procedure. A thorough analysis about the effect of the weight parameter value on the output of the SEMCON is given in subsection A.5.2. The overall score is in the range (0,1]. The overall score is 1 if two terms are the same and 0 when there is no relationship between them.

Finally, a rank cut-off method is applied using a threshold to obtain terms which are closely related to a given term in the ontology. Terms that are above the specified threshold (top-N) are considered to be the relevant terms for enriching the concepts.

A simple example of the SEMCON output, given in Table A.4, shows the top 10 terms obtained as the most relevant terms of *Application* concept. 6 of these terms, namely *Application*, *Program*, *Apps*, *Function*, *Task* and *Software* are amongst the top 10 terms selected by the subjects as the closest terms to concept *Application*.

Table A.4: Top 10 closely related terms of Application concept

Concept	The Top 10 terms obtained by SEMCON model
Application	Apps, Application, Software, Program, Control Task, Part, Master, Operation, Function

A.4 Experimental Procedures

The experiment uses, presentation slides dataset from 5 different domains as shown in Table A.5. The presentations in the database are from domain of Computer, Database, Internet, C++ Programming and Software Engineering. The dataset was limited to a maximum

Table A.5: Dataset used for experimentation

No	Domain name	# of slides	# of terms	# of concepts
1	Computer	7	79	9
2	Database	9	105	8
3	Internet	7	73	7
4	C++_Programming	9	70	10
5	Software_Engineering	7	42	7

of 5 presentations with a restricted number of slides due to the subjective nature of the experiment.

This section presents two approaches to evaluate the performance of the SEMCON. The first one is the subjective evaluation and the second one is the objective evaluation.

A.4.1 Subjective Evaluation

To evaluate the performance of SEMCON, a subjective survey was carried out by publishing an online questionnaire to 15 subjects.

The subjects were all computer science PhD students and Postdocs at the Gjøvik University College. They were asked to select 5 closely related terms from a list of terms for each of the concepts, starting from the most relevant term as their first choice, the second relevant term as the second choice and so on. A screenshot taken from the questionnaire about the computer domain is illustrated in Figure A.5.

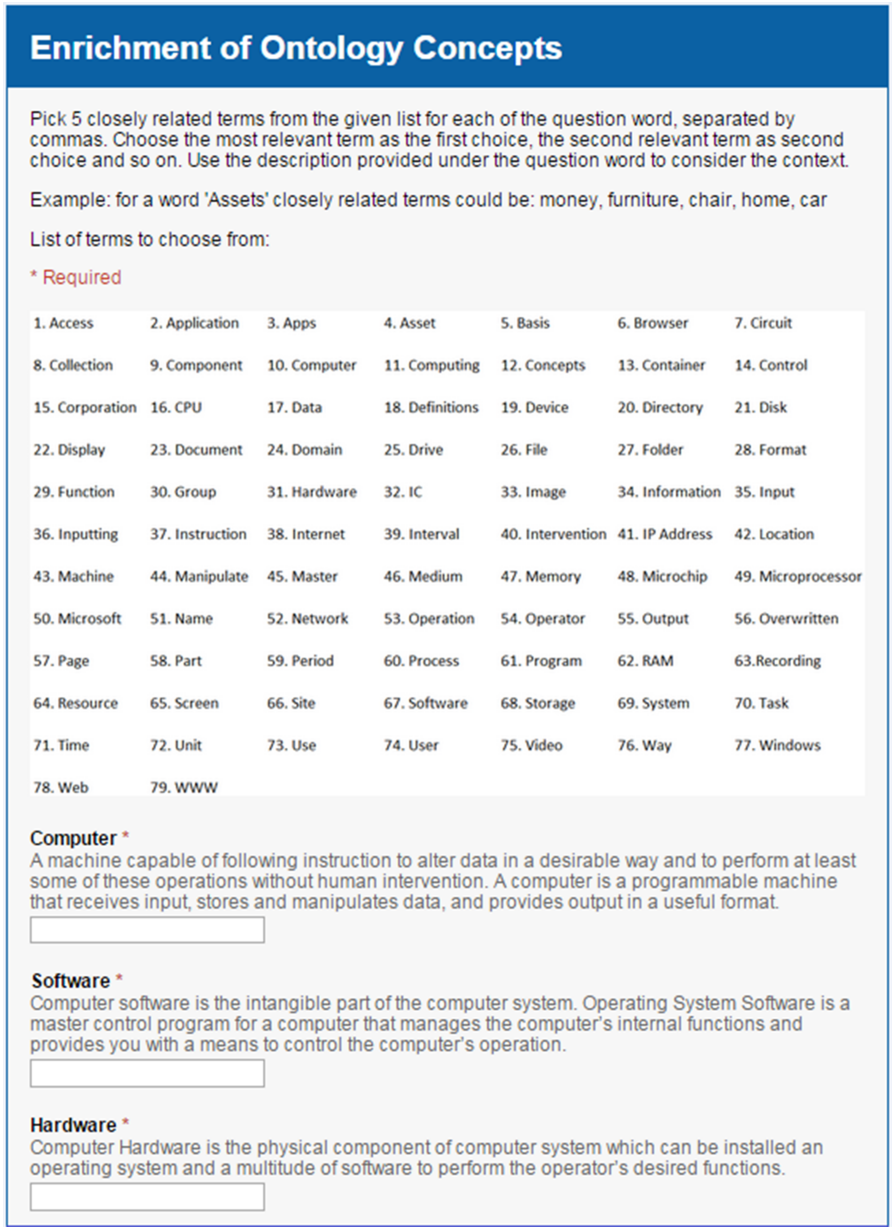
For each of the concepts given in the subjective survey, we obtained the ranking of the corresponding term and its frequency count. An example of ranking terms and calculating the counts of the corresponding term frequencies is given in Table A.6.

Table A.6: Terms selected by subjects for the Application concept

Terms	Pos 1	Pos 2	Pos 3	Pos 4	Pos 5	Total
Apps	2	8	2	1	0	13
Software	3	0	4	3	1	11
Program	0	2	3	1	2	8
Application	8	0	0	0	0	8
User	0	1	1	1	3	6
Task	0	0	2	2	1	5
Windows	1	1	0	2	1	5
Browser	0	0	2	2	0	4
Process	0	0	1	1	2	4
Microsoft	1	1	0	0	0	2
System	0	0	0	0	2	2
Computer	0	1	0	0	0	1
Data	0	0	0	0	1	1
Recording	0	0	0	0	1	1

Pos(n) is the position of the selected term from the term list. It shows how many times a particular term is selected at nth position, e.g. the term *Apps* is chosen by 2 subjects as their 1st choice for the *Application* concept, by 8 subjects as their 2nd choice and so on. The total number of times a particular term being selected by subjects for the *Application* concept is computed by aggregating all these frequencies together.

For each selected term, a single score is computed using the Borda Count method. Borda Count method is an election method used to determine a winner from a voting where voters rank the candidates in order of preference [160]. The mathematical formulation of



Enrichment of Ontology Concepts

Pick 5 closely related terms from the given list for each of the question word, separated by commas. Choose the most relevant term as the first choice, the second relevant term as second choice and so on. Use the description provided under the question word to consider the context.

Example: for a word 'Assets' closely related terms could be: money, furniture, chair, home, car

List of terms to choose from:

* Required

1. Access	2. Application	3. Apps	4. Asset	5. Basis	6. Browser	7. Circuit
8. Collection	9. Component	10. Computer	11. Computing	12. Concepts	13. Container	14. Control
15. Corporation	16. CPU	17. Data	18. Definitions	19. Device	20. Directory	21. Disk
22. Display	23. Document	24. Domain	25. Drive	26. File	27. Folder	28. Format
29. Function	30. Group	31. Hardware	32. IC	33. Image	34. Information	35. Input
36. Inputting	37. Instruction	38. Internet	39. Interval	40. Intervention	41. IP Address	42. Location
43. Machine	44. Manipulate	45. Master	46. Medium	47. Memory	48. Microchip	49. Microprocessor
50. Microsoft	51. Name	52. Network	53. Operation	54. Operator	55. Output	56. Overwritten
57. Page	58. Part	59. Period	60. Process	61. Program	62. RAM	63.Recording
64. Resource	65. Screen	66. Site	67. Software	68. Storage	69. System	70. Task
71. Time	72. Unit	73. Use	74. User	75. Video	76. Way	77. Windows
78. Web	79. WWW					

Computer *
A machine capable of following instruction to alter data in a desirable way and to perform at least some of these operations without human intervention. A computer is a programmable machine that receives input, stores and manipulates data, and provides output in a useful format.

Software *
Computer software is the intangible part of the computer system. Operating System Software is a master control program for a computer that manages the computer's internal functions and provides you with a means to control the computer's operation.

Hardware *
Computer Hardware is the physical component of computer system which can be installed an operating system and a multitude of software to perform the operator's desired functions.

Figure A.5: A screenshot taken from the questionnaire

Borda Count is given in Equation A.7.

$$BC(t) = \sum_{i=1}^{m} [(m + 1 - i) * freq_i(t)] \tag{A.7}$$

where *BC(t)* of a given term *t* is computed by a total sum of the weights of the frequencies $freq_i(t)$. $freq_i(t)$ is the frequency of term *t* chosen at position *i*, and m is the total number of possible positions, in our case 5.

The scores from the Borda Count are then sorted to obtain the top 'n' terms, giving us the refined list of the highest scoring terms. For our experiment, we set n = 10, and this gives us the top 10 terms as shown in Table A.7. This is our ground truth data.

Table A.7: Borda count of subjects' responses for the Application concept

Rank	Term	Borda Count
1	Apps	50
2	Application	40
3	Software	34
4	Program	21
5	Windows	14
6	Task	11
7	Browser	10
8	Function	9
9	User	9
10	Process	7

A.4.2 Objective Evaluation

In addition to the subjective experiment, an objective evaluation is carried out where the results obtained from the SEMCON model are compared with the results obtained from the three state-of-the-art methods namely Term Frequency Inverse Document Frequency (*tf*idf*) [133], χ^2 (Chi square) [85] and Latent Semantic Analysis - LSA [80].

*tf*idf* is a mathematical method which is used to find key vocabulary that best represents the texts. Mathematically, it is given in Equation A.8.

$$tf * idf = tf_{i,j} * \log \frac{N}{df_j} \tag{A.8}$$

where, $tf_{i,j}$ is the term frequency of term *j* that occurs in a passage, *N* is the total number of passages in the corpus and df_j shows the number of passages where the term *j* occurs.

The traditional *tf*idf* considers only the term to document relation and thus it is not appropriate for comparison as it is. Therefore, we modified the existing *tf*idf* in order to take the term to term relation into account. This is achieved using the cosine measure where the dot product between two vectors of *tf*idf* matrix shows the extent to which two terms are similar in the vector space.

χ^2 is a statistical method which computes the relationship between two given terms. Mathematically, it is given in Equation A.9.

$$\chi^2_{t_a,t_b} = \sum_{i \in \{t_a, \neg t_a\}} \sum_{j \in \{t_b, \neg t_b\}} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} \tag{A.9}$$

where, $O_{i,j}$ and $E_{i,j}$ show the co-occurrence and the expected co-occurrence frequency between two terms t_a and t_b. More formally, the co-occurrence frequency between two terms t_a and t_b is the observed frequency $O_{i,j}$ where $i \in \{t_a, \neg t_a\}$ and $j \in \{t_b, \neg t_b\}$. Thus, O_{t_a,t_b} is the observed frequency of passages which contain term t_a and term t_b. $O_{t_a,\neg t_b}$ is the observed frequency of passages which contain term t_a but do not contain term t_b. $O_{\neg t_a,t_b}$ is the observed frequency of passages which do not contain t_a but contain the term t_b. $O_{\neg t_a,\neg t_b}$ is the observed frequency of passages which contain neither term t_a nor term t_b.

Latent semantic analysis (LSA), sometimes referred as latent semantic indexing, is a method for extracting and representing the content of a text using the relationships between terms that occur in similar context.

The first step of LSA is representing the text document as a matrix in which each row denotes a unique term and each column denotes a passage. Each cell contains the frequency of occurrence of one term from the passage.

The second step of LSA is applying a Singular Value Decomposition (SVD). SVD decomposes the rectangular matrix into the product of three matrices. One matrix is term vectors, another denotes a diagonal matrix and the last one denotes passage vectors. More formally, every rectangular matrix M can be decomposed into three matrices T, Σ and P^T, as shown in Equation A.10.

$$M = T\Sigma P^T \tag{A.10}$$

where, T is a term vectors matrix, P^T is a matrix of passage vectors and Σ is a diagonal matrix of decreasing singular values.

The singular values represent the semantic space for terms and passages in a corpus of text. When the matrix Σ contains all the singular values of M, then the original matrix M is reconstructed by multiplying the three matrices T, Σ, and P^T.

The dimensionality of the space of semantic representations can be reduced by deleting some of the singular values, starting with the smallest. The matrix M_k, which is the k dimensional approximation to M, can be built by selecting the k largest singular values. In our case, we set the dimensionality parameter k to 2. The reconstruction of matrix M_k is given in Equation A.11.

$$M_k = T\Sigma_k P^T \tag{A.11}$$

Similarly, the representations of terms and passages by multiplying their corresponding matrix decompositions are obtained. The representations of terms and passages are given in Equation A.12.

$$T_k = T\Sigma_k \qquad P_k^T = \Sigma_k P^T \tag{A.12}$$

Finally, to calculate the similarity between two terms, we used the cosine measure, where the dot product between two vectors of matrix M_k shows the extent to which two terms are similar in the vector space. Cosine similarity measure is given in Equation A.13.

$$Similarity_{LSA}(t_i, t_j) = \frac{t_i \cdot t_i}{\parallel t_i \parallel \parallel t_i \parallel} \tag{A.13}$$

where, t_i and t_i are terms, and $\parallel t_i \parallel$ and $\parallel t_j \parallel$ are the corresponding latent term space vectors.

A.4.3 Measures of the Effectiveness for the Objective Methods

We employed the standard information retrieval measures such as Precision, Recall and F1 [133] to evaluate the effectiveness of objective methods. The objective methods are evaluated against the subjective ones. The evaluation is conducted by taking the 10 top subjective terms as the ground truth and the top-N terms obtained by the objective methods as a relevance list.

The definition of precision and recall is adjusted in order to evaluate top-N terms obtained by objective methods. The definitions adopted are as following.

Precision is the ratio of total number of terms which occur simultaneously in the relevance list and in the ground truth list, to the number of terms in the relevance list. Precision is given in Equation A.14.

$$Precision = \frac{\mid Relevance \cap GroundTruth \mid}{\mid Relevance \mid} * 100 \tag{A.14}$$

Recall is the ratio of total number of terms which occur simultaneously in the relevance list and in the ground truth list, to the number of terms in the ground truth list. Recall is given in Equation A.15.

$$Recall = \frac{\mid Relevance \cap GroundTruth \mid}{\mid GroundTruth \mid} * 100 \tag{A.15}$$

Precision and recall are often inversely related to each other, such that if the number of relevant terms increases, then the value of recall increases, while at the same time precision decreases. Thus, we used the standard F1 measure, which is defined as the average of precision and recall and it is given in Equation A.16.

$$F1 = \frac{2 * (Precision * Recall}{Precision + Recall} * 100 \tag{A.16}$$

A.5 Results and Analysis

The performance of objective methods is evaluated on two criteria. First being how well the objective methods score the top subjective terms. In order to do this, scores for the 10 top terms are taken as the ground truth. The score obtained for these terms using the objective methods are then evaluated. An example for the enrichment of the *Application* concept is observed and the comparison is shown in Table A.8. The final score is in the range of [0,1], where 0 denotes a term with no relatedness and 1 denotes a highly related term for enriching the Application concept.

Table A.8: The overall objective score for the top 10 terms selected by subjects

No	Subjective terms	tf*idf	χ^2	LSA	SEMCON
1	Apps	1.000	1.000	1.000	1.000
2	Application	1.000	1.000	1.000	1.000
3	Software	0.975	0.500	0.981	0.943
4	Program	0.914	0.500	0.894	0.923
5	Windows	0.000	0.028	0.000	0.409
6	Task	1.000	1.000	1.000	0.900
7	Browser	0.000	0.028	0.000	0.479
8	Function	0.569	0.222	0.577	0.701
9	User	0.603	0.417	0.707	0.544
10	Process	0.000	0.067	0.000	0.412

The comparison summarised in Table A.8, shows that SEMCON generally outperforms the *tf*idf*, χ^2 and LSA. The red highlighted values show cases when one method performs better than the other. It can be seen from the red highlighted values that the SEMCON model gives much better results for the terms *Windows*, *Browser* and *Process* in contrast to the *tf*idf* and LSA which scores 0 to these three terms and χ^2 which scores close to 0. This is most likely because these terms did not occur in document/presentation slides that talk contextually about the *Application* concept but they occurred in the WordNet corpus. SEMCON also scores higher for the terms *Program* and *Function*. The term *Task* gets a score of 1.0 by the *tf*idf*, χ^2 and LSA, which means that these three methods would rank the term *Task* as its first term to enrich the *Application* concept. The term *Task* however is ranked as the sixth relevant term to enriching the concept *Application* by subjects as shown in Table A.7.

The second evaluation criteria is to check if the top terms scored by the objective methods are accurate. For this, we compute the precision, recall and F1 measure on the top-15 relevant terms list. Table A.9 shows the resulting precision and recall of objective methods on retrieving and ranking of terms as the most relevant terms for enriching the *Application*

concept in the computer domain. Terms correctly retrieved by the objective methods are highlighted in red in Table A.9.

Table A.9: Precision and recall of Application concept

	Objective terms			
Subjective terms	tf*idf	χ^2	LSA	SEMCON
Apps	Apps	Apps	Application	Apps
Application	Application	Application	Control	Application
Software	Control	Control	Apps	Software
Program	Master	Master	Master	Program
Windows	Part	Part	Part	Control
Task	Task	Task	Task	Task
Browser	Software	Program	Web	Part
Function	Program	Software	File	Master
User	Operation	Operation	Page	Operation
Process	Computer	User	Access	Function
	User	Computer	Asset	Computer
	Function	Function	Browser	System
	System	Component	Collection	User
	Component	System	Concept	Browser
	Access	Device	User	Use
Recall	**70.0**	**70.0**	**50.0**	**80.0**
Precision	**46.7**	**46.7**	**33.3**	**53.3**

In the following paragraph, we are giving an example to show how the precision and recall, shown in Table A.9, are computed. Total number of terms obtained by intersection of ground truth list (column entitled subjective terms) and relevance list (column entitled *tf*idf*) is equal to 7. Number of terms in ground truth list is 10, while number of terms in relevance list is 15. Recall is computed as 7/10*100=70.0% and precision as 7/15*100=46.7%. The example illustrated shows computation of precision and recall for *tf*idf* method but in a similar fashion they are also computed for χ^2, LSA, and SEMCON.

Additionally, Table A.10 and Table A.11 shows precision, recall and F1 results obtained by the SEMCON on retrieving and ranking of terms as the most relevant terms for enriching concepts of computer domain and other domains, respectively.

Table A.10: The performance of SEMCON on computer domain

Domain	P (%)	R (%)	F1 (%)
Computer	26.7	40.0	32.0
Software	46.7	70.0	56.0
Hardware	33.3	50.0	40.0
Web	46.7	70.0	56.0
Storage	46.7	70.0	56.0
Microprocessor	40.0	60.0	48.0
InputAndOutputDevices	33.3	50.0	40.0
Application	53.3	80.0	64.0
Windows	46.7	70.0	56.0
Average	**41.5**	**62.2**	**49.8**

The performance of SEMCON in terms of F1 measure is compared with the performance of *tf*idf*, χ^2 and LSA. The comparison is performed using results of various domains and it shows that SEMCON achieved better results on finding the highly related terms to enrich ontology concepts.

Table A.11: The performance of SEMCON on different domains

Domain	P (%)	R (%)	F1 (%)
Computer	41.5	62.2	49.8
Database	34.2	51.3	41.0
Internet	38.1	57.1	45.7
C++_Programming	37.3	56.0	44.8
Software_Engineering	49.5	74.3	59.4

Table A.12 shows F1 results for computer domain. The results depict that SEMCON achieved the average improvement of 12.2% over the *tf*idf*, 21.8% over the χ^2, and 24.5% over the LSA.

Table A.12: The F1 of objective methods performed on computer domain

Concept	tf*idf (%)	χ^2 (%)	LSA (%)	SEMCON(%)
Computer	24.0	24.0	32.0	32.0
Software	56.0	48.0	40.0	56.0
Hardware	32.0	40.0	32.0	40.0
Web	32.0	32.0	40.0	56.0
Storage	64.0	56.0	64.0	56.0
Microprocessor	48.0	40.0	56.0	48.0
InputAndOutputDevices	32.0	24.0	8.0	40.0
Application	56.0	56.0	40.0	64.0
Windows	56.0	48.0	48.0	56.0
Average	**44.4**	**40.9**	**40.0**	**49.8**

The same comparisons for F1 measure is also conducted for other domains. These results are shown in Tables A.13 - A.16.

Table A.13: The F1 of objective methods performed on SE domain

Concept	tf*idf (%)	χ^2 (%)	LSA (%)	SEMCON(%)
Software	56.0	56.0	40.0	48.0
Cost	40.0	40.0	40.0	48.0
Product	64.0	56.0	48.0	48.0
Attribute	32.0	48.0	32.0	56.0
Process	72.0	48.0	32.0	72.0
Generic	48.0	64.0	64.0	72.0
Hybrid	64.0	56.0	56.0	72.0
Average	**53.7**	**52.6**	**44.6**	**59.4**

Finally, we evaluated the performance of SEMCON and the three other objective methods by comparing the average results of each domain. The obtained results (precision, recall and F1) illustrated in Figure A.6 - A.8 show that SEMCON gives better results than the other three methods for all the domains excepts for the internet domain. This may have happened due to the fact that subjects are making their selections based on the descriptions provided under each concept in the questionnaire, when they were asked to select 5 closely related terms. In other words, subjects might have used contextual information from the description provided in the questionnaire about each concept rather than their existing prior knowledge. As the ground truth list is composed of terms which carry contextual meaning in a document to describe a particular concept, therefore this might have served better for *tf*idf* for Internet domain where people choose terms based on the context

Table A.14: The performance of SEMCON on C++ programming domain

Concept	tf*idf (%)	χ^2 (%)	LSA (%)	SEMCON(%)
C++_Programming	24.0	40.0	40.0	40.0
Syntax	56.0	48.0	48.0	48.0
Technique	24.0	16.0	24.0	24.0
Structure	40.0	40.0	32.0	40.0
Expression	48.0	40.0	40.0	48.0
Operator	24.0	24.0	56.0	24.0
Encapsulation	48.0	64.0	64.0	48.0
Inheritance	64.0	56.0	56.0	56.0
Polymorphism	48.0	48.0	40.0	56.0
Platform	56.0	56.0	48.0	64.0
Average	**43.2**	**43.2**	**44.8**	**44.8**

Table A.15: The F1 of objective methods performed on database domain

Concept	tf*idf (%)	χ^2 (%)	LSA (%)	SEMCON(%)
Database	24.0	16.0	24.0	16.0
Model	48.0	40.0	16.0	48.0
E-R	48.0	48.0	16.0	48.0
User	40.0	16.0	16.0	16.0
SQL	32.0	40.0	16.0	32.0
DDL	64.0	48.0	40.0	64.0
DML	40.0	24.0	32.0	48.0
Administrator	24.0	24.0	16.0	24.0
Average	**40.0**	**32.0**	**22.0**	**41.0**

Table A.16: The F1 of objective methods performed on internet domain

Concept	tf*idf (%)	χ^2 (%)	LSA (%)	SEMCON(%)
Internet	40.0	24.0	48.0	40.0
Application	40.0	40.0	40.0	32.0
Web	32.0	32.0	40.0	32.0
Access	56.0	48.0	40.0	48.0
Browser	64.0	48.0	48.0	64.0
ISP	72.0	48.0	56.0	64.0
HTML	40.0	48.0	56.0	40.0
Average	**49.1**	**41.4**	**46.9**	**45.7**

rather than prior domain knowledge. Nevertheless, there is a significant improvement of results for other domains by SEMCON over other methods.

A.5.1 The Impact of Statistical Features

The SEMCON takes into account the context of a term by computing an observation matrix, which exploits the statistical features such as term font type and term font size besides the frequency of the occurrence of a term. An example of observation matrix, with or without using the statistical features, is shown in Table A.17.

Table A.17 d depicts a part of the observation matrix whose entries are computed without statistical features and with statistical one. These entries are computed for five different terms: *computer*, *data*, *device*, *system*, and *web*. For each term, the first column shows the score obtained using only term's frequency (denoted with No), and the second column

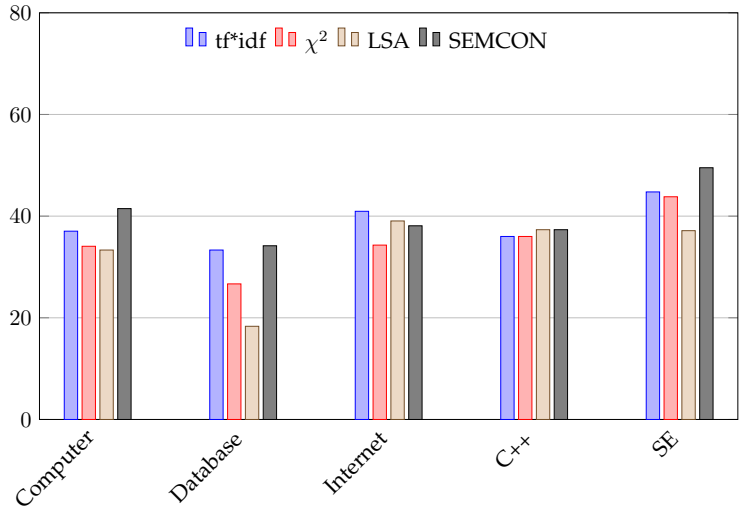


Figure A.6: Precision for 5 different domains

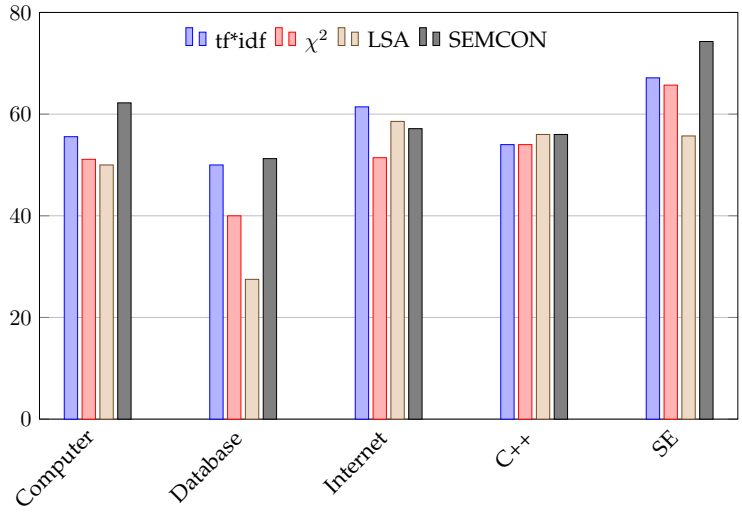


Figure A.7: Recall for 5 different domains

shows the score obtained using statistical features.

It is evident from Table A.17 that statistical features do contribute to observation matrix score but there is a need to investigate into how much each of the statistical features i.e. the font size and font type, contribute to the overall performance of SEMCON. The contribution presented for the computer domain dataset is shown in Table A.18. Furthermore, Table A.18 gives a comparison of Precision, Recall and F1 measures of SEMCON, when the observation matrix is built using the statistical features and when the observation matrix is built

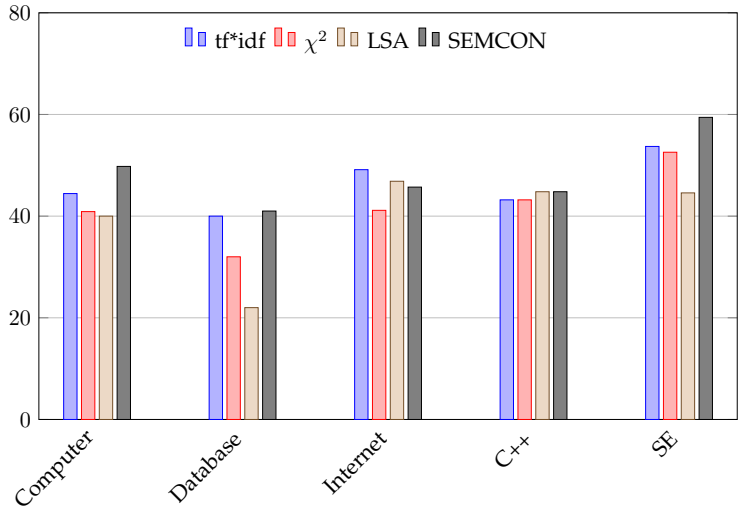


Figure A.8: F1 for 5 different domains

Table A.17: An example of observation matrix with/without using statistical features

Slide	Computer		Data		Device		System		Web	
	No	With	No	With	No	With	No	With	No	With
1	3	6.25	3	5.25	1	1.75	0	0	0	0
2	2	3	0	0	0	0	1	1.5	4	8
3	5	9.25	0	0	4	7	3	5.5	0	0
4	3	5.5	2	3.5	5	8	0	0	0	0
5	3	5	1	1.5	1	1.5	1	2	0	0
6	7	12.25	0	0	0	0	3	6	0	0
7	1	2.25	0	0	0	0	3	6.25	0	0

only using the frequency of the occurrence of a term. The average F1 measure is improved by 3.75% when the observation matrix is built using statistical features. The F1 measures of *Web* and *InputAndOutputDevices* concepts are improved by 16.7% and 25.0%, respectively. This happened due to the fact that these terms occurred very often as level 1 font size and bold font type in the passages, hence statistical features have a high contribution in the value of the overall score.

A.5.2 The Effect of Weight Parameter w

This section investigates into how much each of contextual and semantic components contributes to the overall score. This is achieved by tuning the weight parameter w given in Equation A.6. We conducted the experiments with various w settings from 0.0 to 1.0 with a step size of 0.1. When the w is set to 0.0 the overall score is computed using only the semantic component, while w=1.0 indicates that the contribution is only from the contextual component. The rest of the values shows that the overall score is composed of both the contextual and semantic information. Figure A.9 illustrates the precision with respect to the weight parameter w, obtained by experiments carried out on computer domain data set. It can be seen from the chart diagram that the best result in terms of precision is ob-

Table A.18: The performance of SEMCON with/without statistical features

Concept	Precision (%)		Recall (%)		F1(%)	
	No	With	No	With	No	With
Computer	26.7	26.7	40.0	40.0	32.0	32.0
Software	46.7	46.7	70.0	70.0	56.0	56.0
Hardware	33.3	33.3	50.0	50.0	40.0	40.0
Web	40.0	46.7	60.0	70.0	48.0	56.0
Storage	46.7	46.7	70.0	70.0	56.0	56.0
Microprocessor	40.0	40.0	60.0	60.0	48.0	48.0
InputAndOutputDevices	26.7	33.3	40.0	50.0	32.0	40.0
Application	53.3	53.3	80.0	80.0	64.0	64.0
Windows	46.7	46.7	70.0	70.0	56.0	56.0
Average	**40.0**	**41.5**	**60.0**	**62.2**	**48.0**	**49.8**

tained when the value of weighting parameter w is set to 0.5. The precision starts declining with an increase or a decrease in the value of w. This suggests that both semantic and contextual information should contribute equally to computing the overall score as described in subsection A.3.4.

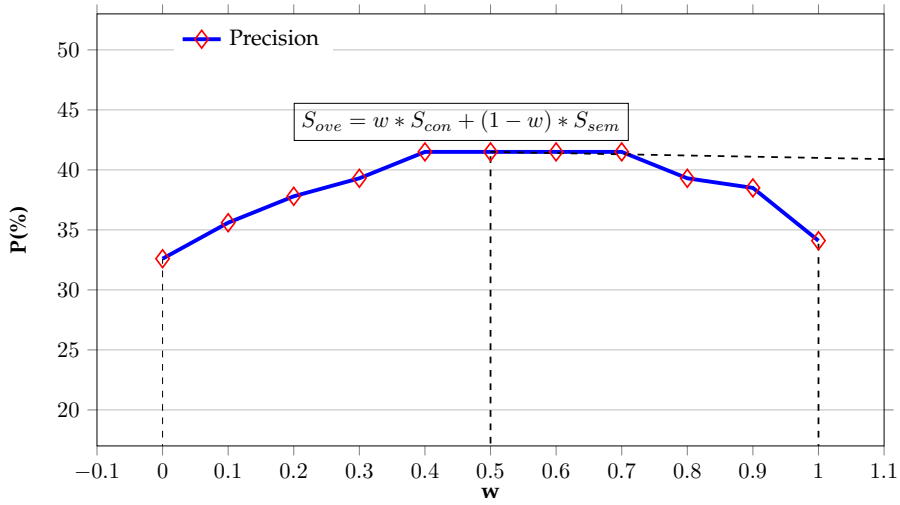


Figure A.9: Precision in function of w

A.6 Recommendations

Despite numerous recent developments in ontology engineering, successful integration of ontologies in today's applications, and automatic extraction of semantic concepts are two important problems. It is important to take advantage of a large number of existing domain-specific ontologies as creating a new ontology is a time-consuming and a laborious process. Using a common ontology is also not feasible in many cases. An ontology can be a source of domain knowledge, therefore, existing ontologies must be capitalized for further populating ontologies with new concepts. Moreover, updating ontologies with automatically extracted semantic concepts is deemed necessary for speeding up the ontology

enrichment process. However, populating ontologies automatically with correct concepts is not a trivial task. One challenging task is the extraction of a relationship between concepts. For domain-specific applications, it is also important to identify the correct sense of a concept. The success of a many large scale industrial applications and semantic web depends upon the successful integration and automation of ontology enrichment process. NLP and ML techniques can play a vital role in this regard. Some recommendations are listed below in light of the above discussion.

1. Make use of the existing domain-specific ontologies for extracting relevant concepts.
2. Choose the right input resource.
3. Use meta-models (domain-specific description) for semantic data integration.
4. Use multiple media modalities to widen the coverage of knowledge base for better concept representation.
5. Use of word sense disambiguation to identify the correct sense of a term.
6. Considering both semantic and contextual information of terms as they do contribute equally in the performance of ontology enrichment.
7. Make use of the statistical features, in addition to the frequency of terms occurrences, for deriving the context.

A.7 Future Research Directions

The knowledge resource explored by SEMCON for identification and acquisition of the relevant terminology for enriching ontologies is basically textual data. There are other knowledge resources, such as image and video which can be exploited by the system in order to identify and acquire the relevant terminology. So, one possible direction to work on in the future is extending SEMCON to exploit diverse knowledge resources including audio, images, and videos for acquiring the relevant terminology.

The two new statistical features introduced in this chapter for deriving the context proved to be useful in terms of improving the performance of the model. In this regard, the future work may further exploit other features for deriving the context and computing the observation matrix. It might be worth to investigate, in addition to the linear model, other nonlinear models, i.e. exponential, for evaluating the weight of font types and font sizes employed in this chapter.

The size of the dataset used to test the performance of the proposed model was small due to the nature of the subjective experiment. A larger dataset is required to thoroughly evaluate the effectiveness and robustness of the model.

Additionally, deep learning techniques such as embedding can be employed to automatically update ontology with new concepts by learning concepts hierarchy and relationship in existing ontology.

A.8 The Applications of SEMCON

SEMCON can be used in many application areas including but not limited to information systems, eLearning platforms, open educational resources (OER), online social network (OSN) analysis, etc., for building dictionaries, classifying documents, enriching ontologies - among many others. For instance, it can be applied for document classification in information systems where each record can be grouped into different categories automatically utilizing context and semantics. The two areas where SEMCON has been applied are as follows:

1. The classification of multimedia documents in the web-based eLearning platforms.
2. The analysis of Online Social Networks (OSNs) for identifying criminal activity and possible suspects.

These applications are discussed briefly in this section.

A.8.1 SEMCON for Web-based eLearning Platforms

Today's eLearning platforms consist of multiple media modalities including presentation slides, lecture videos, transcript files, handouts, and additional documents, delivering thousands of learning objects on a daily basis. These media provides a rich source of information that can be utilized for organizing and structuring learning objects.

Structuring and organizing huge amount of learning objects is a labour intensive, prone to errors and a cumbersome task, however. SEMCON on the other hand can prove useful in automatically organizing pedagogical multimedia content using an automatic classification approach based on the ontology described in [67]. Any new unlabeled learning object can be assigned to a predefined category in an eLearning platform using SEMCON. This is plausible by calculating the similarity between the extracted terms from the learning object and the ontology concepts. The learning object can then be assigned to a category having the highest similarity value with respect to that learning object.

An ontology represents semantic aspects of the learning objects through entities defined within a domain ontology. Therefore, each learning object that uses the ontology is represented as a vector, whose elements indicate the importance of concepts in the ontology.

A.8.2 SEMCON for OSN Analysis for Criminal Activity Detection

Analysing users' behaviour in Online Social Networks (OSNs) for investigating criminal activities is an area of great interest these days. The criminal activity analysis provides a useful source of information for law enforcement and intelligence agencies across the globe. Existing methods monitoring criminal activity normally rely on contextual analysis by computing co-occurrences of terms, which is not much effective.

SEMCON on the other hand can provide useful semantic as well as contextual information in identifying criminal activities by analysing users' posts and data, and by maintaining a history of recent user activities in the digital platforms. The proposed model [72] uses web crawlers suited to retrieve users' data such as posts, feeds, comments from Facebook, and exploits them semantically and contextually using the ontology enhancement objective metric SEMCON. The output of the model is a probability value of a user being a suspect which is computed by finding the similarity between the terms obtained from SEMCON and the concepts of criminal ontology.

A.9 Conclusion

This chapter gives an insight into the ontology concept enrichment process, present readers with an overview of state-of-the-art methods and techniques, review existing approaches and their limitations, contains related literature, and propose solutions to address some limitations of the existing systems.

It also presents a new generic model called SEMCON to enriching the domain ontologies with new concepts by combining contextual and semantic information of terms extracted from the domain documents. SEMCON employs a hybrid ontology learning approach to identify and extract new concepts. This approach involves functionalities from both linguistic and statistical ontology learning approaches. From the former approach, SEMCON utilizes morpho-syntactic analysis to identify and extract noun terms, as a part of speech, which represents the most meaningful terms in a document. While from the

latter approach, SEMCON derives the context using cosine similarity between term vectors whose members are frequencies of terms occurrences. SEMCON uses, besides term frequency, two new statistical features, i.e. term font size and font type to determine the context of a term. In addition to contextual information, SEMCON also incorporates the semantic information of terms using the lexical database WordNet and finally aggregates both contextual and semantic information of this term.

Several experiments on various small data sets are conducted, where results obtained by SEMCON are compared with results obtained by other objective methods such as tf*idf, χ^2 and LSA. The comparison showed that SEMCOM outperforms the three objective methods methods by 12.2% over the *tf*idf*, 21.8% over the χ^2 and 24.5% over the LSA. The chapter also presented experiments about the effect of statistical features on the overall performance of the proposed metric and our findings showed an improved performance. Additionally, we investigated into the amount of contribution made by each of the contextual and semantic components to the overall task of concepts enrichment. The obtained results indicated that a balanced weight between the contextual and semantic components gives the best performance.

Bibliography

[1] ABACHA, A. B., AND ZWEIGENBAUM, P. Automatic Extraction of Semantic Relations Between Medical Entities: A Rule Based Approach. *Journal of Biomedical Semantics 2*, 5 (2011), 1–8. 18, 19

[2] ABDALLA, A., AND YAYILGAN, S. Y. A Review of Using Online Social Networks for Investigative Activities. In *Proceedings of the 6th International Conference on Social Computing and Social Media* (2014), Springer International Publishing, pp. 3–12. 70

[3] ADAMIC, L., AND ADAR, E. Friends and Neighbors on the Web. *Social Networks 25*, 3 (2003), 211–230. 69

[4] AL-AZMI, A.-A. R. Data, Text, and Web Mining for Business Intelligence: A Survey. *International journal of Data Mining and Knowledge Management Process 3*, 2 (2013), 1–26. 81, 87

[5] ALAEE, S., AND TAGHIYAREH, F. A Semantic Ontology-based Document Organizer to Cluster eLearning Documents . In *Proceedings of the 2nd International Conference on Web Research* (New York, NY, USA, 2016), IEEE, pp. 1–7. 25

[6] ALTINEL, B., DIRI, B., AND GANIZ, M. C. A Novel Semantic Smoothing Kernel for Text Classification with Class-based Weighting. *Knowledge-Based Systems 89* (2015), 265–277. 5

[7] ALTINEL, B., GANIZ, M. C., AND DIRI, B. A Corpus-based Semantic Kernel for Text Classification by Using Meaning Values of Terms. *Engineering Applications of Artificial Intelligence 43* (2015), 54–66. 5

[8] ARABSHIAN, K., DANIELSEN, P., AND AFROZ, S. LexOnt: A Semi-Automatic Ontology Creation Tool for Programmable Web. In *Proceedings of AAAI Spring Symposium: Intelligent Web Services Meet Social Computing* (2012), AAAI, pp. 1–8. 17

[9] ARONSON, A. Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program. In *Proceedings of the Annual Symposium* (2001), AMIA, pp. 17–21. 19

[10] BATEMAN, S., GUTWIN, C., AND NACENTA, M. Seeing Things in the Clouds: The Effect of Visual Features on Tag Cloud Selections. In *Proceedings of the ACM Conference on Hypertext and Hypermedia (Hypertext '08)* (Pittsburgh, US, 2008), pp. 193–202. 51, 140

[11] BHAGDEV, R., CHAPMAN, S., CIRAVEGNA, F., LANFRANCHI, V., AND PETRELLI, D. Hybrid Search: Effectively Combining Keywords and Semantic Searches. In *Proceedings of the 5th European Semantic Web Conference* (2008), Springer Berlin Heidelberg, pp. 554–568. 6

[12] BINHAM, C., AND CROFT, J. Twitter Fuels Debate over Super-Injunctions, 2011. Available from: `https://www.ft.com/content/2d4ec938-7a2b-11e0-bc74-00144feabdc0`. 69

[13] BIZER, C., LEHMANN, J., KOBILAROV, G., AUER, S., BECKER, C., CYGANIAK, R., AND HELLMANN, S. DBpedia - A Crystallization Point for the Web of Data. *Web Semantics: Science, Services and Agents on the World Wide Web 7*, 3 (2009), 154–165. 26

[14] BLOEHDORN, S., BASILI, R., CAMMISA, M., AND MOSCHITTI, A. Semantic Kernels for Text Classification Based on Topological Measures of Feature Similarity. In *Proceedings of the 6th International Conference on Data Mining* (2006), IEEE Computer Society, pp. 808–812. 5

[15] BLOEHDORN, S., AND HOTHO, A. Text Classification by Boosting Weak Learners Based on Terms and Concepts. In *Proceedings of the 4th IEEE International Conference on Data Mining* (2004), IEEE Computer Society, pp. 331–334. 6

[16] BORST, W. N. *Construction of Engineering Ontologies*. Ph.D. thesis, Information Systems Department, University of Twente, Enschede, Netherlands, 1997. 13

[17] BRATSAS, C., KOUTKIAS, V., KAIMAKAMIS, E., BAMIDIS, P., AND MAGLAVERAS, N. Ontology Based Vector Space Model and Fuzzy Query Expansion to Retrieve Knowledge on Medical Computational Problem Solutions. In *Proceedings of the 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (2007), IEEE, pp. 3794–3797. 87

[18] BREWSTER, C., IRIA, J., ZHANG, Z., CIRAVEGNA, F., GUTHRIE, L., AND WILKS, Y. Dynamic Iterative Ontology Learning. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing* (2007), Association for Computational Linguistics, pp. 1–5. 18

[19] BRIN, S., AND PAGE, L. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Proceedings of the 7th International Conference on World Wide Web 7* (Amsterdam, The Netherlands, The Netherlands, 1998), Elsevier Science Publishers B. V., pp. 107–117. 84, 91

[20] BRUNZEL, M. The XTREEM Methods for Ontology Learning from Web Documents. In *Proceedings of the International Conference on Ontology Learning and Population: Bridging the Gap Between Text and Knowledge* (Amsterdam, The Netherlands, 2008), IOS Press, pp. 3–26. 18, 137

[21] CAMOUS, F., BLOTT, S., AND SMEATON, A. F. Ontology-based MEDLINE Document Classification. In *Proceedings of the 1st International Conference on Bioinformatics Research and Development* (2007), Springer-Verlag, pp. 439–452. 14, 110

[22] CASTANO, S., ESPINOSA, S., FERRARA, A., KARKALETSIS, V., KAYA, A., MELZER, S., MOLLER, R., MONTANELLI, S., AND PETASIS, G. Ontology Dynamics with Multimedia Information: The BOEMIE Evolution Methodology . In *Proceedings of the International Workshop on Ontology Dynamics* (2007), pp. 1–14. 21

[23] CASTANO, S., ESPINOSA, S., FERRARA, A., KARKALETSIS, V., KAYA, A., MLLER, R., MONTANELLI, S., PETASIS, G., AND WESSEL, M. Multimedia Interpretation for Dynamic Ontology Evolution. *Journal of Logic and Computation, Special Issue on Ontology Dynamics 19*, 5 (2008), 859–897. vii, 21, 22

[24] CASTELLS, P., FERNANDEZ, M., AND VALLET, D. An Adaptation of the Vector Space Model for Ontology Based Information Retrieval. *IEEE Transactions on Knowledge and data engineering 19*, 2 (2007), 261–272. 6, 87, 112

[25] CIMIANO, P. *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. 47, 132

[26] CIMIANO, P., AND VOLKER, J. Text2Onto: A Framework for Ontology Learning and Data-driven Change Discovery. In *Proceedings of the 10th International Conference on Natural Language Processing and Information Systems* (2005), Springer-Verlag, pp. 227–238. 17, 20, 49, 137

[27] CORCHO, O. Sharing and Reuse in Knowledge Discovery. In *The DATA Bonanza: Improving Knowledge Discovery in Science, Engineering, and Business*, Malcolm Atkinson and Rob Baxter and Michelle Galea and Mark Parsons and Peter Brezany and Oscar Corcho and Jano van Hemert and David Snelling, Ed. John Wiley & Sons, New Yourk, USA, 2013. 14

[28] CRAVEN, M., DIPASQUO, D., FREITAG, D., MCCALLUM, A., MITCHELL, T., NIGAM, K., AND SLATTERY, S. Learning to Construct Knowledge Bases from the World Wide Web. *Artificial Intelligence 118*, 1-2 (2000). 17, 21, 48, 137

[29] CROFTS, N. Implementing the CIDOC CRM with a Relational Database. *MCN Spectra 24*, 1 (1999), 1–6. 20

[30] CUNNINGHAM, H. GATE, a General Architecture for Text Engineering. *Computers and the Humanities 36*, 2 (2002), 232–254. 19

[31] DENG, S., AND PENG, H. Document Classification Based on Support Vector Machine Using A Concept Vector Model. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence,* (2006), IEEE, pp. 473–476. 87, 112

[32] DINH, D., AND TAMINE, L. Biomedical Concept Extraction Based on Combining the Content-based and Word Order Similarities. In *Proceedings of the ACM Symposium on Applied Computing* (2011), ACM, pp. 1159–1163. 110

[33] DONG, Z., DONG, Q., AND HAO, C. HowNet and its Computation of Meaning. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations* (2010), Association for Computational Linguistics, pp. 53–56. 14

[34] DRAGONI, M., DA COSTA PEREIRA, C., AND TETTAMANZ, A. G. B. A Conceptual Representation of Documents and Queries for Information Retrieval Systems by Using Light Ontologies. *Expert Systems with Applications 39*, 12 (2012), 10376–10388. 26

[35] DRAGONI, M., DA COSTA PEREIRA, C., AND TETTAMANZI, A. G. B. An Ontological Representation of Documents and Queries for Information Retrieval Systems. In *Proceedings of the 23rd International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems* (Berlin, Heidelberg, 2010), Springer Berlin Heidelberg, pp. 555–564. 24, 26

[36] DRUMOND, L., AND GIRARDI, R. A Survey of Ontology Learning Procedures. In *Proceedings of the 3rd Workshop on Ontologies and their Applications* (2008), CEUR-WS.org, pp. 1–12. 48, 131

[37] DUTHIL, B., TROUSSET, F., ROCHE, M., DRAY, G., PLANTIE, M., MONTMAIN, J., AND PONCELET, P. Towards an Automatic Characterization of Criteria. In *Proceedings of the 22nd International Conference on Database and Expert Systems Applications* (Berlin, Heidelberg, 2011), Springer Berlin Heidelberg, pp. 457–465. 17, 37, 49, 137

[38] ETZIONI, O., CAFARELLA, M., DOWNEY, D., KOK, S., POPESCU, A.-M., SHAKED, T., SODERLAND, S., WELD, D. S., AND YATES, A. Web-scale Information Extraction in Knowitall: (Preliminary Results). In *Proceedings of the 13th International Conference on World Wide Web* (New York, NY, USA, 2004), ACM, pp. 100–110. 18, 49, 136

[39] ETZIONI, O., CAFARELLA, M., DOWNEY, D., KOK, S., POPESCU, A.-M., SHAKED, T., SODERLAND, S., WELD, D. S., AND YATES, A. Unsupervised Named-Entity Extraction from the Web: An Experimental Study. *Artficial Intelligence 165*, 1 (2005), 91–134. 18, 49

[40] FAATZ, A., AND STEINMETZ, R. *Ontology Learning from Text: Methods, Evaluation an Applications* . IOS Press, Amsterdam, Netherlands, 2005. 47, 101, 131

[41] FANG, J., GUO, L., AND NIU, Y. Documents Classification by Using Ontology Reasoning and Similarity Measure. In *Proceedings of the 7th International Conference on Fuzzy Systems and Knowledge Discovery,* (2010), IEEE Computer Society, pp. 1535–1539. 111

[42] FANG, J., GUO, L., WANG, X., AND YANG, N. Ontology-Based Automatic Classification and Ranking for Web Documents. In *Proceedings of the 4th International Conference on Fuzzy Systems and Knowlede Discovery* (2007), IEEE, pp. 627–631. 24, 25, 82, 88

[43] FARD, A. M., AND ESTER, M. Collaborative Mining in Multiple Social Networks Data for Criminal Group Discovery. In *Proceedings of the International Conference on Computational Science and Engineering* (2009), IEEE, pp. 582–587. 70

[44] FELLBAUM, C. *WordNet: An Electronic Lexical Database*. MA: MIT Press, Cambridge, United Kingdom, 1998. 5, 14, 52, 142

[45] FERNANDEZ, M., CANTADOR, I., LOPEZ, V., VALLET, D., CASTELLS, P., AND MOTTA, E. Semantically Enhanced Information Retrieval: An Ontology-based Approach. *Web Semantics: Science, Services and Agents on the World Wide Web 9*, 4 (2011), 434–452. 6

[46] FRANCISCO, V., HERVÁS, R., AND GERVÁS, P. Dependency Analysis and CBR to Bridge the Generation Gap in Template-Based NLG. In *Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Text Processing* (2007), Springer-Verlag, pp. 432–443. 6

[47] FRANTZI, K., ANANIADOU, S., AND MIMA, H. Automatic Recognition of Multiword Terms: the C-value/NC-value Method. *International Journal on Digital Libraries 3*, 2 (2000), 115–130. 21

[48] FROST, H. R., AND MCCRAY, A. T. Markov Chain Ontology Analysis (MCOA). *BMC Bioinformatics 13*, 1 (2012), 1–23. 82, 84

[49] GABRILOVICH, E., AND MARKOVITCH, S. Overcoming the Brittleness Bottleneck Using Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge. In *Proceedings of the 21st National Conference on Artifcial Intelligence* (2006), AAAI Press, pp. 1301–1306. 6

[50] GANGEMI, A., GUARINO, N., MOSOLO, C., OLTRAMARI, A., AND SCHNEIDER, L. Sweetening Ontologies with DOLCE. In *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web* (2002), Springer-Verlag, pp. 166–181. 14

[51] GRUBE, E. Assault Fugitive Who Was Found Via Facebook Is Back In NY, 2010. Available from: `http://newyorkcriminallawyersblog.com/2010/03/assault-criminal-who-was-found-via-facebook-is-back-in-ny.html`. 69

[52] GRUBER, T. A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition 5*, 2 (1993), 199–220. 13

[53] GRUBER, T. Ontology. *Encyclopedia of Database Systems* (2009), 1963–1965. 13

[54] GU, H., AND KUANJIU, Z. Text Classification Based on Domain Ontology. *Journal of Communication and Computer 3*, 5 (2006), 261–272. 24, 25, 82, 88

[55] HAHN, U., AND ROMACKER, M. The SYNDIKATE Text Knowledge Base Generator. In *Proceedings of the First International Conference on Human Language Technology Research* (Stroudsburg, PA, USA, 2001), Association for Computational Linguistics, pp. 1–6. 18, 19, 49, 136

[56] HAHN, U., ROMACKER, M., AND SCHULTZ, S. medSynDiKATe - a Natural Language System for the Extraction of Medical Information from Findings Reports. *International Journal of Medical Informatics 67*, 1-3 (2002), 63–74. 20, 136

[57] HALL, M., FRANK, E., HOLMES, G., PFAHRINGER, B., REUTEMANN, P., AND WITTEN, I. The WEKA Data Mining Software: An Update. *ACM SIGKDD Explorations Newsletter 11*, 1 (2009), 10–18. 94

[58] HALVEY, M., AND KEANE, M. An Assessment of Tag Presentation Techniques. In *Proceedings of the 16th International Conference on World Wide Web* (New York, NY, USA, 2007), ACM, pp. 1313–1314. 38, 51, 140

[59] HAZMAN, M., EL-BELTAGY, S. R., AND RAFEA, A. A Survey of Ontology Learning Approaches. *International Journal of Computer Applications 22*, 9 (2011), 36–43. 48, 131, 134

[60] HEARST, M. A. Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proceedings of the 14th Conference on Computational Linguistics* (1992), Association for Computational Linguistics, pp. 539–545. 18, 135

[61] HEARST, M. A., AND SCHUTZE, H. Customizing a Lexicon to Better Suit a Computational Task. In *Proceedings of the Workshop on Extracting Lexical Knowledge* (1996), MIT press, pp. 55–69. 17

[62] HERSH, W., BUCKLEY, C., LEONE, T., AND HICKMAN, D. OHSUMED: An Interactive Retrieval Evaluation and New Large Test Collection for Research. In *Proceedings of the 7th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval* (1994), Springer London, pp. 192–201. 19

[63] HEYER, G., LAUTER, M., QUASTHOFF, U., WITTIG, T., AND WOLFF, C. Learning Relations Using Collocations. In *Proceedings of the Workshop on Ontology Learning* (2001), CEUR-WS.org. 17, 49, 136

[64] HOTHO, A., STAAB, S., AND STUMME, G. Ontologies Improve Text Document Clustering. In *Proceedings of the 3rd IEEE International Conference on Data Mining* (2003), IEEE Computer Society, pp. 541–544. 5

[65] HYVONEN, E., VALO, A., KOMULAINEN, V., SEPPALA, K., KAUPPINEN, T., RUOTSALO, T., SALMINEN, M., AND YLISALMI, A. Finnish National Ontologies for the Semantic Web: Towards a Content and Service Infrastructure. In *Proceedings of the 2005 International Conference on Dublin Core and Metadata Applications: Vocabularies in Practice* (2005), Dublin Core Metadata Initiative, pp. 219–222. 14

[66] IRIA, J., BREWSTER, C., CIRAVEGNA, F., AND WILKS, Y. An Incremental Tri-Partite Approach To Ontology Learning. In *Proceedings of the 5th International Conference on Language Resources and Evaluation* (2006), European Language Resources Association, pp. 197–202. 18

[67] KASTRATI, Z., IMRAN, A. S., AND YAYILGAN, S. Y. Building Domain Ontologies for Hyperlinked Multimedia Pedagogical Platforms. In *Proceedings of 16th International Conference on Human-Computer Interaction* (2014), Springer International Publishing, pp. 95–100. 8, 65, 156

[68] KASTRATI, Z., IMRAN, A. S., AND YAYILGAN, S. Y. An Improved Concept Vector Space Model for Ontology Based Classification. In *Proceedings of the 11th International Conference on Signal Image Technology & Internet Systems* (2015), IEEE Computer Society, pp. 240–245. 9, 110, 112

[69] KASTRATI, Z., IMRAN, A. S., AND YAYILGAN, S. Y. SEMCON: Semantic and Contextual Objective Metric. In *Proceedings of the 9th IEEE International Conference on Semantic Computing* (2015), IEEE, pp. 65–68. 8, 70, 101, 110

[70] KASTRATI, Z., IMRAN, A. S., AND YAYILGAN, S. Y. SEMCON - A Semantic and Contextual Objective Metric for Enriching Domain Ontology Concepts. *International Journal on Semantic Web and Information Systems 12*, 2 (2016), 1–24. 8, 110

[71] KASTRATI, Z., IMRAN, A. S., AND YAYILGAN, S. Y. Innovations, developments, and applications of semantic web and information systems. IGI Global, Hershey, PA, USA, 2017, ch. A Hybrid Concept Learning Approach to Ontology Enrichment. 9

[72] KASTRATI, Z., IMRAN, A. S., YAYILGAN, S. Y., AND DALIPI, F. Analysis of Online Social Networks Posts to Investigate Suspects Using SEMCON. In *Proceedings of 17th International Conference on Human-Computer Interaction* (2015), Springer International Publishing, pp. 148–157. 8, 65, 156

[73] KASTRATI, Z., AND YAYILGAN, S. Y. Supervised Ontology-Based Document Classification Model. In *Proceedings of the International Conference on Compute and Data Analysis* (2017), ACM, pp. 1–7. 9

[74] KASTRATI, Z., YAYILGAN, S. Y., AND HJELSVOLD, R. Automatically Enriching Domain Ontologies for Document Classification. In *Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics* (New York, NY, USA, 2016), ACM, pp. 1–4. 9

[75] KEIKHA, M., KHONSARI, A., AND OROUMCHIAN, F. Rich Document Rrepresentation and Classification: An Analysis. *Knowledge-Based Systems 22*, 1 (2009), 67–71. 5, 81, 87, 111

[76] KLEIN, D., AND MANNING, C. D. Accurate Unlexicalized Parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1* (2003), Association for Computational Linguistics, pp. 423–430. 19

[77] KNIBBS, K. In the Online Hunt for Criminals, Social Media is the Ultimate Snitch, 2013. Available from: `http://www.digitaltrends.com/social-media/the-new-inside-source-for-police-forces-social-networks/`. 69

[78] KONTOSTATHIS, A., AND POTTENGER, W. M. A Framework for Understanding Latent Semantic Indexing (LSI) Perfomance. *Information Processing and Management 42*, 1 (2006), 56–73. 5, 6

[79] KRUSCHWITZ, U. Exploiting Structure for Intelligent Web Search. In *Proceedings of the 34th International Conference on System Sciences* (2001), IEEE, pp. 1–9. 134

[80] LANDAUER, T. K., FOLTZ, P. W., AND LAHAM, D. An Introduction to Latent Analysis. *Discourse Processes 25*, 2-3 (1998), 259–284. 56, 146

[81] LEHMANN, J. DL-Learner: Learning Concepts in Description Logicsr. *The Journal of Machine Learning Research 10* (2009), 2639–2642. 137

[82] LENAT, D. B. CYC: A Large-scale Investment in Knowledge Infrastructure. *Communications of the ACM 38*, 11 (1995), 33–38. 14

[83] LEXISNEXIS. Social Media Use in Law Enforcement Agencies, 2014. Available from: `http://www.lexisnexis.com/government/investigations/`. 69

[84] LI, H., TIAN, Y., YE, B., AND CAI, Q. Comparison of Current Semantic Similarity Methods in WordNet . In *Proceedings of the International Conference on Computer Application and System Modeling* (2010), IEE, pp. 408–411. 38, 51, 71, 140

[85] LIU, J. N. K., HE, Y.-L., LIM, E. H. Y., AND WANG, X.-Z. A New Method for Knowledge and Information Management Domain Ontology Graph Model. *IEEE Transactions on Systems, Man, and Cybernetics: Systems 43*, 1 (2013), 115–127. 42, 56, 146

[86] LIU, W., WEICHSELBRAUN, A., SCHARL, A., AND CHANG, E. Semi-Automatic Ontology Extension Using Spreading Activation. *Journal of Universal Knowledge Management 0*, 1 (2005), 50–58. 102

[87] LOGESWARI, S., AND PREMALATHA, K. Biomedical Document Clustering Using Ontology Based Concept Weight. In *Proceedings of the International Conference on Computer Communication and Informatics* (New York, NY, USA, 2013), IEEE, pp. 1–4. vii, 24, 25

[88] LUO, Q., CHEN, E., AND XIONG, H. A Semantic Term Weighting Scheme for Text Categorization. *Expert Systems with Applications 38*, 10 (2011), 12708–12716. 5

[89] MAEDCHE, A., PEKAR, V., AND STAAB, S. Ontology Learning Part One - on Discovering Taxonomic Relations from the Web. In *Proceedings of the Web Intelligence* (2003), Springer Berlin Heidelberg, pp. 301–319. 17, 20, 49, 136, 137

[90] MAEDCHE, A., AND STAAB, S. The TEXT-TO-ONTO Ontology Learning Environment. In *Proceedings of the 8th International Conference on Conceptual Structures* (2000), Springer Berlin Heidelberg, pp. 1–4. 49

[91] MAEDCHE, A. D. *Ontology Learning fo the Semantic Web*. Springer Science+Business Media, New York, USA., 2002. 13

[92] MAHN, M., AND BIEMANN, C. Tuning Co-occurrences of Higher Orders for Generating Ontology Extension Candidates. In *ICML-Workshop on Ontology Learning* (2005), pp. 1–5. 102

[93] MANNING, C. D., RAGHAVAN, P., AND SCHUTZE, H. *Introduction to Information Retrieval*. The MIT Press, Cambridge, MA, USA, 2008. 5

[94] MANNING, C. D., AND SCHUTZE, H. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, MA, USA, 1999. 5

[95] MAZANO-MACHO, D., GOMEZ-PEREZ, A., AND BORRAJO, D. Unsupervised and Domain Independent Ontology Learning. Combining Heterogeneous Sources of Evidence. In *Proceedings of 6th International Conference on Language Resources and Evaluation* (2008), pp. 1–8. 134

[96] MCCARTHY, D., KOELING, R., WEEDS, J., AND CARROL, J. Finding Predominant Word Senses in Untagged Text. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics* (Stroudsburg, PA, USA, 2016), Association for Computational Linguistics, pp. 1–8. 125, 143

[97] MCGUINNESS, D. L. Conceptual Modeling for Distributed Ontology Environments. In *Proceedings of the 8th International Conference on Conceptual Structures, ICCS 2000* (2000), Springer Berlin Heidelberg, pp. 100–112. 47, 131

[98] MCGUIRE, M. *Technology, Crime and Justice: The Question Concerning Technomia*. Routledge, 2012. 69

[99] MILNE, D., MEDELYAN, O., AND WITTEN, I. H. Mining Domain-Specific Thesauri from Wikipedia: A Case Study. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence* (2006), IEEE Computer Society, pp. 442–448. 6

[100] MILNE, D., AND WITTEN, I. H. An Effective, Low-Cost Measure of Semantic Relatedness Obtained from Wikipedia Links. In *Proceedings of the 1st AAAI Workshop on Wikipedia and Artificial Intelligence* (2008), pp. 24–30. 15

[101] MIMA, H., ANANIADOU, S., AND NENADIĆ, G. The ATRACT Workbench: Automatic Term Recognition and Clustering for Terms. In *Proceedings of the 4th International Conference on Text, Speech and Dialogue* (2001), Springer Berlin Heidelberg, pp. 126–133. 18, 21

[102] NAGARAJAN, M., SHETH, A., AGUILERA, M., KEETON, K., MERCHANT, A., AND UYSAL, M. Altering Document Term Vectors for Classification Ontologies as Expectations of Co-occurrence. In *Proceedings of the 16th International Conference on World Wide Web* (2007), ACM, pp. 1225–1226. 6

[103] NASIR, J. A., KARIM, A., TSATSARONIS, G., AND VARLAMIS, I. A Knowledge-Based Semantic Kernel for Text Classification. In *Proceedings of the 8th International Symposium on String Processing and Information Retrieval* (2011), Springer Berlin Heidelberg, pp. 261–266. 15

[104] NASIR, J. A., VARLAMIS, I., KARIM, A., AND TSATSARONIS, G. Semantic Smoothing for Text Clustering. *Knowledge-Based Systems 54*, C (2013), 216–229. 5, 15

[105] NAVIGLI, R., AND VELARDI, P. Enriching a Formal Ontology with a Thesaurus: an Application in the Cultural Heritage Domain. In *Proceedings of the 2nd Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge* (2006), Association for Computational Linguistics, pp. 1–9. 18, 20, 23, 24

[106] NAVIGLI, R., AND VELARDI, P. Ontology Enrichment Through Automatic Semantic Annotation of On-Line Glossaries. In *Proceedings of the 15th International Conference on Knowledge Engineering and Knowledge Management* (2006), Springer Berlin Heidelberg, pp. 126–140. 18, 20, 23, 24

[107] NI, Y., XU, Q. K., CAO, F., MASS, Y., SHEINWALD, D., ZHU, H., AND CAO, S. S. Semantic Documents Relatedness Using Concept Graph Representation. In *Proceedings of the 9th ACM International Conference on Web Search and Data Mining* (New York, NY, USA, 2016), ACM, pp. 635–644. 26

[108] NILES, I., AND PEASE, A. Linking Lexicons and Ontologies: Mapping Wordnet to the Suggested Upper Merged Ontology. In *Proceedings of the IEEE Conference on Information and Knowledge Engineering* (2003), CSREA Press 2003, pp. 23–26. 37

[109] NYBERG, K., RAIKO, T., TIINANEN, T., AND HYVONEN, E. Document Classification Utilising Ontologies and Relations Between Documents. In *Proceedings of the 8th Workshop on Mining and Learning with Graphs* (2010), ACM, pp. 86–93. 14, 110

[110] OLTRAMARI, A., PREVOT, L., AND BORGO, S. Theoretical and Practical Aspects of Interfacing Ontologies and Lexical Resources. In *Proceedings of the 2nd Italian workshop on Semantic Web Applications and Perspectives* (2005), CEUR, pp. 1–16. 7

[111] ONTOTEXT. Graphdb workbench users guide, 2014. Available from: http://owlim.ontotext.com/display/GraphDB6/GraphDBWorkbench. 93

[112] PAGE, L., BRIN, S., MOTWANI, R., AND WINOGRAD, T. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report, Stanford InfoLab, 1998. 91, 92

[113] PAREKH, V., GWO, J., AND FININ, T. Mining Domain Specific Texts and Glossaries to Evaluate and Enrich Domain Ontologies. In *Proceedings of the International Conference of Information and Knowledge Engineering* (2004), , pp. 1–7. 102

[114] PEASE, A., NILES, I., AND LI, J. The Suggested Upper Merged Ontology: A Large Ontology for the Semantic Web and its Application. In *In Working Notes of the AAAI-2002 Workshop on Ontologies and the Semantic Web* (2002), pp. 1–4. 14

[115] PEDERSEN, T., PATWARDHAN, S., AND MICHELIZZI, J. WordNet::Similarity: Measuring the Relatedness of Concepts. In *Demonstration Papers at HLT-NAACL* (2004), Association for Computational Linguistics, pp. 38–41. 40

[116] PEREIRA, C. D. C., AND TETTAMANZI, A. G. B. An Evolutionary Approach to Ontology-Based User Model Acquisition. In *Proceedings of the 5th International Workshop on Fuzzy Logic and Applications* (Berlin, Heidelberg, 2006), Springer Berlin Heidelberg, pp. 25–32. 24, 26, 82, 88

[117] PEREIRA, C. D. C., AND TETTAMANZI, A. G. B. An Ontology-Based Method for User Model Acquisition. In *Soft Computing in Ontologies and Semantic Web* (Berlin, Heidelberg, 2006), Springer Berlin Heidelberg, pp. 211–229. 24, 26

[118] PETASIS, G., KARKALETSIS, V., PALIOURAS, G., KRITHARA, A., AND ZAVITSANOS, E. Knowledge-driven multimedia information extraction and ontology evolution. Springer-Verlag, Berlin, Heidelberg, 2011, ch. Ontology Population and Enrichment: State of the Art, pp. 134–166. 137

[119] PREVOT, L., BORGO, S., AND OLTRAMARI, A. Interfacing Ontologies and Lexical Resources. In *Ontologies and Lexical Resources: IJCNLP-05 Workshop* (2005), Cambridge University Press, pp. 1–12. 7

[120] QUAN YANG, X., SUN, N., ZHANG, Y., AND RUN KONG, D. General Framework for Text Classification Based on Domain Ontology. In *Proceedings of the 3rd International Workshop on Semantic Media Adaptation and Personalization* (2008), IEEE, pp. 147–152. 24, 82, 88

[121] RAGHAVAN, P. Extracting and Exploiting Structure in Text Search. In *SIGMOD Conference* (2003), ACM, p. 635. 81, 87

[122] RANWEZ, S., DUTHIL, B., SY, M. F., MONTMAIN, J., AUGEREAU, P., AND RANWEZ, V. How Ontology Based Information Retrieval Systems May Benefit from Lexical Text Analysis. In *New Trends of Research in Ontologies and Lexical Resources: Ideas, Projects, Systems* (2013), Springer Berlin Heidelberg, pp. 209–231. 17, 37, 49, 137

[123] RESNIK, P. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence Research 11*, 1 (1999), 95–130. 19

[124] RESTFB. RestFB facebook graph API, 2015. Available from: https://www.restfb.com. 70

[125] ROCHE, C., CALBERG-CHALLOT, M., DAMAS, L., AND ROUARD, P. Ontoterminology: A New Paradigm for Terminology. In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development* (2009), HAL, pp. 321–326. 37

[126] RUIZ-MARTINEZ, J. M., MINARO-GIMENEZ, J. A., CASTELLANOS-NIEVES, D., GARCIA-SANCHEZ, F., AND VALENCIA-GARCIA, R. Ontology Population: An Application for the E-Tourism Domain. *International Journal of Innovative Computing, Information and Control 7*, 11 (2011), 6115–6133. 19, 23

[127] RUIZ-MARTINEZ, J. M., MINARO-GIMENEZ, J. A., GUILLEN-CARCELES, L., CASTELLANOS-NIEVES, D., VALENCIA-GARCIA, R., GARCIA-SANCHEZ, F., FERNANDEZ-BREIS, J., AND MARTINEZ-BEJAR, R. Populating Ontologies in the eTourism Domain. In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology* (2008), IEEE, pp. 316–319. 19, 23

[128] SALTON, G., AND MCGILL, M. J. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA, 1986. 24

[129] SALTON, G., SINGHAL, A., BUCKLEY, C., AND MITRA, M. Automatic Text Decomposition Using Text Segments and Text Themes. In *Proceedings of the the 7th ACM Conference on Hypertext* (New York, NY, USA, 1996), ACM, pp. 53–65. 51, 138

[130] SANCHEZ-PI, N., MARTI, L., AND GARCIA, A. C. B. Text Classification Techniques in Oil Industry Applications. In *Proceedings of the International Joint Conference SOCO'13-CISIS'13-ICEUTE'13* (2014), Springer International Publishing, pp. 211–220. 6

[131] SANCHEZ-PI, N., MARTI, L., AND GARCIA, A. C. B. Improving Ontology-based Text Classification: An Occupational Health and Security Application. *Journal of Applied Logic 17* (2016), 48–58. vii, 6, 15, 16

[132] SCHMID, H. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *New Methods in Language Processing* (2013), Routledge, pp. 154–164. 51, 71, 139

[133] SEBASTIANI, F. Machine Learning in Automated Text Categorization. *ACM Computing Surveys 34*, 1 (2002), 1–47. 42, 56, 57, 146, 147

[134] SEKIUCHI, R., AOKI, C., KUREMATSU, M., AND YAMAGUCHI, T. DODDLE: A Domain Ontology Rapid Development Environment. In *Proceedings of the 5th Pacific Rim International Conference on Artificial Intelligence* (1998), Springer Berlin Heidelberg, pp. 194–204. 17

[135] SHAMSFARD, M., AND BARFOROUSH, A. A. Learning Ontologies from Natural Language Text. *International Journal of Human-Computer Studies 60*, 1 (2004), 17–63. 18, 19, 49, 136, 138

[136] SHANG, X., AND YUAN, Y. Social Network Analysis in Multiple Social Networks Data for Criminal Group Discovery. In *Proceedings of the International Conference on Cyber-Enabled Distributed Computing and Knowledge Discover* (2012), IEEE, pp. 27–30. 70

[137] SIOLAS, G., AND D'ALCHÉ BUC, F. Support Vector Machines Based on a Semantic-Kernel for Text Categorization. In *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks* (2000), IEEE Computer Society, pp. 205–209. 5

[138] SIRIN, E., PARSIA, B., GRAU, B. C., KALYANPUR, A., AND KATZ, Y. Pellet: A Practical OWL-DL Reasoner. *Web Semantics: Science, Services and Agents on the World Wide Web 5*, 2 (2007), 51–53. 19

[139] SONG, M.-H., LIM, S.-Y., PARK, S.-B., KANG, D.-J., AND LEE, S.-J. An Automatic Approach to Classify Web Documents Using a Domain Ontology. In *Proceedings of the 12th Asia-Pacific Software Engineering Conference* (2005), Springer Berlin Heidelberg, pp. 666–671. vii, 15

[140] SONG, M.-H., LIM, S.-Y., PARK, S.-B., KANG, D.-J., AND LEE, S.-J. Automatic Classification of Web Pages Based on the Concept of Domain Ontology. In *Proceedings of the 12th Asia-Pacific Software Engineering Conference* (2005), IEEE Computer Society, pp. 645–651. 15

[141] SONG, M.-H., LIM, S.-Y., PARK, S.-B., KANG, D.-J., AND LEE, S.-J. Ontology-Based Automatic Classification of Web Pages. In *Applied Soft Computing Technologies: The Challenge of Complexity* (2006), Springer Berlin Heidelberg, pp. 483–493. 15

[142] SPERETTA, M., AND GAUCH, S. Using Text Mining to Enrich the Vocabulary of Domain Ontologies. In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology* (2008), IEEE Computer Society, pp. 549–552. 102

[143] SY, M.-F., RANWEZ, S., MONTMAIN, J., REGNAULT, A., CRAMPES, M., AND RANWEZ, V. User Centered and Ontology Based Information Retrieval System for Life Sciences. *BMC Bioinformatics 13*, 1 (2012), 1–12. 110

[144] TOUMOUH, A., LEHIRECHE, A., WIDDOWS, D., AND MALKI, M. Adapting WordNet to the Medical Domain using Lexicosyntactic Patterns in the Ohsumed Corpus. In *Proceedings of the IEEE International Conference on Computer Systems and Applications* (2006), IEEE, pp. 1029–1036. 19, 23

[145] TSATSARONIS, G., VARLAMIS, I., AND VAZIRGIANNIS, M. Text Relatedness Based on a Word Thesaurus. *Journal of Artificial Intelligence Research 37*, 1 (2010), 1–40. 15

[146] TURNEY, P. D. Mining the Web for Synonyms: PMI-IR Versus LSA on TOEFL. In *Proceedings of the 12th European Conference on Machine Learning* (2001), Springer-Verlag, pp. 491–502. 15, 18

[147] VALARAKOS, A. G., PALIOURAS, G., KARKALETSIS, V., AND VOUROS, G. A Name-Matching Algorithm for Supporting Ontology Enrichment. In *Proceedings of the 3rd Hellenic Conference on Artificial Intelligence* (2004), Springer Berlin Heidelberg, pp. 381–389. 20

[148] VALARAKOS, A. G., PALIOURAS, G., KARKALETSIS, V., AND VOUROS, G. Enhancing Ontological Knowledge Through Ontology Population and Enrichment. In *Proceedings of the 14th International Conference on Knowledge Engineering and Knowledge Management* (2004), Springer Berlin Heidelberg, pp. 144–156. vii, 20, 21, 23

[149] VOSSEN, P. Introduction to EuroWordNet. *Computers and the Humanities 32*, 1 (1998), 73–89. 14

[150] WANG, P., AND DOMENICONI, C. Building Semantic Kernels for Text Classification using Wikipedia. In *Proceedings of the 14th ACM International Conference on Knowledge Discovery and Data Mining* (2008), ACM, pp. 713–721. 5, 6, 109

[151] WANG, P., HU, J., ZENG, H.-J., AND CHEN, Z. Improving Text Classification by Using Encyclopedia Knowledge. In *Proceedings of the 7th IEEE International Conference on Data Mining* (2007), IEEE Computer Society, pp. 332–341. 5, 6

[152] WARIN, M., OXHAMMAR, H., AND VOLK, M. Enriching an Ontology with WordNet Based on Similarity Measures. In *MEANING-2005 Workshop* (2005), Zurich Open Repository and Archive, pp. 1–6. 102

[153] WHITE, S., AND SMYTH, P. Algorithms for Estimating Relative Importance in Networks. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY, USA, 2003), ACM, pp. 266–275. 91

[154] WIJEWICKREMA, C. M. Impact of an Ontology for Automatic Text Classification. *Annals of Library and Information Studies 61* (2014), 263–272. 6

[155] WU, G., LI, J., FENG, L., AND WANG, K. Identifying Potentially Important Concepts and Relations in an Ontology. In *Proceedings of the 7th International Conference on The Semantic Web* (Berlin, Heidelberg, 2008), Springer-Verlag, pp. 33–49. 88, 125

[156] WU, Z., AND PALMER, M. Verbs Semantics and Lexical Selection. In *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics* (1994), Association for Computational Linguistics, pp. 133–138. 40, 53, 72, 104, 142

[157] XU, J. J., AND CHEN, H. CrimeNet Explorer: A Framework for Criminal Network Knowledge Discovery. *ACM Transactions on Information Systems 23*, 2 (2005), 201–226. 70

[158] YAMAGUCHI, T. Acquiring Conceptual Relationships from Domain-Specific Texts. In *Proceedings of the Workshop on Ontology Learning* (2001), CEUR-WS.org, pp. 1–6. 17, 48, 137

[159] YOO, I., AND HU, X. Biomedical Ontology MeSH Improves Document Clustering Qualify on MEDLINE Articles: A Comparison Study. In *Proceedings of the 19th IEEE Symposium on Computer-Based Medical Systems* (New York, NY, USA, 2006), IEEE, pp. 577–582. 14

[160] YOUNG, P. Optimal Voting Rules. *The Journal of Economic Perspectives 9*, 1 (1995), 51–64. 41, 56, 144

[161] ZOUAQ, A., GASEVIC, D., AND HATALA, M. Linguistic Patterns for Information Extraction in OntoCmaps. In *Proceedings of the 3rd International Conference on Ontology Patterns* (2012), CEUR-WS.org, pp. 61–72. 18, 19