DOI: 10.14267/phd.2015065

CORVINUS UNIVERSITY OF BUDAPEST

FROM TEXT MINING TO KNOWLEDGE MINING:

AN INTEGRATED FRAMEWORK OF CONCEPT EXTRACTION AND CATEGORIZATION FOR DOMAIN ONTOLOGY

Ph.D Dissertation

Supervisor: Andrea Kő Ph.D

Saira Andleeb Gillani

Budapest, 2015

DOI: 10.14267/phd.2015065

Saira Andleeb Gillani

From text mining to knowledge mining:

An integrated framework of concept extraction and categorization for domain ontology

DOI: 10.14267/phd.2015065

Department of Information Systems

Supervisor: Andrea Kő, Ph.D.

© Saira Andleeb Gillani

DOI: 10.14267/phd.2015065

Corvinus University of Budapest

Doctoral School of Business Informatics

From Text Mining To Knowledge Mining:

An Integrated Framework of Concept Extraction and Categorization for Domain Ontology

Ph.D Dissertation

Saira Andleeb Gillani

Budapest, 2015

DOI: 10.14267/phd.2015065

Contents

DOI: 10.14267/phd.2015065

DOI: 10.14267/phd.2015065

DOI: 10.14267/phd.2015065

Table of Figures

DOI: 10.14267/phd.2015065

Table of Tables

DOI: 10.14267/phd.2015065

Acknowledgements

First and foremost I would like to express my special appreciation and thanks to my advisor Assoc. Prof. Andrea Kő. It has been an honor to be her first international Ph.D. student. She has taught me, both consciously and unconsciously, how good experimental physics is done. I appreciate all her contributions of time and ideas to make my Ph.D. experience productive and stimulating. The joy and enthusiasm she has for her research was contagious and motivational for me, even during tough times in the Ph.D. pursuit. I am also thankful for the excellent example she has provided as a successful woman researcher and professor.

The research reported in this thesis was supported by Prokex project (EUREKA_HU_12-1-2012-0039), in cooperation with the Corvinno Technology Transfer Center, so I praise the enormous amount of help and teaching by András Gábor who is the director of Pokex project. His genuine caring and concern, and faith in me during the dissertation process enabled me to do this research. He provided insight and expertise that greatly assisted the research. I could not have succeeded without the invaluable support of András Gábor.

I would also like to thank Assoc. Prof. Dr. György Strausz and Dr. Szabó lldikó for serving as my dissertation plan reviewers even at hardship. I also want to thank you for your brilliant comments and suggestions, thanks to you.

A special thanks to my family. Words cannot express how grateful I am to my mother and father for all of the sacrifices that you've made on my behalf. Your endless love, patience, understanding, compassion and prayers for me was what sustained me thus far. I also thank to my brother, Aurang Zeb Gillani and sister, Humaira Gillani for encouraging me throughout this experience. I love them very much.

Finally I thank my ALLAH, for letting me through all the difficulties. I have experienced Your guidance day by day. You are the one who let me finish my degree. I will keep on trusting You for my future. Thank you, Allah.

DOI: 10.14267/phd.2015065

Introduction

1.1 Motivation

Due to the ongoing economic crisis, the management of organizational knowledge is becoming more and more important. This knowledge resides amongst other in knowledge repositories, in business processes and in employees' head. Knowledge repositories contain explicit knowledge while employees have tacit knowledge, which is difficult to extract and codify. Business processes have explicit and tacit knowledge elements as well. Nowadays the efficiency of business processes has become one of the major motivating forces for sustainable businesses. The efficiency can be improved by increasing those people's knowledge who are involved in causing the poor efficiency of business activities.

However, this task poses many challenges, which are summarized as: a) how to effectively find those situations in which employees' performance appear weak and b) how to improve the organizational knowledge especially employees' knowledge by providing the most appropriate learning and/or training materials c) how can we ensure that the knowledge in business processes are the same as in knowledge repositories and employees' head.

However, most regulatory, social and economic environments are complex and changing continuously, causing increased demand for employees in getting the necessary job-role knowledge in right time and right format. Employees have to follow these changes, improve their competencies according it.

More and more knowledge (procedural and declarative) are necessary to run any business. This knowledge is the intellectual capital (intangible asset) of the companies. This intellectual capital resides mostly in human resource, partly in their head, partly in their hand (knowledge, skill).

DOI: 10.14267/phd.2015065

One of the main goals of the organizations is to codify that tacit knowledge in order to preserve it and enrich intellectual capital. The major challenge is that such knowledge is very volatile and scarce (with the hope that the human resource will come back tomorrow). Secondly, the issue is how we have to articulate the knowledge, which knowledge representation is the best and how can we make the knowledge transportable (transfer from one head to another). Even though the majority of regulations may now available in digital form, but due to their complexity and diversity, identifying the ones relevant to a particular context is a non-trivial task. Another important knowledge source is processes, which contain rich, but in many cases embedded or ill articulated, in some cases hidden knowledge. This is a known problem, but yet, there is no any absolutely safe and sound solution to solve this type of problems.

The research reported in this thesis proposal is connected to a two year project “PROKEX”. This PROKEX project (EUREKA, 2013), aims to develop a new knowledge management solution addressing the above mentioned problems and challenges. This solution will provide an integrated platform for process-based knowledge extraction. The purpose of this development is to create an knowledge transfer (often called 'e-learning') solution, what perceive the need for learning, in a real and active work environment (in real-time with processing past events), by monitoring the effectiveness of the work, and if necessary, according to the role of the worker in the process and his/her personal characteristics, it will give him/her most adequate learning material. To achieve this goal, my research problem is dedicated to the following four main areas and challenges:

1. Knowledge extraction from business processes
2. Enhancement in existing knowledge by using text mining techniques

DOI: 10.14267/phd.2015065

3. Develop a new semantic similarity measure to filter irrelevant concepts to keep precision high enough

4. Develop a new method for categorization of knowledge elements.

1.2 Problem Statement and Research Questions

Organizations' business processes are organized according to their strategic, core and support functions. Processes are implemented in organizations and these processes reflex different organizational functions. Behind of a process, there are many architectures and infrastructures of information technology are involved. As it is mentioned earlier, processes are the main sources of organizational knowledge. Processes and their embedded knowledge have significant role in providing a suitable support for employees and at the same time they help to offer improved quality in public services, they facilitate to manage the complexity of the environment. Here, we claim a very important assumption: the organizational knowledge is very much process dependent, or more precisely, task dependent. In our approach, a task is the smallest building block of a process, as long as human resources are allocated to tasks for execution. In this way we can say knowledge is not only process dependent, but rather it is task dependent. Therefore, we can say, in a business process model, competencies (combination of knowledge, skill and attitude) or required knowledge assigned to different tasks. The requirements are usually summarized in the job-role description and stay on a general level. However, the embedded knowledge remains hidden in many cases. To extract the task and process related knowledge means we have to find what pieces of knowledge belongs to a given task. This extraction of knowledge should be in such a manner that it can enhance the existing organizational knowledge base, which is formulated in many cases in domain ontology. A process model is an activity-oriented description while domain ontology is a concept oriented

DOI: 10.14267/phd.2015065

one, but both have to contain the same knowledge elements. In point of fact, the internal organization of the two representation types is different from internal logic point of view because a concept hierarchy means that there is overall concept, many sub-concepts and these are related to one another on a logical basis while the representation of a process is task (flow) oriented. Therefore, from the process point of view it is logical to apply a same concept to several times in a process while from an ontology point of view it is redundant.

The main research issue here is how to populate the domain ontology from a process flow due to the fact that both have different structures. Different solutions are given in literature to extract such knowledge but my research area is dedicated to the challenges of knowledge extraction from business processes through text mining techniques in order to populate, to enrich organizational knowledge base in a systematic and controlled way. In my approach, domain ontology represents the organizational knowledge. By the articulation of organizational knowledge the foreseen solution becomes an efficient tool to manage intellectual capital of the organization.

According to these research challenges, my first research question is investigating the solutions to extract knowledge from process to enrich existing domain ontology.

Research Question 1: How can we use text mining to extract knowledge from processes in order to enhance or populate the existing ontology?

Answering this question starts with clarifying how can we articulate the hidden information/knowledge from business processes. I will review theoretical foundations of related fields, like semantic business process management, process mining and text mining.

DOI: 10.14267/phd.2015065

In literature, process mining is usually used to construct or learn about the nature of a process which is hidden or unknown for the observer. Process mining investigates that what tasks, in what order follow each other, what events trigger which task or connecting process and it deals with event log data (structured data) of business process with the aim of process analysis and improvement. In my thesis emphasis is given to extract information/knowledge from the processes and enrich this organizational information/knowledge by such a paradigm that can disclose the concealed information and interesting patterns by means of methods which automatically handle the processes' data. As mentioned in the introduction, a business process splits into tasks and a task has a number of attributes such as description, responsibility, execution related information (order, triggers, and events) and resources to be used. The overall question to be answered in this thesis: what to do in order to achieve the output (new concepts for ontology population) from the input (process' description)? The emphasis is on the description attribute of a task because it contains explicit and tacit knowledge elements in an embedded way. My research concern is to identify and mine the hidden knowledge elements and put them into a context (domain ontology) to make it knowledge. Text mining is a method or approach of the extraction of new, heretofore unknown information from any text. Through text mining, description can be extracted from a task and then this description can be used to create such a meaningful list or set of concepts which can be applied as an input for contextualization. There are several steps that uses different text mining techniques to change/convert this description into a form which will able to be contextualized in the form of ontology. The overall purpose of this dissertation is to propose a solution to transform embedded knowledge of processes in such a form that will be mapped on an ontology. To use process information for domain ontology population, there is a need to perform advanced analysis of

DOI: 10.14267/phd.2015065

the description (in text form) associated with each activity in log. As it looks from the brief description, the nature of the process modelling is rather ***procedural,*** while many questions raised by the modelling need answers on contextual basis, where the context has a rather ***declarative*** nature (ontology). This declarative nature is provided in my research by Studio ontology(Vas, 2007). Studio ontology is a domain ontology developed by Corvinno Technology Transfer Center (Corvinno Center, 2011). Studio ontology has a certain meta structure, the main meta classes are knowledge area and sub knowledge area, basic concept. Until now 1500 concepts are in the ontology. The core knowledge area is business informatics.

The text mining application with its simple or sophisticated procedures bridges, connect the two different approaches, processes the concepts and transport them to the ontology for enhancement of ontology. The purpose of ontology building and enhancement is not for the sake of own, but provides the contextual background (it is often referred as business logic) what is necessary to the process modelling (later improvement and optimization). Therefore, to infer useful insight from this description of a task, text mining can be the best solution. Thus, in this thesis, my emphasis is to search such a text mining solution that can dig out hidden information from business tasks.

The underlying motivation drives the research to create a text mining framework that can extract this information from business processes as well from domain related available documents and transform this information into knowledge.

Text and data mining methodologies provide the methods and tools to identify key concepts mainly on statistical basis. Another challenge lies to provide a cyclic incremental **semantic**

DOI: 10.14267/phd.2015065

process that will extract information from the process to enrich existing ontology and after ontology enrichment also enrich these organizational processes as well.

Answering the following research questions will help to answer the first one.

Research Question 2: what methodology and concept extraction methods are available to enhance the existing knowledge that captured from the business process? Whether a text mining based solution can be used for ontology learning in the context of machine learning?

To answer this question, different existing information extraction (IE) methods will be analyzed. It is difficult to use existing information extraction methods to enhance the existing knowledge of business process for domain ontology learning. It becomes more appalling when we also want to use this enhanced information again to enrich these business processes. Therefore, there is an increasing need to propose some method or approach in order to automate extraction of information from business processes and enhance this extracted information in such a way that can contribute to enrich existing domain ontologies. Devising such a mechanism for information extraction from business process is challenging due to different nature of both domains (process modeling and context as in our case a domain ontology). It is also challenging to provide a concept extraction mechanism that can provide reasonable performance in terms of precision and recall as well as show good quality in providing a precise, shared, and well-founded, distinction between the classification of a term in an individual or in a concept. Another limitation is the description of business process tasks is not sufficient to grasp all domain specific concepts.

DOI: 10.14267/phd.2015065

Already developed information extraction tools can, with reasonable accuracy, extract information from text with somewhat regularized structure. Wilks defined (Wilks, 1997) an information extraction method, which extract some specific information from natural language texts into predefined, structured representation, or templates. These IE systems identify the instances of facts names/entities, relations and events from semi-structured documents or semi-structured text[1]. For example, programs that read in resumes and extract out people's names, addresses, job skills, dynamic event tracking and so on, can get accuracies in the high 80 percent. The overall goal of such tools is to present overall trends in textual data. To find relationships between clinically important entities, a system is proposed (Abdel-moneim, Abdel-Aziz, & Hassan, 2013). This system performs two major tasks: first from patient narratives, it identifies the relationships between important entities while the second task is to apply statistical machine learning (ML) approaches to extract relationship. In this technique, authors proposed a system for Clinical Relationships extraction techniques. Some advanced IE systems are also proposed with the help of some outsources such as Gelfand et al. (Gelfand, Wulfekuler, & Punch, 1998) have developed a method based on the Semantic Relation Graph to extract concepts from a whole document. With the help of lexical knowledge base, they created a directed Semantic Relationship Graph (SRG) by identifying relationships between words. But this system also needs an initial list of words on which SRG will develop. Thus, the major problem with these approaches is that we have to know in advance what kind of semantic

[1] In this thesis, for text I have used term semi-structured not unstructured because its structure implicitly reflects a schema and according to Herbert Simon, If no (hidden) structure exists, no chance to discover anything. One of major goals of this dissertation is to explicitly recover meaningful structures (concepts) in semi-structured text corpora.

DOI: 10.14267/phd.2015065

information we are looking for therefore, it is difficult to find new knowledge without any prior predefined structure.

In contrast, for this thesis, my purpose is to extract new knowledge in the form of new concepts from business processes to enrich domain ontologies as the domain ontology formally represents concepts and the relations between them. Therefore, I can't rely on such already developed IE extraction tools. However with the help of text mining automatic knowledge extraction systems can be developed because as aforementioned above text mining discovers new pieces of knowledge. In literature there are many tools developed that first extract information from plain text and then applied different text mining techniques on this extracted information to fetch hidden knowledge. To find interesting relationships in text (Nahm & Mooney, 2002) presented a text mining framework, DxscoTEX (Discovery from Text EXtraction). This framework used the information that is extracted through IE module and this module used a template which specifies a list of slots to be filled. Karanikas et al. (Karanikas, Tjortjis, & Theodoulidis, 2000) used IE for term and event extraction and then applied text mining algorithms to label the documents, in order to cluster them. For labeling they used extracted terms and events. Concept extraction is a sub task of text mining (Sarnovský, Butka, & Paralič, 2009) in which the main focus is on extraction of ideas or concepts rather than information as in IE. A variety of text mining approaches have been devised to address the concept extraction problem. Successful techniques include rule based (Mykowiecka, Marciniak, & Kupść, 2009; H. Xu et al., 2010) and statistical methods (Bunescu et al., 2005; F. Wu & Weld, 2007). There are different ontology learning approaches such as Text2Onto tool (Cimiano & Völker, 2005), OntoLT (P Buitelaar, Olejnik, & Sintek, 2003), OntoBuilder (Gal, Modica, & Jamil, 2004), DODDLE-OWL (Morita, Fukuta, Izumi, & Yamaguchi, 2006) and

DOI: 10.14267/phd.2015065

OntoGen tool (Fortuna, Grobelnik, & Mladenic, 2007) were developed which semi automatically extract concepts for ontology. However, the quality of extracted concepts is low in the previously mentioned solutions as those concepts do not represent the domain well and mostly approaches used traditional ranking metrics (e.g., TF-IDF) thus, they do not show promising results. There is need to use some latest text mining approaches to extract and rank the new concepts.

I deal aforementioned issues of existing approaches. However, to my knowledge, there have been no studies done to address in connecting text mining to process management in context of extracting new concepts of business tasks to enrich domain ontology (defined in this thesis) for ontology learning. This proposed automatic information extraction method will comprise two basic phases: In a first phase, system will extract information from business process and in second phase it will enhance the extracted information with the help of other sources such as WordNet, Wiktionary and corpus and this enhanced information will enrich the root ontology.

Research Question 3: A modified semantic similarity measure will improve significantly the efficiency and quality of a domain ontology enhancement/ enrichment.

In this thesis, the main goal is the concept extraction from unstructured data (in form of text) of business processes. In literature, some general frameworks are introduced for concept extractions which are less flexible and adaptable. These ontology engineering frameworks rely on shallow NLP techniques for concept extraction. Their emphasis is not on contextual information extraction. Consequently, they neglect to handle semantic phenomena and resulted concepts from such schemes are overly cosmopolitan. From a user perspective, concepts in a context become more meaningful to take decisions. Some state of the art techniques (Ercan &

DOI: 10.14267/phd.2015065

Cicekli, 2007; Gelfand et al., 1998; Hassanpour, O'Connor, & Das, 2013; Kok & Domingos, 2008) proposed for semantic concept extraction. A semantic similarity detection technique can allow additional matches to be found for specific concepts not already present in knowledge bases. It is believed that measures of semantic similarity and relatedness can improve the performance in form of quality of such systems. However, these past semantic-based methods, fall short in resolving the main issue: helping users to identify specific concepts related to any business process, not just the presence of domain concepts, within relevant text. The difficulty of semantic similarity is increased when there is a reduced quantity of text like in my case where business process' have not enough domain related data. Therefore, semantic concept extraction is still an open issue in ontology construction and there is a need to implement NLP and text mining techniques in more detail. I propose a new semantic similarity measure which will help in concept extraction and that will overcome the problems of existing semantic similarity measures.

Research Question 4: whether taking top categories from the existing ontology will improve the result of text mining solution to help ontology enrichment process or not?

Manual ontology population and enrichment is a complex and time-consuming task that require professional experience involving a lot of expert discussions and efforts. In this thesis, my concern is to propose a semi-automatic solution for ontology population and enrichment. Ontology enrichment is the task of extending an existing ontology with additional concepts and relations and placing them at the correct context in the ontology. Ontology population, on the other hand, is the task of adding new instances of concepts to the ontology. The process of ontology population does not change the structure of an ontology, i.e., the concept hierarchy

DOI: 10.14267/phd.2015065

and non-taxonomic relations are not modified. What changes is the set of realization (instances) of concepts and relations in the domain. Ontology learning, enrichment and maintenance is an ongoing and complex process, with several challenges (Shamsfard & Abdollahzadeh Barforoush, 2003; Wong, Liu, & Bennamoun, 2012; Zouaq, Gasevic, & Hatala, 2011). It has a key role in ontology management; it tackles the issues to turn facts and patterns from the content into shareable high-level constructs or ontologies. Various approaches based on information extraction methods have already been used for ontology population. An interesting aspect of ontology population, which is not addressed adequately in the literature, is the handling of redundancy. If an ontology is populated with an instance without checking if the real object or event represented by the instance already exists in the ontology, then redundant instances will be inserted. Therefore, consistency maintenance or redundancy elimination are main issues in ontology population. My proposed solution will handle all these issues by using lexical resources.

1.3 Research Objectives

The context of my research is the maintenance of the organizational knowledge. There is an explicit form of this knowledge in knowledge repositories, ontologies and a tacit one in the head of employees. A special knowledge representation form is the business process, which contains knowledge in embedded way. My text mining based knowledge management solution help to connect these knowledge representation forms to each other. One of the key contributions of my text mining solution is to automate the whole process, namely knowledge extraction from business process and to facilitate domain ontology enrichment If there is any alteration in a process or in the knowledge required by the task we throw to modify only the process on high level, and this new or changed knowledge can be easily presented to the employees.

DOI: 10.14267/phd.2015065

The aim of the thesis is represented by the identification of methods and techniques used for knowledge extraction from organizational process and domain related documents available in digital form.

This research underscores the importance of concept extraction from business processes and ontology development as knowledge representation. Text mining solutions identifies the relationship between private activities and job-specific knowledge, soft skills (usually called competencies). The social system of job competencies, and organizational structure and business processes can be paired with each other. Relations between concepts which are identified through association rules constitute the basis for ontology evolution. Thus, the output from this text mining framework is expected to be helpful in the maturation of an ontology which will constitute the foundation of the content structure. So, this knowledge will hold employees to easily take their job-role specific knowledge.

1.4 Research Contributions

The main contributions of this dissertation are fourfold.

- The first key contribution of this thesis is to provide a generic text mining solution/framework that build bridges between two different approaches; process modeling that is procedural in nature and context/ontology that is declarative in nature.
    - The major contribution is concept extraction and enrichment with the help our resources such as WordNet, Wiktionary and domain corpus. The state-of-the-art text mining and NLP techniques are used in information extraction solution.
    - Developing of a generic method of discovering useful knowledge in terms of single key term and compound terms to some key issues discussed in the textual databases.
    - For concept filtration, I proposed a new method which is a combination of statistical and

DOI: 10.14267/phd.2015065

semantic measures. Through this process more semantic and contextual concepts of any sphere can be elicited from a given text data. This proposed method is a hybrid similarity measure.

- Another important contribution is to design a concept categorization method for ontology population.
- The proposal of novel integration of text mining techniques to capture knowledge elements from organizational processes and disseminate these elements in terms of new concepts in domain ontology.

Besides these primary contributions, some algorithms are also advised to make concept learning more effective. An algorithm is designed for compound word extraction from text. Another algorithm of stemming is also aimed to sweep over the restrictions of the Porter stemming algorithm.

1.5 Research Methodology

Despite a lack of consensus on the research methodologies to be used in the domain of IT (Information Technology) and computer science, a large number of thesis or dissertations are exploring and using various methods to perform research in these disciplines. Both are increasingly broad and diverse fields. Research in these fields combine aspects of different sciences such as it uses mathematical reasoning to prove properties of the proposed system, engineering methodologies to design a solution of a practical problem, and the empirical approaches (Qualitative and Quantitative Methods) of the scientific method. Therefore, in the process of writing this thesis and carrying out research, I had to follow the pattern of traditional research methodologies which is the requirement of the PhD School as well as explored models to perform the research in the domain of computing. A complete research process in solvable tasks has been defined in this thesis and these tasks are time depended and measurable. In order

DOI: 10.14267/phd.2015065

to achieve these solvable tasks, I had to define the problem statement and research questions instead of devising hypothesis.

The Business Informatics Ph.D. School of Budapest Corvinus University belongs to the doctoral schools of social sciences in the university and has been classified to the IT discipline as well, therefore applying research methods in a kind of 'hybrid' way can hopefully be considered to be accepted.

1.5.1 Fundamental of social science and computer science research

There are four basic categories of science: natural sciences that includes physics, chemistry and biology, formal sciences (mathematics), social sciences such as economics, psychology and sociology and applied sciences. Computer science combines several branches of science such as mathematics, engineering and soft science. Rather, Information Technology (IT) is the application of computer programs in various disciplines including but not limited to business, health, education and transportation. IT relates to social sciences.

Scientific research is a scientific systematic inquiry to establish facts about a particular question or a problem. The purpose of research is to explore new and innovative aspects of any branch of knowledge and postulate theories. One another purpose is to prove already discovered unproved theories. This scientific inquiry may be in form of theoretical research or empirical research. In theoretical research, theoretical concepts are developed about natural or social phenomenon while in empirical research testing is performed on theoretical concepts. Empirical research methods can be divided into two main categories: qualitative and quantitative. Research in computer science is of both types, depends on the nature of the problem. Sometimes, it is theoretical such as to investigate a complex theory or to design and analyze an

DOI: 10.14267/phd.2015065

algorithm. And sometimes this research can be empirical that involves experiments, design, implementation, and testing.

1.5.2 Qualitative and Quantitative Research

Choice of research methodology; qualitative, quantitative or their combination is closely tied with nature of research problem, research design and approach used towards the solutions. Main purpose of the quantitative research is the measurable quantification of organized data. It focuses mainly on computation and classification of features, statistical models and numbers to explain the findings. It also generalizes the outcome from a small sample to a larger population and during the process various parameters are computed and analyzed. Tools can be used to gather data including survey, automated application, questionnaire and measurements via equipment. Quantitative research is objective in its nature because it tend to work with exact numerical measurements and detailed analysis is done to find answers of research problem under consideration.

Qualitative research deals mainly with gathering of rhetorical data instead of numerical measurements and the analysis is done for interpretations of that data for the purpose of exploration. Qualitative research is more suitable to gain detailed understanding about a new research problem in early stage and used to lay a solid basis for quantitative research to follow up. Data gathering is done by the researchers themselves. Qualitative methods are subjective in their approach and try to understand the human behavior and rationale behind that behavior.

Due to the recent emergence of new sciences and disciplines we now have more diverse research problems that overlap many different fields. To triangulations are used to address such complex research problems. Triangulation refers to the process of combining multiple research

DOI: 10.14267/phd.2015065

methodologies to study a complex research problem. Cohen and Manion defined triangulation as an attempt to map out, or explain more fully, the richness and complexity of human behavior by studying it from more than one standpoint.

1.5.3 Research based on case studies

Case study involves an in-depth behavioral analysis of one or more characters under observations. A case study present truthful, diverse, subjective and highly contextual situations of subjects and also presents problems that are faced by these subjects. Professor Paul Lawrence defined a good case study as:

"The vehicle by which a chunk of reality is brought into the classroom to be worked over by the class and the instructor. A good case keeps the class discussion grounded upon some of the stubborn facts that must be faced in real life situations."

In prospective case studies one or more individuals are observed for behavioral analysis and conclude outcomes but in retrospective case study, historical information is analyzed to find out reasons to support a particular outcome. There are many different sources for collecting data and may involve (Yin-1994, Stake-1995):

Observations: Refers to the process of observing one or more subjects in real life setting. There is no limitation on number of observer being used, that may be one or group of observers.

Documentations: Studies based on documented matter that may include but not limited to applications, papers, letters, processing records, surveys and newspapers.

DOI: 10.14267/phd.2015065

Interviews: Most commonly used tools for real life data gathering for studies on subjects and it may include structured or un-structured questions.

Physical artifacts: Refers to the study of observing any historical or cultural object.

Researcher observation: Researchers act as participants to gather information on outcomes.

Archival records: Involves summary of previously done surveys, census records etc.

There are many advantages of using case studies as research tool: It gives a bigger picture of the overall situation and provides an in depth and thorough understanding of the problem under study. Uniqueness if using case study method lies in the fact that it reveals relationships that may not be revealed by using any other method (Babbie-89, Galliers-92). Bensabatet et al. highlights potential features of case study research strategies as follows:

- Observation of subjects in their natural setting
- Using multiple methods for data gathering
- Observation collection on one or more subjects
- Exploration of complex situations
- Analysis without manipulated experimentation
- Without independent and dependent variables
- Analysis dependency on researchers ability of perceive
- Nature of normal phenomenon of a routine procedure

Case study methodology has some strength over other methods. Case study allows a detailed data collection that surely gives an in depth understanding, such a detailed data gathering may

DOI: 10.14267/phd.2015065

not be possible using other methods specifically experimentations. It is quite effective in problems where there are very less number of subjects available for study and it is rare to find subjects and it also allows conducting scientific experiments within case study. Some of the disadvantage of case study include that outcome may not be easily generalized on larger population keeping in mind the fact that it was gathered on very small subjects. Other drawback includes many case studies not being scientific and it is also difficult to draw any clear effect from case study.

1.6 Proposed Research Methodology

To solve afore mentioned research problem, I have developed such a research methodology that consists of existing processes and findings from design research. This defined process combines qualitative and quantitative research methods. The main design research phases applied in this thesis are as follows;

- **Problem Awareness:** This involves reviewing the literature to analyze the existing techniques of knowledge extraction and ontology enrichment. This phase also confirms the lack of a general framework that can automatically extract knowledge from organizational processes and then after enriching this information use it for ontology enrichment.
- **Extensive Literature Review:** in this phase a thorough literature study took up and identification of unresolved publications of relevant subjects. The purpose of this phase is a detail study about knowledge extraction methods ontology enrichment and concept categorization is a part of this phase and patterns and ontology learning and also level out some issues of these arenas. Nevertheless, the research of this dissertation is part of an ongoing project

DOI: 10.14267/phd.2015065

(project (PROKEX project, EUREKA_HU_12-1-2012-0039), so it is intended to keep exploring different techniques/domains to cope with changing requirements/conditions.

- **Developing Research Questions:** After a through literature survey, the next step is to develop research questions from the general purpose statement. The focus is narrow down to specific questions. These questions to be answered in later study. This is very important phase because these questions act like driving force behind the research from beginning to end. Four research questions have been developed in this study in which first question RQ1 is the central question and other three are associated sub-questions.
- **Conceptual Framework:** conceptual framework help to clarify and map out the key research issues in the research area. A tentative idea is produced in this phase for further research. This conceptual mapping is in a pictorial form. Here, it is decided that how to select suitable preprocessing techniques on extracted data from processes and how to apply knowledge extraction techniques. This idea should suggest that how can we use text mining techniques for information extraction, how we can use different machine learning techniques for concept enrichment by using different out resources and how to filter out irrelevant concepts by using different statistical or semantic measures. An initial conceptual framework has been projected after a bit of brainstorming and reviewing sessions.
- **Development:** The development of the solution will be achieved by building the design artefact. Here, intend is to develop a working prototype using JAVA that will work as a module within the said project. By immersing in the build activity of this prototype the understanding of the problem becomes more clarify and new suggestions come to mind that helps to improve the next build and evaluate cycle. I have implemented first build with basic modules of our conceptual framework that is for knowledge extraction by using JAVA.

DOI: 10.14267/phd.2015065

- **Evaluation:** in this phase an assessment method is to develop to assess the quality and effectiveness of the designed artefact (March & Smith, 1995). The purpose of evaluation phase is to consider that what actually occurred. Whether the development met the expected results or not and also check the links between the program as it was delivered and the outcome of the program (Balbach, 1999). The proposed framework, ProMine, is evaluated for coverage of the domain and for accuracy. Both qualitative and quantitative evaluation methods have been established to quantify the performance of the proposed framework. For qualitative evaluation method I have selected three cases from three different domains: 1) Food Chain Safety, 2) Insurance, 3) IT audit. The understanding of selecting these domains is the Byzantine complexity of interrelated tasks and the problems occurring during their everyday performance. I will also compare my framework performance through controlled experiments with existing enterprise ontologies in which one is manually engineered ontology (STUDIO).
- **Conclusions:** This is the final phase of the Design Research cycle that covers the overall contribution made by the research. In literature, this phase is given different names like conclusion, results analysis or communication. Conclusions should appropriately supported by evidence. Limitations of my proposed solution and future work will also present in this research part.

DOI: 10.14267/phd.2015065

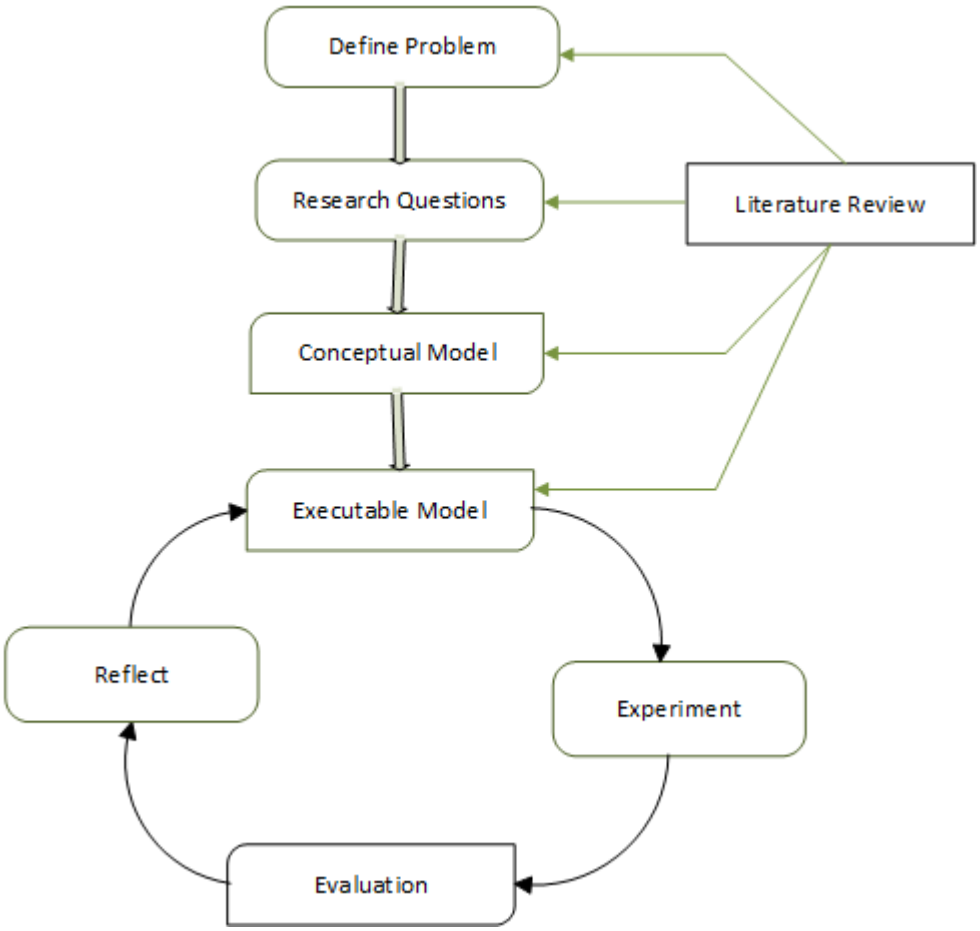

Figure 1-1 The Research Cycle

By following above mentioned research method, the initial prototype (ProMine) is prepared. In order to deliver the final version of ProMine, Design Science approach (March & Smith, 1995) have been used . According to this approach, there are two basic activities, build and evaluate; build is the process of constructing and artifact for a specific reason and evaluation is the process of determining the performance of the artifact. In this research, these activities are executed in an iterative incremental Design Research manner consisting of number of iterations as follows:

 Iteration 1 – Core framework development is done including key term extraction technique and concept enrichment module by using WordNet. For concept filtering, statistical measure,

DOI: 10.14267/phd.2015065

information gain is used. Evaluate the technique and tool by using real dataset and evaluating the extracted concepts with the identified evaluation metrics.

 Iteration 2 – For good results of concept extraction, extending the framework and add Wiktionary along with WordNet. For concept filtering, I will devise a semantic measure with the combination of statistical measure. This will improve results in better form. A link is also created between extracted concepts and business processes.

 Iteration 3 – In third iteration, I will improve the concept extraction technique by modifying compound word selection method and the main development of this iteration will be concept categorization module that will be prepared by using seed ontology. This categorization module will help in ontology enrichment.

 Iteration 4 – Validate the framework by applying and evaluating the extraction method across other domains. The generality of the framework and ProMine tool will be demonstrated through comparing evaluation measures for three different domains.

Through this iterative procedure of the text mining prototype, three real case scenarios have been run to illustrate the effectiveness and provide a live proof of the proposed method (ProMine) and as the means by efficiencies and improvements are identified. Determining whether progress is made by the extraction method and tool is evaluated by applying the appropriate metrics from the knowledge base to measure the accuracy and coverage of the learned domain ontology model.

DOI: 10.14267/phd.2015065

Determining whether progress is made by the extraction method and tool is evaluated by applying the appropriate metrics from the knowledge base to measure the accuracy and coverage of the learned domain ontology model.

Table 1-1 Research questions, aspects of framework, research process and research methods

Research Questions	**Aspects**	**Research Process**	**Methods**
RQ1	Actions: acquiring information, analyzing information, analyzing information extraction from processes	Analysis	Literature analysis
RQ2	Tools: *NLP, TM, External resources: WordNet & Wiktionary, analysis tools*	Analysis, Design, Development	Literature analysis, Exploratory software development
RQ3	Tools: *NLP, TM, ML*	Development	Literature analysis, exploratory software development
RQ4	Actions: concept categorization, developing algorithms	Development	Literature analysis, exploratory software development

DOI: 10.14267/phd.2015065

	Tools: NLP, TM		

DOI: 10.14267/phd.2015065

Literature Review

In this chapter, relevant literature, as pertaining to the aim of this study, is systematically analyzed to offer the necessary background and terminology as the foundation for the rest of this thesis. Literature was acquired by different approaches of text extraction. As it is mentioned in problem statement that the main purpose of this research is to extract knowledge elements from organization processes, so in literature review Business Process Management (BPM) is also discussed with respect to information extraction. Here, the difference between information retrieval and information extraction is also described. A review of the existing literature about similarity measures and techniques has been conducted before proposing a new hybrid similarity measure. I thoroughly reviewed existing approaches and methodologies that focus on knowledge extraction, ontology learning and ontology enrichment. I also talked about algorithms and software that are being applied for detection of useful pattern of interest from text streams. The research community is exercising to develop many text mining approaches for text analysis and knowledge extraction.

2.1 Business Process Management (BPM)

Business Process Management (BPM) is a systematic approach that combines the knowledge from information technology and knowledge from management sciences and use this knowledge for operational business processes (Van Der Aalst, 2004). Business Process Management (BPM) is an approach that manage the execution of IT-supported business operations and this is done from a business expert's view rather than from a technical perspective (Hepp, Leymann, Domingue, Wahler, & Fensel, 2005; Smith & Fingar, 2003). In the same way, two roles can be distinguished in the BPM lifecycle: business manager who

DOI: 10.14267/phd.2015065

create and analyze business process models from the business point of view while IT engineers are involved in the implementation and execution phases. Business process management has been originated from the global business trends, as a management method to facilitate strategic alignment by streamlining business processes, and harmonizing organization and technology. BPM describes business processes in a complex modelling tool and implement the process in supporting applications (ERP, workflows). The emphasis is on the effective use of models for automatic generation of IT applications. Strategic alignment is a dynamic process: continuous adjustment of strategy, organizational structure, technology platform and skills (knowledge) is a key issue in today's business environment, more important than ever. Frequent changes in the environment (regulation, requirements of compliance, changing user needs, shorter product life-cycles, customization, emerging new technologies and "superconductivity" of markets) makes the challenging task of harmonization between process, skills, human resource and technology. BPM is traditionally an effective tool for revitalizing outdated business process, increase their productivity and improve quality. In literature, different phases of BPM are described such as process modeling, process implementation, process execution and process analysis. In recent years, different BPM systems are developed to improve operational business processes in the form of cost, time or error. However, this improvement is not in the way to improve individual activities rather it manages the flow of events of a process that ultimately adds value to the organization. As it is mentioned earlier that BPM has roots in information technology and historic view on information systems' development also illustrates that BPM systems can be used to push "process logic" out of the application (Van der Aalst, 1998).

DOI: 10.14267/phd.2015065

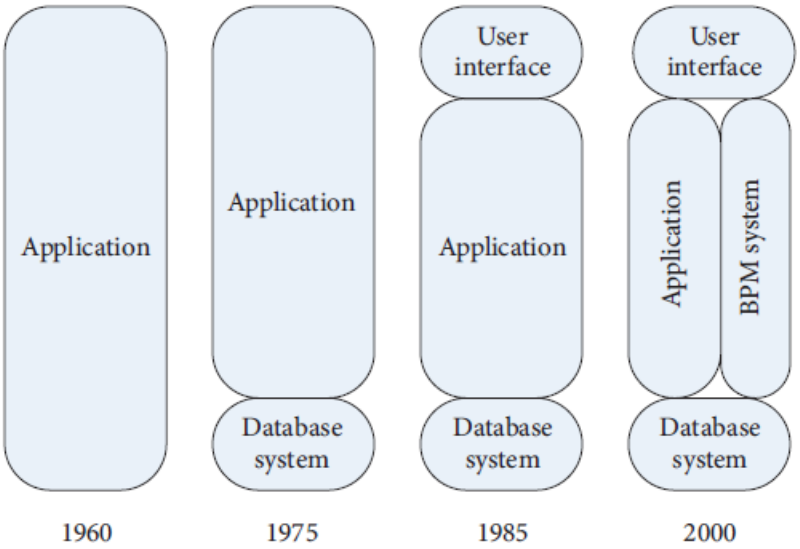

Figure 2-1 Business Process management systems in a historical perspective (adapted from (Van der Aalst, 1998).

2.1.1 BPM standards

BPM standards can be categorized with respect to similar functions and characteristics. According to features that define the process design and process enactment phase, BPM standards can be divided into four main types of standards (Ko, Lee, & Wah Lee, 2009).

Business Process Model and Notation (BPMN): According to this standard the business processes and their flow are represented through diagrams. These standards have highest level of expression of business processes. These diagrams based on flowcharting technique that is very similar to activity diagrams of Unified Modeling Language (UML).

DOI: 10.14267/phd.2015065

Execution standards: It computerizes the deployment and automation of business processes. There are currently two prominent execution standards: BPML and BPEL. Of the two, BPEL is more widely adopted in several prominent software suites.

Interchange standards: This type of standards facilitates portability of data. This standard and above mentioned standards as well address the process design and process enactment stage of the BPM life cycle. This standard acts as a bridge between graphical standard and execution standard. However, sometimes this translation can be imperfect because both standards are conceptually different.

Diagnosis standards: Diagnosis standards address the diagnosis stage of the BPM life cycle. These standards govern the management and optimization of business processes. Diagnosis standards are the most under-developed of all standards.

2.2 Semantic Business Process Management

Due to its IT oriented nature the major challenge in BPM is the ability of seamless translation between the business requirements view and the IT systems and resources. Semantic Business Process Management (SBPM) is a new approach that can increase the level of automation in the translation between these two domains (Hepp et al., 2005). For the improvement of BPM lifecycle, semantic technologies in particular ontologies, reasoners and semantic web services (SWS) are integrated with BPM tools. These semantic technologies increase the automation degree within the BPM phases. Through these semantic technologies, Semantic Business Process Management (SBPM) close the Business-IT gap by using semantic technologies. In SBPM analysis two features can be distinguished; one is process monitoring and other is process mining and both these operate on the event log which is written during process execution. The process mining analyzes executed process instances for the improvement of process model. Through process mining the performance of business models in the form of cost and duration can be improved (Wetzstein et al., 2007).

DOI: 10.14267/phd.2015065

2.3 Business Process Management and Knowledge Management

The process modelling supports the design and implementation of IT applications. The maintenance and the systematic integration of models is a great burden. BPM describes organizational knowledge about operations in form of models, model based solutions, measurement and controlling methods, and organizational arrangements (roles, responsibilities, etc.). From this aspect, BPM is a form of organizational learning: about strategy, organizational structure, IT and knowledge necessary for operations. Therefore, each phase of the BPM life cycle is knowledge dependent and should be supported by knowledge management methods and tools (Gábor & Szabó, 2013). BPM phases and activities are dependent from organizational knowledge; BPM can be aligned to knowledge management. Figure 2-2 presents the relation of knowledge management life cycle phases and business process management fields. The external cycle details the steps of knowledge management life cycle, while the internal cycle deals with process management life cycle phases. Using semantic technologies, knowledge management tools can be implemented to facilitate the management of process and job related knowledge elements, enabling customized training programs and the efficient maintenance of knowledge.

DOI: 10.14267/phd.2015065

Figure 2-2 Relation of knowledge management life cycle phases and business process management fields

2.4 Text Mining

Process mining is usually used to construct or learn about the nature of a process which is hidden or unknown for the observer. Process mining investigates that what tasks, in what order follow each other, what events trigger which task or connecting process and it deals with event log data (structured data) of business process with the aim of process analysis and improvement. Process mining techniques allow knowledge extraction from events stored by information systems. The objective of my thesis is to device such a paradigm that can extract information/knowledge from the processes and use this knowledge to make knowledge base (domain ontology). As mentioned above that a business process splits into tasks and a task has a number of attributes such as description, responsibility. My focus is on the description attribute that contains information about that task in the form of text. To find knowledge from this text that can be

DOI: 10.14267/phd.2015065

used in the enrichment of domain ontology so there is need to perform advanced analysis on this process' text, in this case, I have to apply text mining on the description field specifically to infer useful insights. Usually text mining is performed on unstructured data like long sentences and comments while process mining is performed on event log that has structured fields such as caseID, activity, timestamp and actor. Text mining is a method or approach of the extraction of new, heretofore unknown information from any text. In literature, text mining is applied to event logs for different purposes such as (Peng, Li, & Ma, 2005) applied text mining techniques to categorize messages in log files into common situations, improve categorization accuracy by considering the temporal characteristics of log messages. (Bembenik, Skonieczny, Rybinski, Kryszkiewicz, & Niezgodka, 2013) used text mining for advanced event log classification.

2.5 Text Preprocessing

A huge collection of our digital data is in textual form. This textual data is in natural language which is semi-structured form. It is really hard to extract rules from semi-structured and thus, such data cannot be used for prediction or any other useful function. When text mining techniques are given to such semi-structured huge data, another problem of pattern overabundance can create, in which immense number of practices are generated and is very hard to encounter out the only relevant result sets of a user. Thus, to perform text mining, it is necessary to run this data through a process in which different refinement techniques are given to this data and this operation will make sophisticated refinements in data. Then this refined data will be translated in such a shape that will be more appropriate to extract knowledgeable data for users. This process is called "preprocessing" of data. Research community working on different preprocessing techniques which are all different from knowledge discovery

DOI: 10.14267/phd.2015065

preprocessing techniques which set up structured data for data mining operations. Some techniques of preprocessing are discussed here.

Wang (Yanbo Wang, 2004) described five different preprocessing approaches to create an intermediate form of text documents for text mining. These approaches are: Full Text Approach in which a set of words is selected from a text document and this set of words represents that whole document. This band of words representation is also addressed as a suitcase of words. Many researchers (Ahonen-Myka, 1999, 2002; Ahonen, 1999; Yanjun Li & Chung, 2005) used this approach for text mining. For further refinement in data stop words can be removed from this set of words. The second approach is Keywords / index data approach (Delgado, Martín-Bautista, Sánchez, & Vila, 2002; Feldman, Dagan, & Hirsh, 1998; Feldman & Hirsh, 1996). This approach refines each bag of words by referring a keyword list. Keywords are extracted by using different schemes like frequency based weighing scheme (Salton & Buckley, 1988) or key word extraction based on CRF (conditional random fields) (C. Zhang, 2008). One drawback of this approach is removal of rich data of the document which can be useful for text mining. Prototypical document is a third approach of preprocessing. This is the full text approach and composed of two components. One is Part of Speech (POS) tagging and other is Term Extraction. In POS, automatically, tags are assigned to words in a document. The second component Term Extraction is domain dependent. This term may be a word or a phrase. Another preprocessing approach is Multi-Term text phrase (Ahonen-Myka, Heinonen, Klemettinen, & Verkamo, 1999) is co-occurrence of a set of words in raw data. Wang (Yanbo Wang, 2004) described the Concept approach as a last preprocessing approach. After extracting key terms and their syntax relationship from raw text, more semantically meaningful concepts are extracted.

DOI: 10.14267/phd.2015065

Hotho (Hotho, Nürnberger, & Paaß, 2005) defined three main methods of preprocessing of text data. First is tokenization, in which stream of words is made by removing different grammatical symbols (punctuation marks), clean spaces and tabs from each text document. The resulting document of words is called a dictionary. Then to reduce words, some other methods, filtering, lemmatization, stemming and keywords selection is given to this dictionary. By filtering, stop words are taken away. Lemmatization is a procedure in which noun words are mapped into singular form and verbs are mapped into infinite tenses. But this operation is error prone, and then mostly used stemming method for this function. In stemming word is cut back to its root word. Porter (Porter, 1980) is well known stemming algorithm. The third method of preprocessing is keyword choice. This method is likewise employed to further cut down words from the lexicon. Different techniques are used for keywords selection. Hotho (Hotho et al., 2005) used entropy for this purpose. Words having low entropy, mean frequently occur in text files. So an importance of a word in a given domain can be found out by finding entropy. In some other keyword selection technique, distance between every two words is measured and most related keywords having minimum distance are selected (Kardan, Farahmandnia, & Omidvar, 2013).

Mathiak and Eckstein (Mathiak & Eckstein, 2004) delineated a text mining method for any biomedical application into five steps. Preprocessing is one of five steps. The authors defined tokenization, POS tagging, frequency count and stemming methods as the substance of preprocessing. By using these preprocessing steps, a lot of time can be saved by extracting most interesting terms from text documents for further mining process.

DOI: 10.14267/phd.2015065

Wang (Y Wang, 2012) proposed three text preprocessing approaches. First is a pruned bag of single-words approach which is different from a conventional bag of single-words approach in two aspects. It gets rid of common words from all documents because these words do not generate useful classification rules. It also removes very rare words that are present in some documents. A minimal and maximum threshold values are taken for removing such words. The second approach is Emerging pattern based bag of single-row. This approach is based on traditional emerging pattern approach in which frequency of an itemset can be increased by moving this itemset from one database to another database. In text mining, a document consists of different books and each disc represents a document with its predefined class label. By splitting the document base into small databases with regard to their predefined classes, emerging patterns can be taken out. A third approach is bag of frequent itemset. In this approach single word is regarded as a single point and frequent worst is as frequent itemset. From given document base frequent itemset can be brought forth.

In this paper (Torunoglu, Cakirman, Ganiz, Akyokus, & Gurbuz, 2011) authors, analysed the effect of preprocessing on Turkish text. For this role they need two large data sets from Turkish newspapers and used some basic pre-processing techniques stop word removal, stemming, term weighting and then separate the information. Solutions showed that stemming has great impact on Turkish Information Extraction (IR). While, stop word filtering and stemming have less impact on classification accuracies. Same work is likewise performed by (Sohail & Hassanain, 2012), but authors find out the impact of preprocessing techniques on Arabic language. In this respect, they showed a comprehensive study about different preprocessing techniques which researchers are using for handling text pre-processing applications based on the Arabic language.

DOI: 10.14267/phd.2015065

All above research study demonstrates that most text mining techniques are founded on the thought that a text document can be presented by a set of words. If each document is to be considered a linear vector, then for each word of the document, a numerical value is stored which shows its importance in the document (linear vector). All above research study done in preprocessing field, indicates that various preprocessing methods gradually find out a representative lot of words on which text mining techniques can be employed to find out interesting patterns of cognition. Some major preprocessing techniques are discussed in succeeding.

2.5.1 Tokenization

Tokenization is the beginning measure of text preprocessing. At the source, textual data is in the pattern of a collection of characters while any text mining techniques is applied to speech. Word is a sequence of meaningful characters. So, thither is a need to convert this collection of graphic symbols into words for further processing. Tokenization performs this chore to get all words from the given textbook. In tokenization, a stream of text is converted into a stream of processing units (words) or terms(Hassler & Fliedl, 2006). There are three major goals of tokenization: first is to split the input text into tokens, second is to identify meaningful keywords and the third is to recognize sentence and word boundaries. Tokens may be words, phrases, keywords or symbols. Tokens are separated by a single space and punctuation marks may or may not include. Tokenization plays a significant role in lexical analysis. Hassler et. al (Hassler & Fliedl, 2006) proposed a rulebased Extended Tokenization method. This extended tokenization method has two levels. At first level, word or sentence boundaries are defined and single tokens are produced which may be words, numbers or abbreviations. On the second point, multi tokens are semantically or contextually motivated groups of tokens are made which may

DOI: 10.14267/phd.2015065

include named entities, idioms or special format tokens. This second level tokens are through rule based typing. The second level improves the accuracy of first level tokens. Leaman and Gonzalez (Leaman & Gonzalez, 2008) developed a 3-stage pipeline architecture BANNER for tokenization. This architecture takes one sentence at a time and output simple, non-alphanumeric tokens. And these token are converted into features. This stream of features is then labeled and each token has one label. There are numerous tokenization methods (Attia, 2007; Huang, Šimon, Hsieh, & Prévot, 2007; Labadié & Prince, 2008; Meknavin, Charoenpornsawat, & Kijsirikul, 1997) are proposed according to different languages. Rehman et al. (Rehman, Anwar, Bajwa, Xuan, & Chaoying, 2013) discussed the issue of boundary detection of compound words in tokenization. The authors proposed a morpheme matching based approach for Urdu text tokenization. According to the authors, in Urdu language words are written in continuation without space so general tokenization methods which make tokens from a string on the basis of space between words cannot be applied. In morpheme matching based approach, they used three matching algorithms; first is forward maximum matching algorithm, second is dynamic maximum matching algorithm and the third is Dynamic maximum matching along with the maximum likelihood approach. Experimental results showed up to 97% precision, but their employment is extremely dependent on the dataset and if unseen data comes then precision can be lessened.

2.5.2 Stop Words Filtering

As a result of tokenization, an aggregation of words adds up and there is a need to reduce the dimensionality of resulting data. Consequently, filters are applied to the data and stop word filter is a standard filtering method that is applied as a second step of preprocessing method. The stop words filter removes words that have littered or no content information. First time, in

DOI: 10.14267/phd.2015065

1958, Hans Peter Luhn used the term “stop words” for non-keywords in his proposed Keyword-in-Context (KWIC) indexing technique (Blanchard, 2007). Stop words are those words which are useful to make sentences or phrases, but have little information content. Nevertheless, such words cause a heavy fraction of the text in documents so, in preprocessing, stop word filters are used to eliminate such words from the content. For this purpose, a word of the list is built, this list can be provided by the user or system can automatically build it. Some schemes (Lo, He, & Ounis, 2005; Rose, 2013) are also proposed for automatic generation of stop word lists. A general or classic stop word list of English language can be habituated. A stop word list contains words which have low discrimination value or hold no content information. The stop word list contains articles such as ‘the’, ‘a’ and ‘an’, conjunctions like, ‘and’, ‘or’, ‘but’, and ‘yet’, prepositions like ‘by’, ‘from’, ‘about’, ‘below’, ‘in’ and ‘onto’ etc. This list also contains very frequent non-significant words which occur extremely often, but have little information content to make interesting decisions and this list have also words which occur rarely so have no significant statistical importance (Hotho et al., 2005). Stop word list is contextual in nature (Pennete, 2014) or domain dependent, so according to application requirements this list can be customized.

1.1.1 Part-Of-Speech Tagging (POS)

Part-of-Speech is to be considered a sub process of linguistic preprocessing rather text preprocessing (Hotho et al., 2005). Actually, to apply POS tagging depends on the application. For example, corpus annotation projects, information extraction, speech synthesis, term extraction, and many other applications need this preprocessing step in which every token is assigned a part of speech (noun, verb, adjective, etc.). POS tagger usually contains 50 and 150 tags, but size may be fluctuations in different languages. For morphological rich languages like

DOI: 10.14267/phd.2015065

German language, tagger uses several hundred tags(Voutilainen, 2003). Different approaches such as rule based (Alfred, Mujat, & Obit, 2013; Gimpel et al., 2011; Manning, 2011), stochastic or probabilistic (Antony, Mohan, & Soman, 2010; Owoputi et al., 2013) and transformation-based learning (Ferraro et al., 2013; Naz, Anwar, Bajwa, & Munir, 2012) approaches have been applied for POS tagging. The aim of each approach is to get to a tagger that can designate part of language to each word according to its context in the textbook. In rule-based tagging, tags are assigned to each word by using some hand written rules which are usually based on the contextual framework(Hasan, UzZaman, & Khan, 2007). The accuracy of such taggers is not pretty good and these are also not robust while taggers based on stochastic approaches outperforms, their accuracy is much higher as compared to rule based taggers. Most stochastic approaches are based on Markov model (Hasan et al., 2007). Transformation based approaches are the combination of rule based and stochastic approaches. It's a method to induce constraints from tagged corpus. Transformation based taggers are easier to develop and they are also do not depend on language or tag sets.

2.5.3 Lemmatization

Lemmatization is a major preprocessing step of many text mining techniques. It is also a preprocessing step of Natural Language Processing (NLP) techniques. Lemmatizers are actually filters as a stop word filter, by using them more distinctive and relevant words or candidate terms can be found from the given data. For this purpose, lemmatizers use frequencies or different statistical methods on word form or lemma level. Lemmatizer maps a token into its lexical headword or baseword (lemma) as verbs are mapped to the infinitive form and nouns are mapped to nominative singular form. This mapping transforms the word into its normalized form. For this mapping, it is necessary to know about part of speech of each word that is possibly

DOI: 10.14267/phd.2015065

done by applying POS tagging in the last preprocessing step. A general lemmatizer consists of three parts: one is a set of rules, the other is a lexicon of basic (normalized) words and the last is lemmatization algorithm. These principles will fix that which word suffix should be withdrawn and/or added to get words in normalized form. These rules can be in if-then rules (Kanis & Müller, 2004) or Ripple Down Rules (if-then-else)(Plisson, Lavrac, & Mladenić, 2004). Lemmatizers can be categorized into two major types: one is a manual approach (Hajic et al., 2006; Lefever & Hoste, 2010) that is based on handcrafted lexicon of lemmas and manually created affix rules. The other is an automatic approach (Kanis & Müller, 2005; Plisson et al., 2004)in which dictionary of lemmas and a set of affix rules are inferred from training data. For different languages, researches proposed different approaches of lemmatization (Aduriz et al., 1995; Al-Shammari & Lin, 2008; Perera & Witte, 2005; Plisson et al., 2004). There are some limitations of lemmatization: it is so difficult to handle inflected natural languages have many words of the same normalized word, lemmatization of large dictionary is very time consuming strategy and lemmatization can be ambiguous as left can be normalized as left (adjective) or can normalize as leave(verb).

2.5.4 Stemming

The main purpose of lemmatization and stemming is same to reduce different inflectional forms. There are some differences between stemming and lemmatization. Lemmatization reduces inflectional forms to basic word or lemma, while stemming chops off the ending of words to their stems. The result of lemmatization may be a substitute word (synonym) but this is not in stemming. The aim of lemmatization is to normalize word while the aim of stemming is just stem finding. To remove disambiguation, lemmatization takes context into account while stemmer does not resolve this trouble. Lemmatization only deals with inflectional variance

DOI: 10.14267/phd.2015065

while stemming also deal with derivational affixes. Lemmatization needs a comprehensive dictionary that has all base words with inflected forms or rules to drives inflected forms while stemmers do not call for this lexicon. Lemmatization is more useful for morphological complex languages while in most information retrieval applications, stemming performs better where data size is immense. From an implementation perspective, stemming is easier than lemmatization.

To surmount the limitations of lemmatization, stemming is used in information retrieving and other text mining applications. Stemming is a procedure to reduce inflated words to their stems. In this manner, by using stemming is applied to trim the data set size. In literature, different approaches are proposed for stemming algorithms. Some researchers (Jivani, 2011; Smirnov, 2008) defines three basic approaches of stemming algorithms.

One approach is truncating stemming algorithms, these algorithms are also called affix removal stemmers. In these algorithms, suffixes or prefixes are removed from words and produce a word in basic or root form which is called the stem. These are the simplest stemming algorithms. The first ever published description of stemmer was also close to a truncate stemmer of J.B Lovins. The world famous Porter stemming algorithm is also an example of truncate stemming algorithms (Dawson, 1974; Lovins, 1968; Paice, 1990; Porter, 1980, 2001). In all these algorithms, some transformation rules are applied to cutoff known prefixes or suffixes. These algorithms are easy to implement. However, these algorithms need prior language knowledge to form transformation rules and there are chances of over stemming or under stemming.

The second approach of stemming algorithms is statistical approach. This type of algorithms are not dependent on language. These stemmers usually strip of words, but after putting on some

DOI: 10.14267/phd.2015065

statistical methods. Instances of such stemming algorithms are N-gram stemmer (Mayfield & McNamee, 2003),

HMM stemmer (Melucci & Orio, 2003) that is based on Hidden Morkov model. This algorithm uses unsupervised training and stemming is performed by calculating the most probable path (stem) for any input word. This is fully automated system where no need for prior knowledge of the speech or a training set of manually stems words. All the same, this statistical approach is less complex and there are chances of overstemming. Another statistical stemming algorithm is proposed (Rogati, McCarley, & Yang, 2003) for non-English language which is based on statistical machine translation. This is an unsupervised learning approach. For training purpose, it uses an English steamer and a parallel corpus. Yet Another Suffix Stripper (YASS) is a statistical stemmer presented in 2007 (Majumder et al., 2007). The author used a hierarchical clustering to discover classes of root (stem) words and their variants. For long matching prefixes, the author defined a lot of string distance measure. GRAS (Paik, Mitra, Parui, & Järvelin, 2011) is a graph based statistical stemming algorithm. A set of co-occurring suffix pair is automatically identified from the lexicon. In each category, a pivot (central word) is identified and other associated words are neighbors of pivot in the division. And for every word (except pivot) most of its neighbors are also neighbors of pivot. With low computation GRAS performs better than other rule based stemmers.

The third approach of stemming algorithms is mixed, which include some inflectional and derivational algorithms, some corpus based algorithms and some contextual sensitive stemming algorithms. Inflectional and derivational algorithms need a large corpus so this subcategory is also being seen equally a part of corpus based approach algorithms. All classical stemmers like

DOI: 10.14267/phd.2015065

porter stemmer or Lovin's stemmer conflate words having same syntax, but different semantics because they do not consider corpus. There are hazards that one stemmer gives more precise solutions for one corpus but not for other corpus like stock, stocks, stocked and stocking may have extra meaning in the sheath of the Wall Street Journal. Seeing all this, Jinxi Xu and W. Bruce Croft (J. Xu & Croft, 1998) proposed a new approach corpus based stemming. In Corpus based stemming, equivalent classes of words are automatically modified that suits the characteristics of a given principal. The primary idea of this report is the conflated forms of a word co-occurs in the documents of that principal. Their approach uses corpus statistics to refine conflation. Initial set of conflated words that generated from stemming can be changed by using co-occurrence measure which is similar to expected mutual information. Done this manner, over stemming and under stemming can be manipulated to some extent. Nevertheless, use of trigram can reduce the chance of conflation. Another purely corpus based stemming approach is proposed (Paik & Parui, 2011) which uses lexicon of the corpus. In this unsupervised stemming algorithm, they collected words from the corpus and grouped together the words having same suffixes. The number of the words in a group show the frequency of the corresponding suffix. Then they select potential suffixes that have a larger frequency than a defined threshold. In this way, they identified a set of potential suffix of length n (n=1,2,….). So if a set of words (w1, w2,…. wn) is generated by a root word w then its variants or suffixes of w1,w2,wn that induced from w belong to the set of potential suffixes. Then equivalence classes are generated on the basis of information about common prefix and potential suffix. After generating these equivalence classes, then they pass their own proposed algorithm. This algorithm checked the strength of each class by this formula: $\text{Strength}\,(\text{R}) = \frac{\text{size of the potential}-\text{class(R)}}{\text{size of the generated}-\text{class(R)}}$. If strength is greater than the defined threshold, then this class

DOI: 10.14267/phd.2015065

becomes the equivalence class and its longest common prefix becomes the root of the class. Otherwise, all words are checked iteratively for a valid root and any member whose strength is equal to or greater than the threshold, it will become the potential root and generated class will be equivalence class. They applied their proposed stemmer on four languages, Bengali, Marathi, Hungarian and English. They compared their experiments' results with other stemmers like YASS, Oard, n-gram, porter and snowball. Results showed that performance time of their scheme is much smaller than all other schemes. Computational overhead is also very low as compared to YASS that is also clustered based approach for stemming. Besides all this, this approach showed not good results for query based stemming.

2.6 Information Extraction

Many researchers are integrating different data extraction methods, natural language processing techniques and Knowledge Discovery from Databases (KDD) to find useful knowledge from unstructured or semi-structured information.

Karanikas and Tjortjis (Karanikas et al., 2000) presented a new text mining approach, TextMiner which involved Information Extraction and Data Mining. This approach consists of two main components, one is text analysis and other is data mining. In text analysis component after applying preprocessing techniques on text they extracted events and terms from a document and then convert this information in structured form. And in second component they performed data mining by building up a novel clustering algorithm to see structure within the text file. They used their approach for financial knowledge base. So they drew out key terms and issues referred to finance. Every event has some attributes. All this information kept in a table. After pulling out such information from all documents, the resulting database contains all

DOI: 10.14267/phd.2015065

text files as records of the database and effects are as attributes of the record and this database is utilized as an input for clustering algorithm. They cause some alterations in the ROCK clustering algorithm which is used for categorical data. For substantiation of their attack they used some classification techniques by using description derived from clustering as class attribute. In that respect are some restrictions in their proposed scheme. One is of list of events that a user will define so there may be chances of errors because it is manual work and there is no guarantee that the list will cover all events of that domain. In this way, this approach is not fully automatic. The other limitation is about clustering algorithm. They used ROCK algorithm with some changes, but ROCK algorithm has some has limitation, like it is designed as it scales the interconnectivity according to static user specified interconnectivity model so if a model is under or overestimate the interconnectivity then incorrect decisions can be made. One other drawback of using ROCK algorithm is its similarity function which is dependent on document length.

Nahm and Mooney (Nahm & Mooney, 2002) integrated the text mining with information extraction and data mining. They proposed an automatically learned information extraction system, DxscoTEX (Discovery from Text EXtraction) to extract structured database from document corpus and then KDD tools are applied to mine this database. In this framework, they also developed an alternative rule induction system called TEXTRISE for partial matching in text mining. TEXTRISE combined the instance-based learning and rule induction. This system suggested a method of learning word to word relationship across fields by integrating data mining and data extraction. By utilizing this system, experiments were led to construct soft matching rules from textual databases via information extraction. The primary drawback of Mooney's system is time complexity because TEXTRISE is based on RISE algorithm and time

DOI: 10.14267/phd.2015065

complexity of RISE is quadratic$O(e^2, a^2)$ where ‘e’ represents no. of examples and ‘a’ represents no. of attributes. So as dataset will increase in size, efficiency will be diminished. The second limitation of this system is about the information extraction system that is RAPIER. RAPIER does not deal with the relationships of attributes because it is a field level extraction system (Chang, Kayed, Girgis, & Shaalan, 2006). RAPIER also handles only single slot extraction of semi structured data.

Popowich (Popowich, 2005) developed an enterprise healthcare application which processed on structured and unstructured data associated with medical insurance claims. In this study, the author combined text mining with Natural Language Processing (NLP) techniques to detect dependencies between different entities with textual data. By using NLP techniques, the author defined a text-based conceptual creation algorithm and developed an NLP Concept Matcher. Through this This NLP Concept Matcher considered only those concepts that extend the selected relevant key words and in this way when the text is tagged then these tags used as indicators for further mining process. This report also proved that structured linguistic analysis gives more accuracy than using pure stochastic analysis. This organization is not fully automated because human involvement is taken in the investigation of resulting claims.

Interactions between biological factors and effects of these interactions are real important in biological systems. Such interactions and effects are called events (Ananiadou, Pyysalo, Tsujii, & Kell, 2010). This paper described all techniques of text mining that are available for event extraction. Automatic extraction of events causes an extensive scope of biological applications. Searching is one instance of such applications that use text mining techniques. MEDIE (Miyao et al., 2006) is a search system that uses events for searching. This system automatically

DOI: 10.14267/phd.2015065

identifies semantic types by using syntactic and semantic analysis. This scheme is based on parsing technology so in this way this search system is more powerful and specific than key based search systems. In the past, for network construction or linking pathways to literature, text mining systems rely on the extraction of binary interactions that don't show coherent interpretation of reported facts. Thus, extraction of context around the events is also important to represent pathways. For this design, some text mining annotation and visualization tools are incorporated. PathText (Kemper et al., 2010) is the best example of such integration. The writer also identifies different approaches that draw out events from bio text. Pattern matching approaches which extract sentences. These sentences contain match patterns. Generalization is limited in such approaches and is not transferable to another user cases. Rule based approaches require sublanguage grammars and dictionaries which describe the constraints of the area. GENIES (Friedman, Kra, Yu, Krauthammer, & Rzhetsky, 2001) is an example of the rule based system that is fully parsed. To establish such systems is expensive.

2.7 Ontology Learning

We are going to develop such a text mining solution that can extract knowledge from business processes in order to automatically or semi-automatically enhance or populate the existing domain ontology. Therefore in this section, we will discuss an ontology learning process in general, which steps are included in this process. Till so far how many efforts have been made in this context.

DOI: 10.14267/phd.2015065

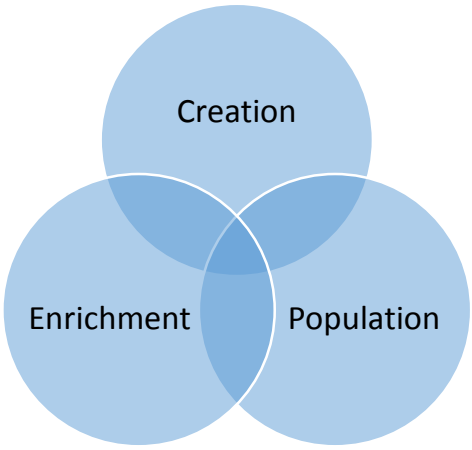

Figure 2-3 Ontology Learning

The most cited definition of an ontology is, "an ontology is a formal specification of a conceptualization" (Gruber, 1993). While, ontology learning refers to the process of creating an ontology in an automatic or semi-automatic way with a limited human exert. It also referred as a process to extract conceptual knowledge from several sources and building or creation of an ontology from scratch, enriching, or populating an existing ontology. The creation of an ontology can be represent by a touple <C, H, R, A> (Zouaq, 2011) where C represents the set of classes, H represents the set of hierarchical links between the concepts, R is the set of conceptual links and A represents the set of axioms. Acquiring knowledge from a specific domain is also called ontology learning (Santoso, Haw, & Abdul-Mehdi, 2011). George. et al. (George, Vangelis, Anastasia, Georgios, & Constantine) divide ontology learning into six major subtasks; term identification, synonym identification, concept identification, taxonomic relation identification, non-taxonomic relation identification, rule acquisition. Maedche and Staab described a conceptual model KAON Text-To-Onto system (Maedche & Staab, 2004) that consists of four general modules of ontology learning; ontology management component deals ontologies manually, resource processing component is about preprocessing of input data that

DOI: 10.14267/phd.2015065

will pass to algorithm library component which is the next component which acts like a backbone of ontology learning framework and responsible for extraction and maintenance, the last component is coordination component in which ontology engineer select the input data and choose method from resource processing module and algorithm from algorithm library. This framework performs ontology import, extraction, pruning, and refinement. A flexible framework OntoLancs (Gacitua, Sawyer, & Rayson, 2008) for ontology learning is presented. This framework presents a cyclic process that have four phases. Phase one is part-of-speech (POS) and semantic annotation phase in which domain corpus text is tagged morpho-syntactically and semantically. The second phase is extraction of concepts where a list of candidate concepts are extracted from the tagged domain corpus by applying a set of NLP and machine learning techniques. In third domain ontology construction phase, a domain lexicon is built with the help of some outsources (WordNet, Webster) and last phase extracted concepts are added to a bootstrap ontology. Fourth and last phase of the framework is domain ontology edition phase in which boot strap ontology is converted into light OWL language and then ontology editor is used to modify/improve this domain ontology. In another study (Nie & Zhou, 2008), authors explained ontology learning into three subtasks; extraction of concepts, extraction of relations and extraction of axioms. To perform these tasks they proposed an ontology learning framework OntoExtractor to construct ontologies from corpus. The main steps of OntoExtractor are seed concept extraction, syntactic analysis, new seed concept extraction and semantic analysis based on templates. Barforush and Rahnama taled about the creation of ontologies and they described four main steps that are employed for ontology building; (i) concept learning (ii) taxonomic relation learning (iii) non-taxonomic relation learning (iv) axiom and rule learning (Barforush & Rahnama, 2012)

DOI: 10.14267/phd.2015065

In the same line of research, this chapter proposes a text mining solution that is based on a set of methods that are contributing in aforementioned all three major ontology learning processes.

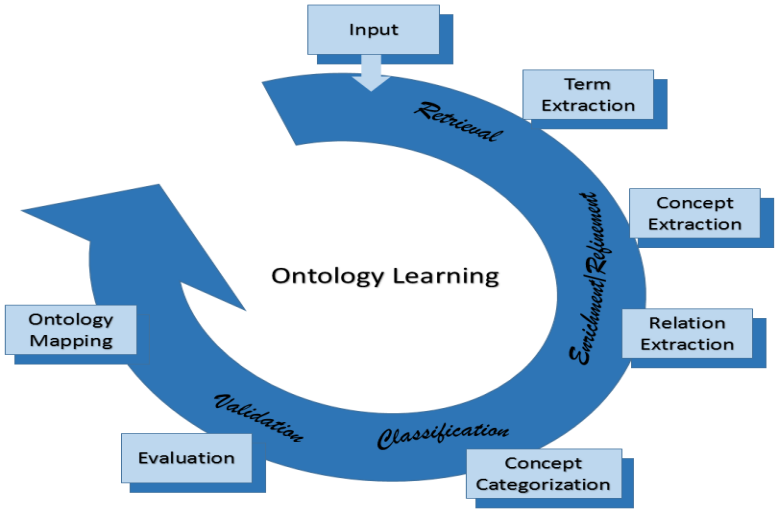

Figure 2-4 Ontology Learning Process

2.8 Ontology Extraction Tools

Since manual ontology construction is costly, time-consuming and error-prone work therefore during last decades, several semi and automatic ontology tools are presented to make ontology learning process more effective and more efficient. However, most ontology tools deal while there are few tools that cover whole ontology learning process. These tools can be broadly classified into two major categories. First those which mainly deal with plain text for ontology building while second category tools use semi structured text (Barforush & Rahnama, 2012). These ontology learning tools are divided into three types by Park et al. (Park, Cho, & Rho, 2010) and these three parts are ontology editing tools which provide help to ontology engineer in the acquiring, organizing, and visualizing domain knowledge, ontology merging tools are used to make one coherent ontology from two or more existing ontologies and the third type

DOI: 10.14267/phd.2015065

tools are ontology extraction tools which extract concepts and/or relations by applying some NLP or machine learning techniques. In this section I will discuss some of such tools.

Though ontology editing tools (Auer, 2005; Farquhar, Fikes, & Rice, 1997; Islam, Siddiqui, & Shaikh, 2010; Noy et al., 2001; Sure, Angele, & Staab, 2002) and ontology merging tools (Noy & Musen, 2003; Raunich & Rahm, 2011) also reduce the ontology building time but ontology extraction tools play more promising role in ontology automation. In this chapter our focus is on ontology extraction tools as I earlier mentioned that acquiring domain knowledge for constructing ontologies is an error prone and time-consuming task, thus, automated or semi-automated ontology extraction is necessary. In last two decades, for this purpose many ontology extraction tools have been developed.

Text2Onto(Cimiano & Völker, 2005) is an ontology learning framework that is a successor (completely redesign) of TextToOnto (Maedche & Staab, 2000). Text2Onto combines machine learning and NPL techniques to extract concepts and relations. In first phase through NPL techniques like tokenization and sentence splitter are applied to find an annotation set on which POS tagger is applied and then this POS tagger assigns a syntactic category to each token. After this, machine learning and linguistic heuristics are applied to derive concepts and relations from the corpus. During this process Text2Onto apply different measures to find the relevance of a term with respect to corpus and result of this whole extraction process is a domain ontology. The whole process is monitored by ontology engineers. This cyclic process has some shortcomings. One of these shortcomings is the difficulty to make compound words due to lack of deep semantic analysis and due to stochastic methods Text2Onto generates very shallow and light weight ontologies (Zouaq et al., 2011). Text2Onto also lacks ontology change management and validation (Zablith, 2008).

DOI: 10.14267/phd.2015065

Jiang and Tan (Jiang & Tan, 2010) proposed a system, Concept-Relation-Concept Tuple based Ontology Learning (CRCTOL) for ontology learning. This system follows a multiple corpus based approach for key concept extraction. CRCTOL, automatically extracts semantically rich knowledge of domain related documents. For this determination, this arrangement utilizes a full text parsing technique and employ both linguistic and statistical methods to identify key concepts. The authors also proposed a rule based algorithm to find out semantic relations (include both systematic and non-taxonomic relations) between key concepts. An association rule mining algorithm is used for pruning unimportant relation during ontology building. For evaluation, they applied this system on two domains of terrorism and sports and compare the results with Text-To-Onto and Text2Onto. Results showed that ontologies built by CRCTOL are more concise and contain rich semantics as compared to other ontology learning systems. In that respect there are some limitations of this arrangement just as other automatic learning ontology systems, this organization also observes general concepts only and ignores whole-part relations that are likewise important in ontology building. The resulting ontology is based on domain specific documents so this ontology is not the comprehensive and accurate representation of given domain so, there are hazards that such ontology will not be useful for different applications of that knowledge base. The third limitation of this system is time expensive because it performs full text parsing. To identify domain relevant concepts, this system uses term frequency measure (Domain Relevance Measure) which computes the frequency in documents of the target area and contrasting domain documents. For accurate key concept extraction this approach involves a significant number of documents from both domains (target and contrast). However, less matured domains can have a small number of relevant

DOI: 10.14267/phd.2015065

documents and this leads to a high skewness in key concepts and the overall performance of system may affect.

For ontology learning, Kang et al. (Kang, Haghighi, & Burstein, 2014b) introduced a novel method called CFinder, that extracts key concept for an ontology of a domain of interest. The authors described main four approaches that are oftentimes utilized for key concept extraction in literature. These are: (i) Machine learning approaches; (ii) Multiple corpus based approaches; (iii) Glossary based approaches; and (iv) Heuristic based approaches. They highlighted the problems of all these approaches such as machine learning approaches strongly depend on quality and amount of training documents prior to learn. Multiple corpus based approaches (Jiang & Tan, 2010) can face problem in performance when different domains have different corpus size. In glossary based approaches, a set of key concepts is provided, but it is not sure that all terms of glossary carry important information of the domain, because some new or too general terms may also present in this set; so it can be hard to find key concepts of corpus on the basis of such provided terms. The writers claimed that their system overcomes all these troubles. CFinder, find domain specific single-word terms that all are nouns. Hence, derived compound phrases by using a statistical method that mixes these single words. In this process, CFinder ignores the non-adjacent noun phrases. To work out weights for these candidate key concepts of the domain, CFinder combines statistical knowledge with field specific knowledge and inner structural pattern of these extracted candidate key concepts. This area specific knowledge obtained from the domain specific glossary list that is furnished by the author (domain expert) or already available glossary of that area. This list contains domain related terms. CFinder use this list to assign a high score to domain specific key concepts. They evaluated the effectiveness of CFinder against the three state of the art methods of key

DOI: 10.14267/phd.2015065

extraction. Results showed that CFinder outperforms in comparison of other key extraction methods. In the beginning, the authors pointed out the drawback of glossary based approaches, but their system also uses domain specific glossary. Although, at the conclusion, they remarked that without domain specific cognition, their system can also do well, but they did not pass on any experimental proof for this call. In spite of its seeming limitations, key concept extraction is a major step of ontology learning, but notwithstanding, it is a question how to estimate semantic relations between these extracted key concepts.

OntoCmaps (Zouaq et al., 2011) is a domain-independent unsupervised ontology learning tool that extracts deep semantic representations from unstructured text in the form of concept maps. This ontology learning tool is based on three phases: 1) a knowledge extraction phase which relies on a deep semantic analysis based on syntactic dependency patterns; 2) the integration phase builds concept maps, which are composed of terms and labeled relationships, and uses basic disambiguation techniques such as stemming, synonym detection. These concept maps form a concept map around domain terms; and finally 3) the filtering phase where various metrics rank the items (terms and relationships) in concept maps and acts as a sieve to filter out irrelevant or too general terms from candidates. The good thing of this ontology extraction tool is this it does not rely on any predefined template for its semantic representation and knowledge extraction is performed on each key sentence. An improvement in this work is presented (Ghadfi, Béchet, & Berio, 2014) a flexible language (DTPL—Dependency Tree Patterns Language) for expressing patterns as syntactic dependency trees to extract semantic relations and through this DTPL each time they extract one kind of relation from a pattern because extraction of more than one kind of relations from a pattern indicates nested patterns to

DOI: 10.14267/phd.2015065

differentiate by specifying dependency bindings (each dependency binding consists of a dependency link, the governor and the dependent) that should not exist when a match occurs.

To overcome these literature gaps, in this chapter an ontology extraction tool ProMine is presented that will extract concepts from business tasks for domain ontology. To my knowledge, there have been no studies done to address in connecting text mining to process management in context of extracting new concepts of business tasks to enrich domain ontology (defined in this chapter) for ontology learning. This proposed automatic information extraction method will comprise two basic phases: In a first phase, system will extract information from business process and in second phase it will enhance the extracted information with the help of other sources such as WordNet, Wiktionary and corpus and this enhanced information will enrich the root ontology.

2.9 Concept Extraction to Enrich Ontologies

Despite many developments in ontological tools, knowledge acquisitions can a hindering fact in ontology building process due to the fact that it is still very manual, delaying, tedious and quite complex activity. Ontological development process needs assistance from technological resources like web, corpus, structured / semi-structured sources and dictionaries to minimize the time delay and efforts required for manual ontology learning process. Therefore, this has now become a comprehensive research field. Manual ontology building is expensive, time consuming, error-prone, biased towards their developer, inflexible and specific to the purpose of construction (Goldsmith & Messarovitch, 1994; Gómez-Pérez & Manzano-Macho, 2003; Meyer & Gurevych, 2012; Shamsfard & Abdollahzadeh Barforoush, 2003; Stanford, 2014). During the past decade, researchers proposed several methods to support semi-automatic or

DOI: 10.14267/phd.2015065

automatic methods for building ontologies. Ontology learning refers to the process of integration of a set of methods and techniques used for ontology engineering from scratch, enriching, or adapting an existing ontology in a semi-automatic fashion using several sources (Meyer & Gurevych, 2012). Ontology learning is a research area, which deals with the challenges to turn facts and patterns from the content into shareable high-level constructs or ontologies(Wong et al., 2012). Various ontology learning methods are discussed in the literature, which can be differentiated by several characteristics. One of the first surveys, which discussed the ontology learning related challenges, was published by the OntoWeb Consortium in 2003 (Gómez-Pérez & Manzano-Macho, 2003). They investigated 36 approaches for ontology learning from text. Their report pointed out that there is no fully automated ontology learning system available and numerous require user involvement to extract knowledge from the corpus. They concluded that there is a need for a general approach for evaluating the accuracy of ontology learning and for comparing the results produced by different systems.

Ontologies are used to find interesting on topic knowledge and they also improve the functioning of knowledge discovery. An ontology consists of concepts and relationships among them in a specific area. In recent years, a great deal of study has been performed in the text mining field to get these concepts from domain specific documents. In that respect there are different techniques like linguistic, statistical and machine learning are involved to extract new concepts and semantic relations among them. Linguistic techniques (Hamish Cunningham, 2005; H Cunningham, Maynard, Bontcheva, & Tablan, 2002; Hobbs, Appelt, Bear, & Tyson, 1992) based on the premise that by using syntactic analysis, the relationship between words can be made out. Statistical approaches (Cimiano & Völker, 2005; Jiang & Tan, 2010; Wong, Liu, & Bennamoun, 2007) used statistical measures to find out frequent terms in domain related

DOI: 10.14267/phd.2015065

documents. These frequent terms represent important concepts and frequent occurrence of these concepts indicates the relationship among them. On the other hand machine learning techniques use a set of algorithms to discover concepts and relations in an automated way. To find this type of knowledge, machine learning methods can use linguistic or statistical methods together. After discovering concepts and relationships, for an ontology building, different mining techniques are used to make connections and connections among these concepts. However, the performance of current concept extraction tools restrict them to be more suitable for ontology's with low execution and medium quality requirements and these tools are not suitable for achieving high conceptualizations efficiency according to good practices in the ontological modeling domain. For example, available term extraction algorithmic tools aid acceptable efficiency in term of precision and recall but they do not possess ability to differentiate between term and concept.

Jiang and Tan (Jiang & Tan, 2010) proposed a system, Concept-Relation-Concept Tuple based Ontology Learning (CRCTOL) for ontology learning. This system follows a multiple corpus based approach for key concept extraction. CRCTOL, automatically extracts semantically rich knowledge of domain related documents. For this determination, this arrangement utilizes a full text parsing technique and employs both linguistic and statistical methods to identify key concepts. The authors also proposed a rule based algorithm to find out semantic relations (include both systematic and non-taxonomic relations) between key concepts. An association rule mining algorithm is used for pruning unimportant relation during ontology building. For evaluation, they applied this system on two domains of terrorism and sports and compare the results with Text-To-Onto and Text2Onto. Results showed that ontologies built by CRCTOL are more concise and contain rich semantics as compared to other ontology learning systems.

DOI: 10.14267/phd.2015065

In that respect are some limitations of this arrangement. Like other automatic learning ontology systems, this organization also looks at general concepts only and ignores whole-part relations that are likewise important in ontology building. The resulting ontology is based on domain specific documents so this ontology is not the comprehensive and accurate representation of given domain so, there are hazards that such ontology will not useful for different applications of that knowledge base. The third limitation of this system is time expensive because it performs full text parsing. To identify domain relevant concepts, this system uses term frequency measure (Domain Relevance Measure) which computes the frequency in documents of the target area and contrasting domain documents. To accurate key concept extraction this approach involves a significant number of documents from both domains (target and contrast). However, less matured domains can have a small number of relevant documents and this leads to a high skewness in key concepts and the overall performance of system may affect.

For ontology learning, Kang et al. (Kang, Haghighi, & Burstein, 2014a) introduced a novel method called CFinder, that extracts key concept for an ontology of a domain of interest. The authors described main four approaches that are oftentimes utilized for key concept extraction in literature. These are: (i) Machine learning approaches; (ii) Multiple corpus based approaches; (iii) Glossary based approaches; and (iv) Heuristic based approaches. They highlighted the problems of all these approaches like machine learning approaches strongly depend on quality and amount of training documents prior to learn, multiple corpus based approaches (Jiang & Tan, 2010) can face problem in performance when different domains have different corpus size, and in glossary based approaches, a set of key concepts is provided but it is not sure that all terms of glossary are given important information of the domain because some new or too general terms may also present in this set so it can be hard to find key concepts of corpus on the

DOI: 10.14267/phd.2015065

basis of such provided terms. The writers claimed that their system overcomes all these troubles. CFinder, find domain specific single-word terms that all are nouns. Then, derived compound phrases by using a statistical method that mixes these single words. In this process this system ignores the non-adjacent noun phrases. To work out weights for these candidate key concepts of the domain, CFinder combines statistical knowledge with field specific knowledge and inner structural pattern of these extracted candidate key concepts. This area specific knowledge obtained from the domain specific glossary list that is furnished by the author (domain expert) or already available glossary of that area. This list contains domain related terms. CFinder use this list to assign a high score to domain specific key concepts. They evaluated the effectiveness of CFinder against the three state of the art methods of key extraction. Results showed that CFinder outperforms in comparison of other key extraction methods. In the beginning, the authors pointed out the drawback of glossary based approaches, but their system also uses domain specific glossary. Though, at the conclusion, they remarked that without domain specific cognition, their system can also do well, but they did not pass on any experimental proof for this call.

2.10 Similarity Measures for Ontology Learning

Similarity measures determine the degree of overlap between terms or words (entities) and this measurement is based on some pre-defined factors such as statistical information about these entities or semantic structure of these words or taxonomic relationships between these entities. The computation of the similarity between terms is at the core of ontology learning. In literature similarity measures are used for different applications of ontology learning such as some researchers used similarity measures to compare the similarities between the concepts in the

DOI: 10.14267/phd.2015065

different ontologies, some used for the task of detecting and retrieving relevant ontologies while Saleena et. al (Saleena & Srivatsa, 2015) proposed a similarity measure for adaptive e-Learning systems by comparing the concepts in cross ontology. There have been many attempts to determine similar term pairs from text corpora. It is assumed that if terms occur in similar context then they have similar meanings (Bekkerman, El-Yaniv, Tishby, & Winter, 2001; Dagan, Pereira, & Lee, 1994). The context can be defined in diverse ways such as it can be represented by co-occurrence of words within grammatical relationships. Some measures of similarity is employed to assign terms into groups for discovering concepts or constructing hierarchy [Linden and Piitulainen 2004].

In this chapter, my focus is concept extraction for ontology development. Therefore I saw the literature related to similarity measures used for concept extraction and process in ontology development. Aforementioned ontology extraction tools extract concepts from text by using NPL or text mining techniques and during this process many irrelevant results also come out. Majority of concept extraction approaches that are reported in literature are domain independent and few of them generally address these issues using traditional information theory metrics. In order to identify the most relevant terms, it is necessary to filter out noisy data (split words or words having no meaning) and general and irrelevant terms. The output of information extraction tool is usually a long list of words. Therefore, ranking is needed to compare several alternatives to find the best. The result of this ranking process, by applying a threshold, noisy and irrelevant words eliminate automatically. To present adequate results to users, a filtering process is applied on the extracted knowledge. These filtering methods use different statistical and semantic measures to obtain better results. By using semantic similarity measures,

DOI: 10.14267/phd.2015065

important concepts and relationships (elements of domain ontology) are filter out by comparing different candidate terms.

Researchers proposed different filtering and ranking methods based on different metrics such as co-occurrence measures, relevance measures and similarity measures to rank concepts and after ranking select the most relevant concepts. For term ranking, Buitelaar and Sacaleanu (Paul Buitelaar & Sacaleanu, 2001) taken a relevance measure from information extraction. Their relevance measure is an adaptive form of standard tf.idf. Their approach is task independent and complete automatic. They evaluated their method of ranking using human judgment by selecting 100 top most concepts. Results showed that 80 to 90 percent accurate prediction of domain specific concepts. Schutz and Buitelaar (Schutz & Buitelaar, 2005) developed a system (RelExt) that is capable to identify the most related pairs of concepts and relations from a domain specific text. For this purpose they used linguistic measures such as concept tagging and statistical measures such as relevance measure (χ^2 test) and co-occurrence measure. Wang et al (G. Wang, Yu, & Zhu, 2007) used entity features for filtering. From extraction method a large number of the entity pairs are generated and thus it is inefficient if they are directly classified so it is necessary to eliminate irrelevant entity pairs. (X. Wu & Bolivar, 2008) developed an advertising keyword extraction system. This system uses machine learning approach for ranking contextually relevant keywords. To model relevance score, linear and logistic regression models are used and experiments are executed with large set of features to obtain keyword ranking score. Text2Onto (Cimiano & Völker, 2005) relies on a distributional similarity measure to extract context vectors for instances and concepts from the text collection. To find the relevance of a term, different measures such as Relative Term Frequency (RTF), TFIDF (Term Frequency Inverted Document Frequency), Entropy and the C-value/NC-value

DOI: 10.14267/phd.2015065

are used. They also defined a Probabilistic Ontology Model (POM) that represents the results of the system by attaching a probability to them. In OntoCmaps (Zouaq et al., 2011), a set of metrics are defined to find the importance of a term such as Degree centrality, the Betweenness centrality and the Eigen-vector centrality. Betweenness is calculated by the ratio of shortest paths between any two terms. On the basis of these metrics, the author defined a number of voting schemes to improve the precision of terms filtering process.

Statistical measures face problem of sparsity when corpus size is small or of specialized domains. At that time there is a need to apply semantic measures to tackle such issues. However, it is also a difficult task to extract suitable semantic information from such corpus. In a semantic similarity measure, two concepts are taken as input and a numeric value is returned as an output which describes how much these concepts are alike (Pedersen, Pakhomov, Patwardhan, & Chute, 2007). These semantic similarity measures are used to find common characteristics between two concepts/terms. A number of semantic similarity measures have been developed in last two decades. These measures can be classified into four categories: i) Corpus-based similarity measures, ii) Knowledge-based similarity measures, iii) Featured-based similarity measures and iv) Hybrid similarity measures as shown in figure 3.

DOI: 10.14267/phd.2015065

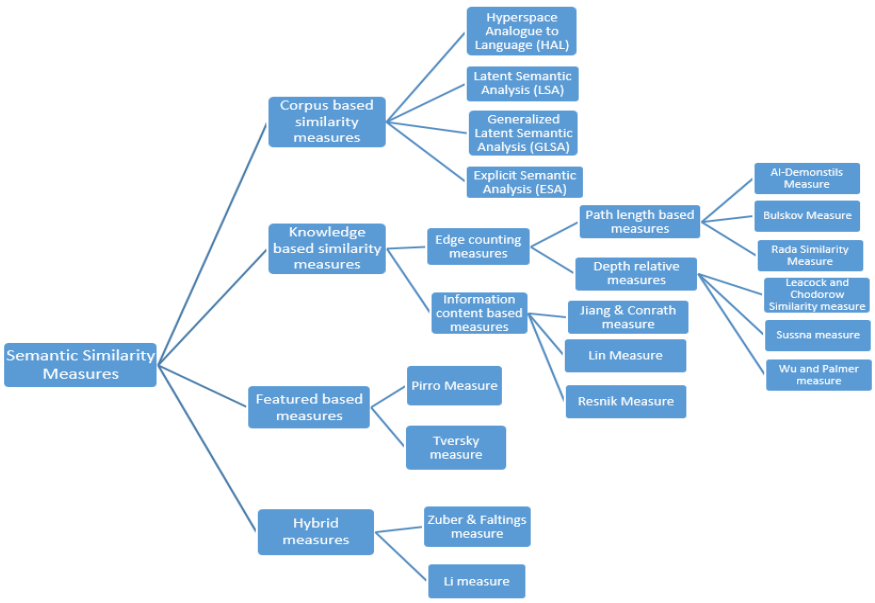

Figure 2-5 Semantic Similarity Measures

Corpus-based measures find the similarity between concepts/terms on the basis of information that derived from a corpus. Two well know corpus-based similarity measures are Latent Semantic Analysis (LSA) (Guo & Diab, 2012; Landauer, Foltz, & Laham, 1998) and Hyperspace Analogues to Language (HAL) model [Burgess et al. 1998]. In LSA, this is assumed that words that are close in meanings occur in similar pieces of text. LSA is a high-dimensional linear association model that generates a representation of a corpus and through this representation similarity between words is counted. In HAL, on the bases of word co-occurrences, a semantic space is created. Word ordering information (from a corpus) is also recorded in HAL.

Knowledge-based measures used semantic networks to measure the degree of similarity between words. Semantic networks are the networks that describe the semantic relation between words, the most famous semantic network is WordNet. In such type of networks, information is in the form of graphs where nodes represents concepts and vertices represent edges. On the

DOI: 10.14267/phd.2015065

basis of this semantic network many similarity measures has been proposed. Such measures can be also further categorized into two types: i) edge counting measures and ii) information content based measures. In edge counting measures, the similarity is determined by the path length measures (Euzenat & Shvaiko, 2007; Nagar & Al-Mubaid, 2008; Rada, Mili, Bicknell, & Blettner, 1989) in which shortest path between two concepts are measured. Edge counting measure can also find similarity through depth relative measures(Qin, Lu, Yan, & Wu, 2009; Sussna, 1997; Z. Wu & Palmer, 1994) in which depth of a particular node is calculated.

The information content is the information that a concept contains in a context in which it appears. Therefore, main idea of information content-based measures (Formica, 2008; Pirró, 2009; Resnik, 1995; Sánchez, Batet, & Isern, 2011) is to use this information content of the concepts. The more common information is shared between two concepts they are more similar to each other. If two concepts have no common information then it means they considered maximally different. This information content can get from corpus or from a knowledge base (WordNet). Information content calculation based on WordNet performs better than corpus based information context approaches (Sánchez et al., 2011) because sparse data problem cannot be avoided in corpus based information content similarity measures.

Hybrid measures combine the above mentioned approaches to find more accuracy such as these hybrid measures combine methods of length based measure and depth based measures. Zhou (Meng, Huang, & Gu, 2013)has proposed a hybrid measure that combined the information content based measures and path based measures. Some researchers (Meng et al., 2013; Slimani, 2013) evaluated aforementioned measures and conclude that every semantic similarity measure has some advantages and some shortcomings also. Path based measures are simple to implement but local density of pair concept cannot be reflected. Information content based measures cannot

DOI: 10.14267/phd.2015065

reflect structure information though they are simple and effective. Hybrid measures give more accuracy as compared to other measures though these measures are more complex than others and these measures also need turning of parameters.

2.11 Semantic Similarity Measure for Concept Filtering

The sole input of most knowledge extraction tools is a corpus. This corpus can be of a specific domain or web that is independent of domain. The corpus text comes from different websites, pdf documents, word documents, scanned documents or different domain glossaries converted automatically into plain text. This conversion process generate noise in the data. The second cause of noisy data is knowledge extraction approach. Zouaq et al. (Zouaq et al., 2011) described that unsupervised or blind knowledge extraction approaches generate a significant amount of noisy data due to improper syntactic or semantic analysis. Besides this noisy knowledge, many knowledge extraction tools generate irrelevant knowledge (in form of terms or concepts). However, domain ontology demands more relevant and noiseless knowledge.

Therefore, for ontology learning, a solution is required that should answer the aforementioned shortcomings by carefully choosing the source of knowledge (corpus), adopting deep syntactic or semantic analysis techniques and filtering the extracted knowledge. Majority of concept extraction approaches that are reported in literature are domain independent and few of them generally address these issues using traditional information theory metrics. In order to identify the most relevant terms, it is necessary to filter out noisy data (split words or words having no meaning) and general and irrelevant terms. The output of information extraction tool is usually a long list of words. Therefore, ranking is needed to compare several alternatives to find the best. The result of this ranking process, by applying a threshold, noisy and irrelevant words

DOI: 10.14267/phd.2015065

eliminate automatically. Researchers proposed different filtering and ranking methods based on different metrics such as co-occurrence measures, relevance measures and similarity measures to rank concepts and after ranking select the most relevant concepts. For term ranking, Buitelaar and Sacaleanu (Paul Buitelaar & Sacaleanu, 2001) taken a relevance measure from information extraction. Their relevance measure is an adaptive form of standard tf.idf. Their approach is task independent and complete automatic. They evaluated their method of ranking using human judgement by selecting 100 top most concepts. Results showed that 80 to 90 percent accurate prediction of domain specific concepts. Schutz and Buitelaar (Schutz & Buitelaar, 2005) developed a system (RelExt) that is capable to identify the most related pairs of concepts and relations from a domain specific text. For this purpose they used linguistic measures such as concept tagging and statistical measures such as relevance measure (χ^2 test) and co-occurrence measure. Wang et al (G. Wang et al., 2007) used entity features for filtering. From extraction method a large number of the entity pairs are generated and thus it is inefficient if they are directly classified so it is necessary to eliminate irrelevant entity pairs. (X. Wu & Bolivar, 2008) developed an advertising keyword extraction system. This system uses machine learning approach for ranking contextually relevant keywords. To model relevance score, linear and logistic regression models are used and experiments are executed with large set of features to obtain keyword ranking score.

2.12 Concept Categorization for Ontology Population/Enrichment

After concept extraction process the next step in ontology learning is to populate the ontology with these newly extracted concepts. In recent years, this research area has gain a significant attention, therefore though it is an emerging field, several approaches have been proposed and many practical systems have been developed. In this process, ontology structure does not

DOI: 10.14267/phd.2015065

change rather just new instances in the form concepts and relations are added into an existing ontology.

A detail review about ontology population systems are given in some research articles (Cimiano, 2005; Petasis, Karkaletsis, Paliouras, Krithara, & Zavitsanos, 2011; Z. Zhang & Ciravegna, 2011). In literature, three types of ontology population tools can be seen: First those ontology population systems which populate an ontology with instances of both concepts and relations (Kara et al., 2012; Packer & Embley, 2013; Ruiz-Martınez, Minarro-Giménez, Castellanos-Nieves, Garcıa-Sánchez, & Valencia-Garcia, 2011), some systems just populate relations in the ontology (Brewster, Ciravegna, & Wilks, 2002; Suchanek, Ifrim, & Weikum, 2006) while some other systems only populate ontology through concepts (Etzioni et al., 2005; Yates, 2004).

ISOLDE (Weber & Buitelaar, 2006) is a system to populate a base domain ontology with new concepts and relations by combining a domain corpus, a general purpose NER and web resources like Wikipedia, Wiktioonary and a German online dictionary (DWDS). In the first place, this system extracts instances from a base ontology with the help of NER system. New concepts are generated by applying lexico-syntactic patterns from base ontology class candidates. For filtration of these concepts, web resources are exploited and to determine the relevance between these concepts, a statistical measure $x^2$of are used. Their results showed that semi-structured data resources seem worthwhile. This approach is basically aim at a taxonomy rather than a complete ontology. Generally, there are more error chances in automatic taxonomy construction.

DOI: 10.14267/phd.2015065

In (Meyer & Gurevych, 2012), authors used Wiktionary for the construction of an upper level ontology. They highlighted the limitations of Wikipedia to use it as a knowledge base. They described a two stepped approach "OntoWiktionary", one is harvesting knowledge to extract knowledge from Wiktionary and other step is ontologizing knowledge for formation of concepts and relations. In first phase, they developed a new adapter, Wikokit, which allows to extract data from Wiktionary. In ontologizing phase, this system defines concepts of OntOWiktiOnary with the help of ontoWordNet. This two phase approach automatically enriched the extracted data and this data can be used to improve NLP solutions. My work is different in that I have seed ontology and want to enrich this seed ontology with new domain concepts. It means taxonomy or categories are already defined in the seed ontology and I want to enrich these categories with new domain concepts. Although, my proposed solution is also used Wiktionary as a part with other resources for concept extraction and categorization. However, focus of my work is different from OntoWiktionary. My system is semi-automatic because I believe that complete automatic system will be error prone.

(Janik & Kochut, 2008) presented a text categorization method based on Wikipedia categories using a thematic graph construction. This method basically used for document categorization. They described their approach into three parts; in first semantic graph construction part, a document is converted into a semantic graph after matching entity labels of ontology in a document. Edges between these nodes in the semantic graph are created based on the relationships existing in the ontology, in this part, an initial weight is also assigned to each node, in second part, a sub-graph of the semantic graph that is related to the main topic of the document is selected and this is called dominant thematic graph. In third part, each entity of dominant thematic graph is assigned a set of classes, according to the entity's classification in

DOI: 10.14267/phd.2015065

the ontology. For experiments, they used the ontology derived from Wikipedia, where the category pages were converted into the internal ontology classes.

(Chifu & Le Ia, 2008) described an unsupervised framework that is based on a kind of neural network, Growing Hierarchical Self-organizing Maps (GHSOM) for domain ontology enrichment. This framework functions as hierarchical backbone of an existing ontology and enriches it with new domain-specific concepts extracted from the corpus. The whole process can be divided into two main parts; the term extraction and the taxonomy enrichment. The term extraction process is based on recognizing linguistic patterns in the domain corpus documents. In taxonomy enrichment phase, the terms extracted from the corpus are mapped to classes of the existing taxonomy. They defined an algorithm for enrichment that populates the given taxonomy with the extracted terms. They used distributional similarity in which similar concepts are expressed by similar vectors in the distributional vector space. While the dissimilarity between vectors are computed through the Euclidean measure. Every new concept is attached as successor of an intermediate or a leaf node of the given taxonomy and becomes a hyponym of that node.

To use a vector space model for ontology enrichment (Chifu & Le Ia, 2008) can give wrong results due to a data sparseness problem. Concepts or terms represented by sparse vectors have an increased chance to be wrongly classified, because of the reduced power of attraction towards the correct branches and nodes of the taxonomy. The other problem that can be faced by using a vector space model is that sometimes distributional similarity is ignored or sometimes hard to interpret (e.g., friend and enemy are distributionally similar).

DOI: 10.14267/phd.2015065

Literature showed that Wikipedia became an important resource for ontology learning (Janik & Kochut, 2008; Suchanek, Kasneci, & Weikum, 2008). Although, Wikipedia yields a densely connected taxonomy of concepts, Ponzetto and Strube (Ponzetto & Strube, 2007) point out that the Wikipedia categories "do not form a taxonomy with a fully-fledged subsumption hierarchy." but represents the domain the concept is used in. one more problem of a Wikipedia-based ontology lies in the lexicalizations of concepts. In order to reduce redundancy, each concept is encoded only once within Wikipedia and thus described within the article with the most common lexicalization of the concept. Another problem with Wikipedia is that it is a very huge, rich database so, there are chances that any two entities have more than one relations with different strengths. To assign weights according to their strength is quite tricky task and a heuristic algorithm with a lot of computation is required.

To overcome these literature gaps, ProMine framework uses Wiktionary as an external resource for semantic concept categorization module. It takes categories from the seed ontology with some basic concepts and enriches these categories with the help of Wiktionary. The objective of the framework is to facilitate ontology learning from texts in real-world settings through two main basic tasks; one is concept extraction from text and the other is semantic concept categorization.

DOI: 10.14267/phd.2015065

ProMine: The Proposed Framework

This section proposes a flexible framework called ProMine for ontology learning from text. This system involves the successive application of various NLP techniques and learning algorithms for concept extraction, filtration and ontology enrichment. ProMine development is a part of PROKEX project (EUREKA, 2013) to build ontologies semi-automatically by processing a collection of texts of different domains. ProMine uses many text mining and data mining techniques for ontology learning which led to the development and enrichment of a domain ontology. Therefore, with the help of this framework an ontology can be built rather to enrich and populate an existing ontology.

I have developed a prototype workbench that performs three basic tasks; one is knowledge element extraction from the domain document corpus and other sources, and then concept filtering to find most relevant terms of a domain from the extracted knowledge elements. The third task is semantic concept categorization with these extracted knowledge elements that will help the enrichment and population of domain ontology. Initial work of ProMine concept extraction is presented in (Gillani & Kő, 2014). This prototype shows that proposed framework's efficacy as a workbench for testing and evaluating semantic concept extraction, filtering and categorization.

DOI: 10.14267/phd.2015065

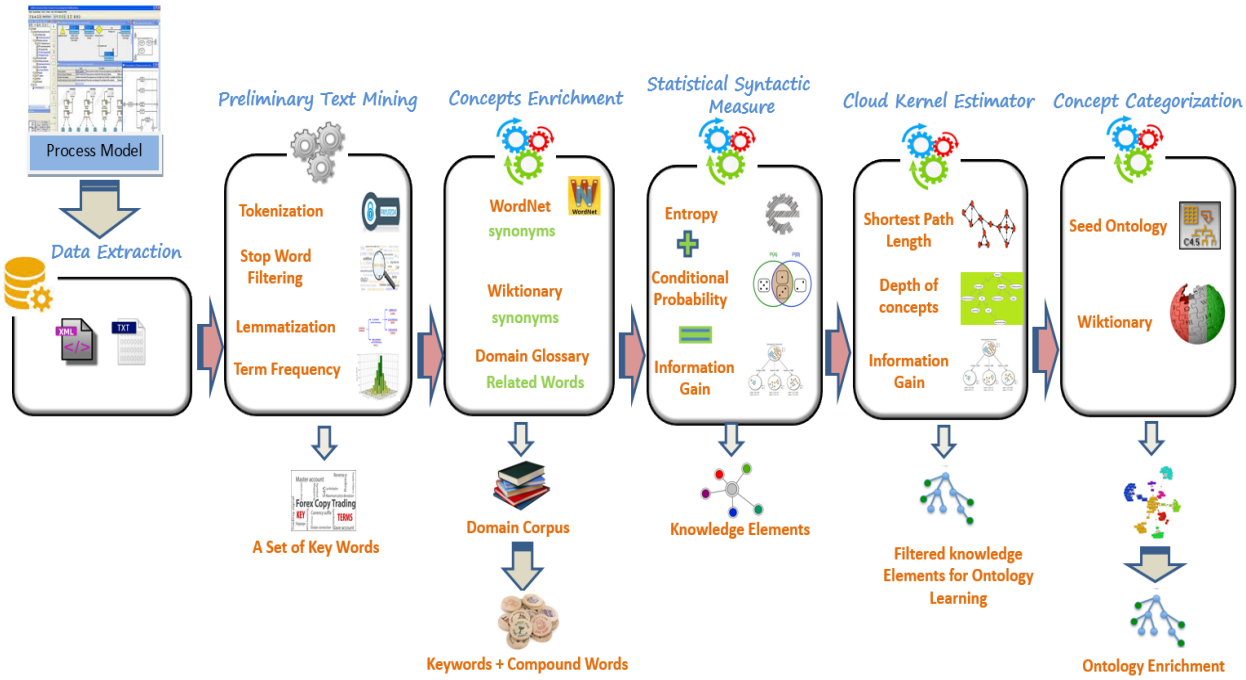

Figure 3-1 ProMine: A Functional Framework

This chapter embodies ProMine as a framework to extract knowledge elements from the context data as illustrated in Figure 3-1. The workflow of our ontology framework proceeds through the phases of (i) Data extraction from organizational process (ii) Text preprocessing on extracted data by applying natural language processing (NLP) techniques; (iii) Concept Enrichment Phase to extract concepts from domain corpus and other sources; (iv) focuses on the filtering process, introduces our proposed new hybrid semantic similarity measure and in section. (v) The last phase or module of ProMine framework is a semantic concept categorization that is used to categorize extracted knowledge elements in different categories (based on seed ontology) for ontology enrichment. Below I provide detailed descriptions of these phases.

DOI: 10.14267/phd.2015065

3.1 Data Extraction

A major difference between existing ontology extraction tools and ProMine is the data extraction phase that starts from a small sized input file while in already developed ontology learning tools the input is large sized corpus or any existing ontology. ProMine's input file is actually the output file of an organizational process by using a process model. As mentioned earlier, a process can split into different tasks. These tasks have different attributes such as description, responsibility, execution related information (order, triggers, and events) and information about all attributes are in this input file. Our focus is on the description attribute of a task because it contains explicit and tacit knowledge elements about tasks in an embedded way. This input file is in the form of XML. At the first step of this framework (data extraction phase), the pertinent information from this input file is extracted automatically by ProMine. After extracting specific text from the input files, this text is saved into text files according to all tasks.

DOI: 10.14267/phd.2015065

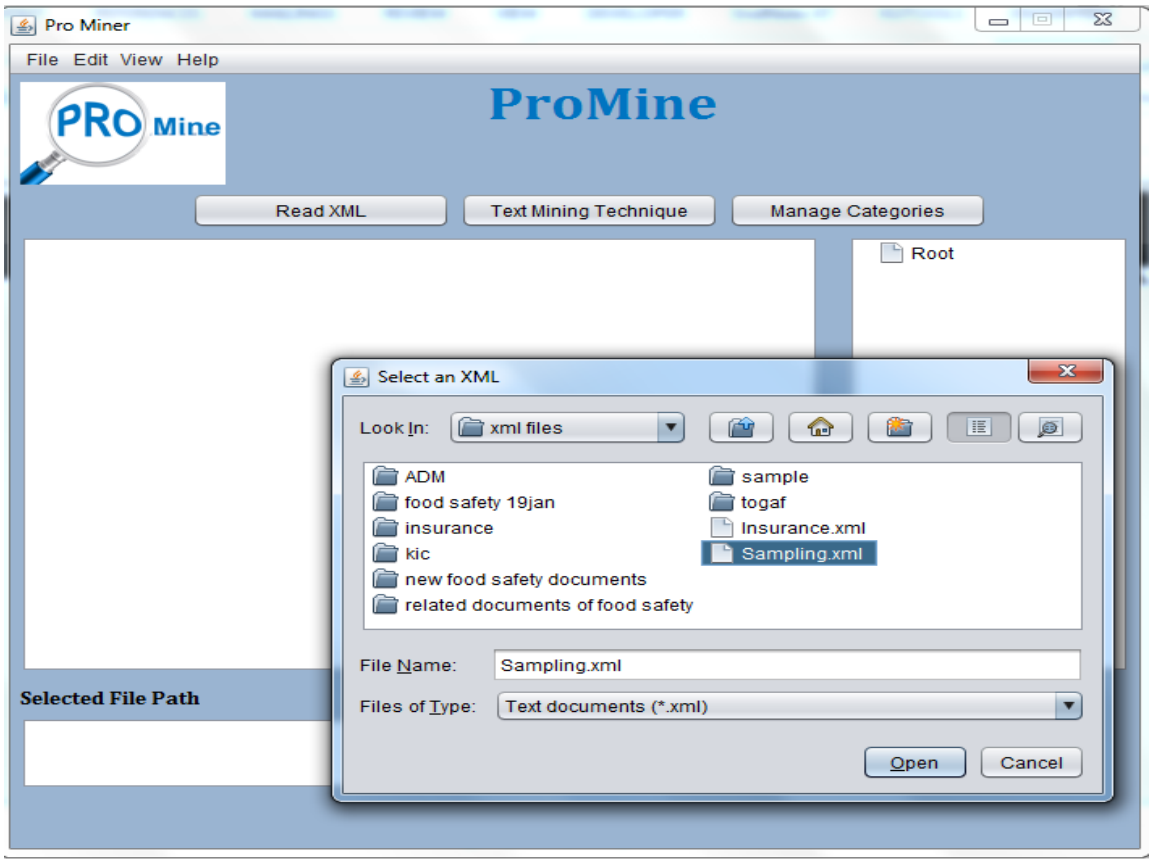

Figure 3-2 Data Extraction Module of ProMine

3.2 Preprocessing of Data

After text extraction, the most crucial part, cleaning of extracted text starts. Preprocessing portrays any sort of transformation performed on unstructured text to set it up in such format that it will be easily and efficiently processed. This preprocessing module ensures that data are prepared for subsequent activities, which are discussed later in the description. Text preprocessing is an integral part of natural language processing (NLP) system. Text preprocessing include in general different NLP and text mining techniques such as tokenization, stop word removal, part-of-speech (POS) tagging, stemming or lemmatization and frequency count.

DOI: 10.14267/phd.2015065

By applying these techniques, the input text is transformed into term vector and the weight of each term is based on the frequency of the term in an input file. For multivariate text analysis, in ProMine, the following preprocessing techniques have been implemented.

Tokenization: In this process unstructured text is segmented into discrete words that called tokens and these words are our processing units. At this stage, word boundaries are defined and this process is totally domain dependent. There are different ways to define these boundaries. For English language text, white spaces or punctuation characters. This process is also called sentence segmentation.

Stop Words Filtering: To reduce the dimensionality of tokenized data, stop word filter is applied. In this process most frequent but unimportant words that have no semantic content relative to a specific domain are removed from the data. . Such type of data have little impact on the final results so can be removed. This list of words is user defined so modification is possible in this list of words. This process is applied to save storage space and to increase the processing.

Part-of-speech (POS) Tagging: This process helps in tokenization and it is necessary to identify valid candidate terms based on predefined PoS patterns. POS removes the disambiguation between homographs and especially in ProMine it will also provide help in the next coming phase of concept enrichment. More detailed description is in 3.3.

Key Term Extraction: At the end of this phase a well-known statistical filter of frequency count is applied to find more interesting terms. For extracting maximum important key terms, I have set minimum threshold.

DOI: 10.14267/phd.2015065

Progression to a subsequent phase of ProMine framework depends on successful progression through the previous phase in order to produce optimal results. Therefore, poor text preprocessing performance will have a detrimental effect on downstream processing.

3.3 Concept Enrichment

At the end of preliminary phase, a set of unique key words is created against each organizational task. This phase can be divided into two steps; first step extracted synonyms from different lexical resources and in second step compound words are made with the help of domain corpus.

A set of key words that came from the description attribute of a task from the input file may not provide enough information to generate knowledge elements for ontology enrichment because this description attribute contains little information about the task. In order to enrich vocabulary of required knowledge elements, some language engineering tools such as WordNet (Miller, 1995) and Wiktionary are used. WordNet is a semantic lexical database that contains a synset against each word and words in this synset are linked by semantic relations (Luong, Gauch, & Wang, 2012). As of the current version 3.0, WordNet contains 82,115 synsets for 117,798 unique nouns. The second lexical database that I have used is Wiktionary (Contributors, 2002), is larger in size as compared to WordNet. Like Wikipedia, any web user can edit it that causes to grow its content very quickly. However, semantic relations in parsed Wiktionary are less than WordNet. Therefore, I have used both WordNet & Wiktionary as an external resources to expand a concept's vocabulary. For every key word that has been extracted after a first phase, I get a set of synonyms from WordNet and Wiktionary. The synonyms are the semantic variants of a given word. In ProMine, there is a flexibility to add domain lexical resource, for example in one experiment on food safety domain, AGROVOC multilingual agricultural thesaurus is

DOI: 10.14267/phd.2015065

also used for capturing more domain related concepts. At the end of this step a combine list of synset is produced against each key word.

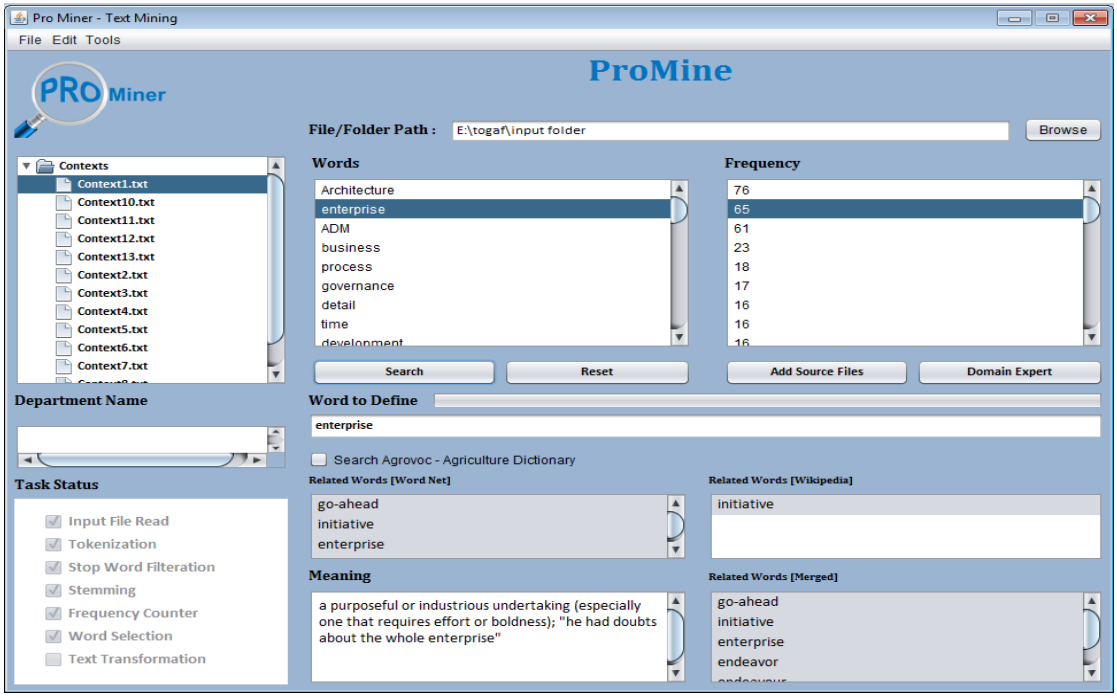

Figure 3-3 ProMine: Text Preprocessing and Concept Enrichment

In the next step of concept enrichment phase, a domain corpus is used. WordNet and Wiktionary are not domain dependent lexical databases. WordNet has different senses of a word so there are chances that many irrelevant words (have semantically same meaning, but can not a part of specific domain) may also generated. There is a need to eliminate such words from this synonym list. For this purpose, a domain corpus is used that includes domain glossaries or legal documents or any domain related published or unpublished documents. ProMine, prepares this corpus by itself. It takes different format (pdf, word, ppt) files and transform them into a text file. After transformation, preprocessing techniques which are described in section 5.1.2 are applied to this domain corpus. Now, a procedure of few steps is applied to this preprocessed domain corpus to filter out above mentioned ambiguities. An important function of this

DOI: 10.14267/phd.2015065

procedure is to extract automatically, a set of domain-specific key-concepts in the form of compound words. Concepts can be more informative in compound or multi-word terms as compared to single words. However, WordNet database provides only few compound words/multiword terms. Therefore, at this step, multiword terms are also stretched from the given corpus because these multiword terms represent concepts that are more important to get meaningful knowledge elements. However, WordNet has different senses of a word so; many irrelevant words (have semantically same meaning, but can not a part of specific domain) are also generated. I have to filter out such candidate words. For this purpose, a domain corpus is required that may include domain glossaries or legal documents or any type of domain related documents. To overcome such ambiguity, I apply a procedure of few steps that will filter out such words as well automatically extracts a set of domain-specific key-concepts in the form of compound words from the domain corpus. WordNet database does not provide all compound words/multiword terms. Concepts can be more informative in compound or multi-word terms as compared to single words. Therefore, at this step, multiword terms are also stretched from the given corpus because these multiword terms represent concepts that are more important to get meaningful knowledge elements. For example for “governance” keyword, the multiword terms can be: "enterprise governance", "corporate governance" and "IT Governance". These multiword terms have different meanings: "Corporate governance broadly refers to the mechanisms, processes and relations by which corporations are controlled and directed" (Shailer, 2004), while “enterprise governance is the set of responsibilities and practices exercised by the board and executive management with the goal of providing strategic direction, ensuring that objectives are achieved, ascertaining that risks are managed appropriately and verifying that the enterprise’s resources are used responsibly" (Selig, 2008) and “IT

DOI: 10.14267/phd.2015065

governance" is the responsibility of executives and the board of directors, and consists of the leadership, organizational structures and processes that ensure that the enterprise's IT sustains and extends the organization's strategies and objectives (ITGI, 2007). All these words have different meanings and all multiword concepts are important in IT service management.

All candidate words with the key-word will pass through the following described procedure.

1. As preprocessing has been applied on the corpus. Two-word noun compounds (bigram) via the POS tags are extracted from the corpus. The noun - noun compound is a common type of multiword expression in English. From these two-word noun compound, one word is our candidate word from a candidate list of words and other word is from the corpus. Every candidate word is passed in the corpus and if a noun is found either its right or left, it is joined to the candidate word and make a compound (bigram) word. If no noun word is found on the right side or left side of the candidate word, it keeps it as a single word (unigram). During joining it is noted that if nouns are separated by full stop or comma (punctuation marks) then system will not join such two nouns.
2. Once the compound words are identified automatically, the next step is to count the frequency of all words including unigram and bigram. Then a user defines a threshold frequency. If any candidate word does not occur in the corpus or its recurrence is below than a defined threshold, then this word will be dropped from the list. In this way all irrelevant words from the list of synonyms are also dropped because if some synonyms are not found in the corpus, they automatically eliminated from the output list. If any compound word is below the threshold, then our system will check the other content word (not candidate word) and if it passes the frequency threshold, then it will remain

DOI: 10.14267/phd.2015065

in the list, but if the second content word will not pass the frequency threshold then it will remove from the list.

3. As a result of this phase, a rich list of concepts against each key word will be generated.

I also did a trigram compound word experiment but it didn't bring any valuable information. I already get more information with the two word nouns (bigram) selection.

3.4 Concept Filtering based on Semantic Similarity Measure

Though, till last phase unrelated terms (conceptually, not related to a specific domain) from a set of synonyms terms (from WordNet & Wiktionary) of a given key term has been removed. However, the resultant word list consists of lexical terms which are hundreds in number. This high dimensionality of the feature space is the major particularity of text domain. These unique concepts or potential concepts are considered as feature space, these lists of concepts can be considered as high dimensional and sparse vectors. In our proposed framework, at this stage, I am reducing feature space by selecting more informative concepts from this concept list by using a concept filtering method. Conventionally, in most ontology learning tools, statistical measures such as TF-IDF, RTF, entropy or probability methods are used for filtering process (Cimiano & Völker, 2005). To identify important lexical terms, ProMine used an innovative approach that is the combination of statistical and semantical measures. I have proposed a new hybrid semantic similarity measure to identify relevant ontological structures for a given organizational process. This module consists of two phases; in first phase for each candidate concept its information gain (IG) is calculated by using domain corpus and in second phase to find more semantically representative candidate concepts I proposed our hybrid semantic

DOI: 10.14267/phd.2015065

similarity measure that uses different information sources such as lexical semantic network (WordNet) and domain corpus.

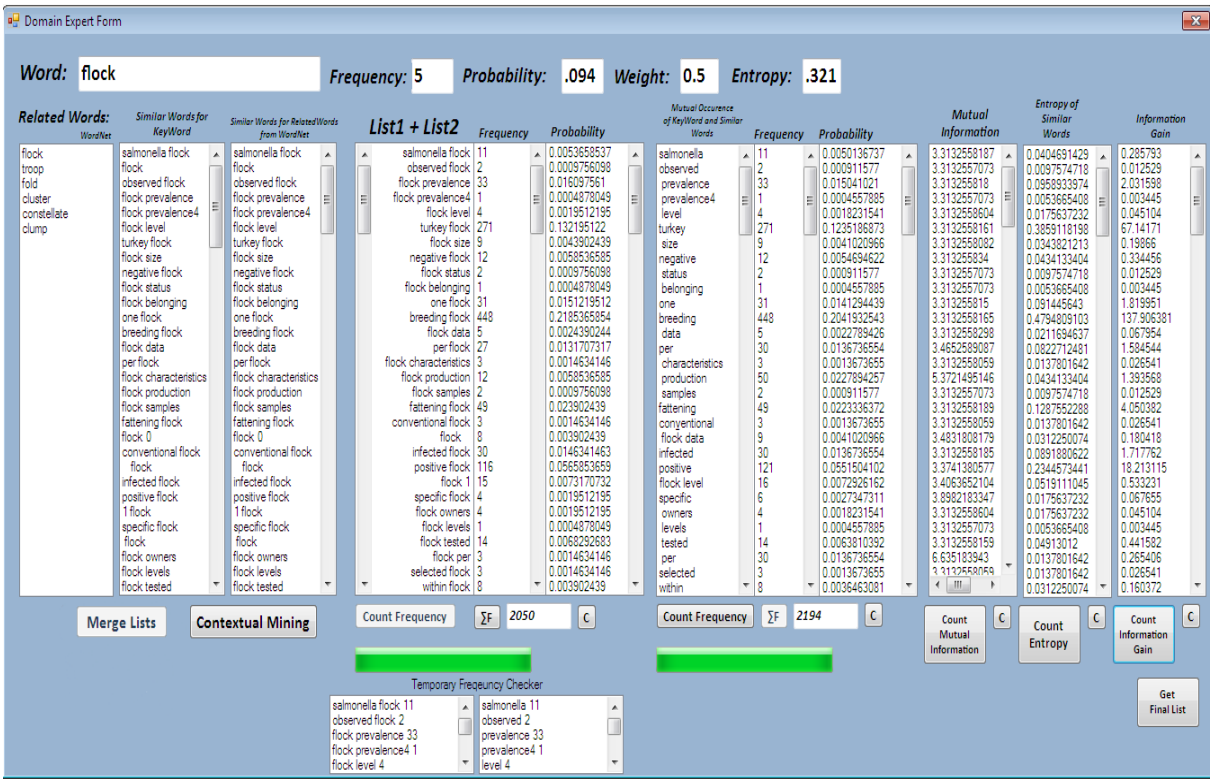

Figure 3-4 ProMine: Concept Filtering

3.4.1 Statistical Syntactic Measure (Information Gain)

ProMine uses Information Gain as a term goodness criterion (Yang & Pedersen, 1997) . I find out IG for all potential terms. First, I calculate entropy, which is the measure of unpredictability and provide the foundation of IG. Entropy is defined as

$$Entropy = -\sum_{i=1}^{m} P_r\,(c_i) log Pr(c_i) \qquad (1)$$

Where $\{c_i\}_{i=1}^{m}$ is the set of words in the target space (synonym set of key word).

After calculating entropy, I have to find out probability with respect to candidate concept by following equation

DOI: 10.14267/phd.2015065

$$P_r(t) \sum_{i=1}^{m} P_r\,(c_i|t) log Pr(c_i|t) \qquad (2)$$

Where $P_r(t)$ represents candidate concept.

Now with the help of equation1 and equation 2 information gain (IG) will be found out. On the basis of information gain (IG) all candidate concepts are ranked and concepts with lowest information gain will be removed by defining a threshold value. The information gain $IG(t)$ of a candidate concept with respect to the key term is defined as

$$IG(t) = -\sum_{i=1}^{m} P_r\,(c_i) log Pr(c_i) + P_r(t) \sum_{i=1}^{m} P_r\,(c_i|t) log Pr(c_i|t) \qquad (3)$$

At the end of this step, I have information gain for all candidate concepts. This information gain of each candidate is used in our proposed hybrid similarity measure.

3.4.2 A New Hybrid Semantic Similarity Measure (Cloud Kernel Estimator)

WordNet is a lexical semantic database and the graph nature of WordNet makes this resource a perfect candidate to find semantic similarity between concepts. I view this database as a semantic graph whose nodes are concepts and to find the similarity between concepts we measure the length of path between these concepts. For example, in figure5 the semantic distance between node {boy, child, male child} and {girl, female child, little girl} is 4, {boy, child, male child} and {teacher} is 6. LCS (least common subsumer) is the most specific common concept of the two synsets. I consider LCS to take shortest path between two concepts. A longer path length indicates less similarity, so ‘boy’ is more similar to ‘girl’ as compared to ‘boy’s similarity with ‘teacher’.

DOI: 10.14267/phd.2015065

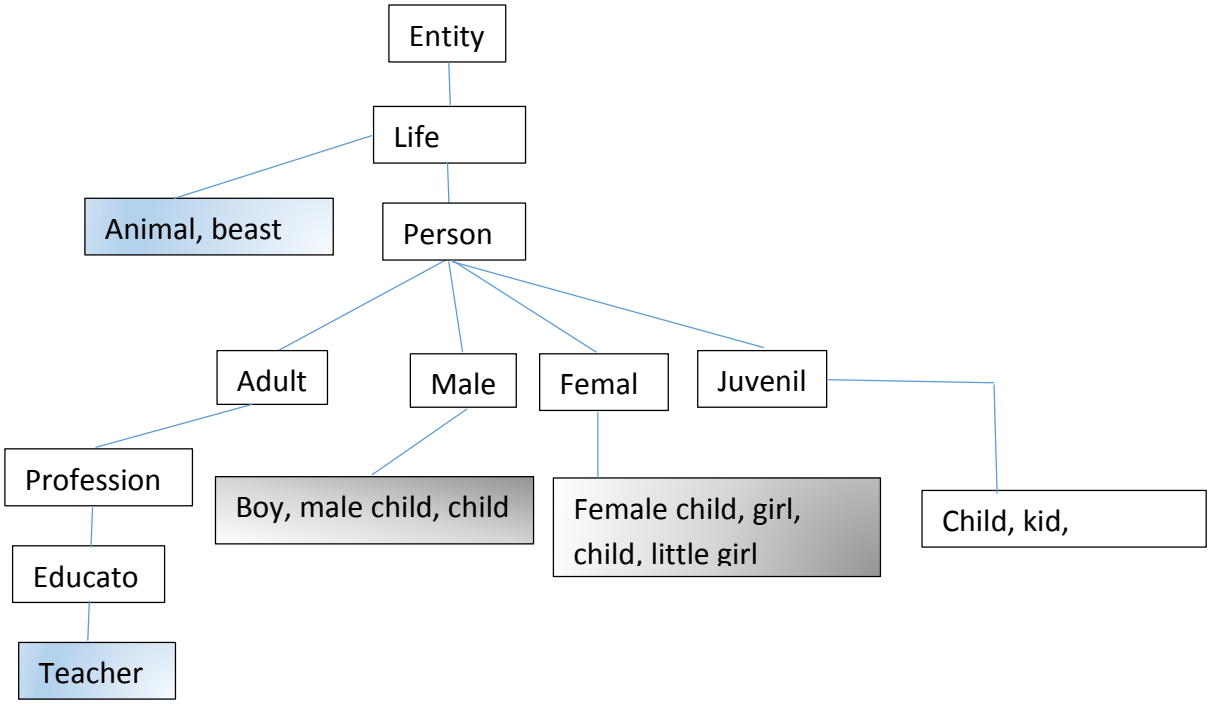

Figure 3-5Hierarchical semantic knowledge base

To find similarity based on only shortest path count attribute cannot suitable for larger networks (Miller, 1995). For example, the shortest path from boy to animal is 4 that is less than from boy to teacher, but, it can't say that boy is more similar to animal than teacher so to overcome this weakness, some other attribute must be added that can provide more information from the hierarchical semantic nets. The higher levels of WordNet hierarchy have more general concepts with weaker similarity while, concepts at lower levels of the hierarchy have more specific concepts with stronger similarity (Yuhua Li, Bandar, & McLean, 2003). Thus, to consider these hierarchical levels should be considered. This depth of concept in the hierarchy can be an influential attribute for similarity measure. As aforementioned, the main objective of this research is to enrich and populate the domain ontology so there is need to incorporate domain corpus statistics to identify the degree of similarity between concepts. Therefore, in my

DOI: 10.14267/phd.2015065

proposed semantic similarity measure the third attribute is the statistical information (Information Gain) of the domain corpus. This statistical information of concepts in a huge corpus allows our method to be adaptable to different domains. In literature, different methods are used to calculate corpus statistics, for example latent semantic analysis (LSA) (Dennis, Landauer, Kintsch, & Quesada, 2003; Foltz, Kintsch, & Landauer, 1998), Hyperspace Analogues to Language (HAL) (Burgess, Livesay, & Lund, 1998; Turney, 2001) and information content ((Yuhua Li et al., 2003; Resnik, 1995). Based on these considerations and empirical results of all three aforementioned individual attributes, I proposed a new hybrid similarity measure which combines the shortest path length, depth and information gain of concepts. Since, it is believed that semantic similarity depends not only on multiple information attributes rather these attributes should be properly combined processed.

The main idea of this cloud kernel estimator is that the similarity between two concepts c1 and c2 is a function of the attributes path length, depth and information gain (IG) as follows:

$$Similarity(c1, c2) = f(len, depth, IG) \qquad (4)$$

Where,

len is the conceptual distance between two nodes $(c1, c2)$ which is also known as the shortest path length between $c1$ and $c2$.

$depth$ is the depth of concept nodes

IG is the information gain of $c1$ and $c2$

We assume that (4) can be rewritten using three independent functions as:

DOI: 10.14267/phd.2015065

$$Similarity(c1, c2) = f(f1(len), f2(depth), f3(IG)) \quad (5)$$

Path length and depth is calculated from lexical database WordNet while IG is derived from the domain corpus as mentioned in section 5.4.1. The details of these $f1$, $f2$ and $f3$ are as follow:

Definition 1: The conceptual distance between two concepts (herein known as Path Length Attribute) is proportional to the number of edges separating the two concepts in the hierarchy.

$$f1_{length}(c1, c2) = (2 * deep\_max - len(c1, c2))/2 * deep\_max \quad (6)$$

$len(c1, c2)$ is the length of the shortest path from $c1$ to $c2$ and

$deep\_max$ is the maximum depth of the semantic hierarchy that is a fixed value for a specific version of lexical database (WordNet). $2 * deep\_max$ is the maximum value that $f1(len)$ can get.

Definition 2: D Depth is another factor that affects the similarity between words. As we know that concepts at upper layers of the hierarchy in semantic networks (WordNet) have more general semantics and less similarity between them, while concepts at lower layers have more concrete semantics and stronger similarity. This shows the importance of depth attribute to find similarity between concepts. Our measure's this attribute is based on Wu and Palmer (Z. Wu & Palmer, 1994) measure that is simple, and gives good performance.

$$f2_{depth}(c1, c2) = \frac{2 \times depth(LCS(c1,c2))}{depth(c1)+depth(c2)+2 \times depth(LCS(c1,c2))} \quad (7)$$

Where $depth\ (c1)$ and $depth\ (c2)$ are the depths of $c1$ and $c2$ on the path through lowest common subsume (LCS) of $c1$ and $c2$.

DOI: 10.14267/phd.2015065

$depth(LCS(c1, c2)$ is the distance which separates the lowest common subsume of C1 and C2 from the root node

For example if we want to find similarity between boy and teacher and boy and animal. This calculation is done as follows:

Depth (boy, teacher) = $\frac{2\times2}{2+4+2\times2} = 0.4$

Depth (boy, animal) = $\frac{2\times1}{3+1+2\times1} = 0.33$

This example can show the importance of this attribute in our cloud kernel estimator because as I mention earlier that shortest path length value between boy and teacher is greater than shortest path length between boy and animal but it doesn't mean that boy is more similar to animal than a teacher so for such cases our measure's depth attribute provide help for more accuracy in results. As definition 2 shows that 0 < score <= 1. The score can never be zero because the depth of the LCS is never zero (the depth of the root of a taxonomy is one). The score is one if the two input synsets are the same.

Definition 3: In our approach, to measure the similarity of two concepts, we use information gain (IG) measure as a third attribute. In section 3.3.1 the detail description of IG is given. Information gain computes a relevance score that measures the similarity between the key concept and candidate concept.

The information gain $IG(t)$ can be rewritten as $f3$ as:

$$f3(IG(t)) = -\sum_{i=1}^{m} P_r\,(c_i) log Pr(c_i) + P_r(t) \sum_{i=1}^{m} P_r\,(c_i|t) log Pr(c_i|t) \qquad (3)$$

At the end of this step, we have information gain for all candidate concepts.

DOI: 10.14267/phd.2015065

By applying this similarity function I filtered out important terms which are the potential concepts for the domain ontology. Here I called them knowledge elements.

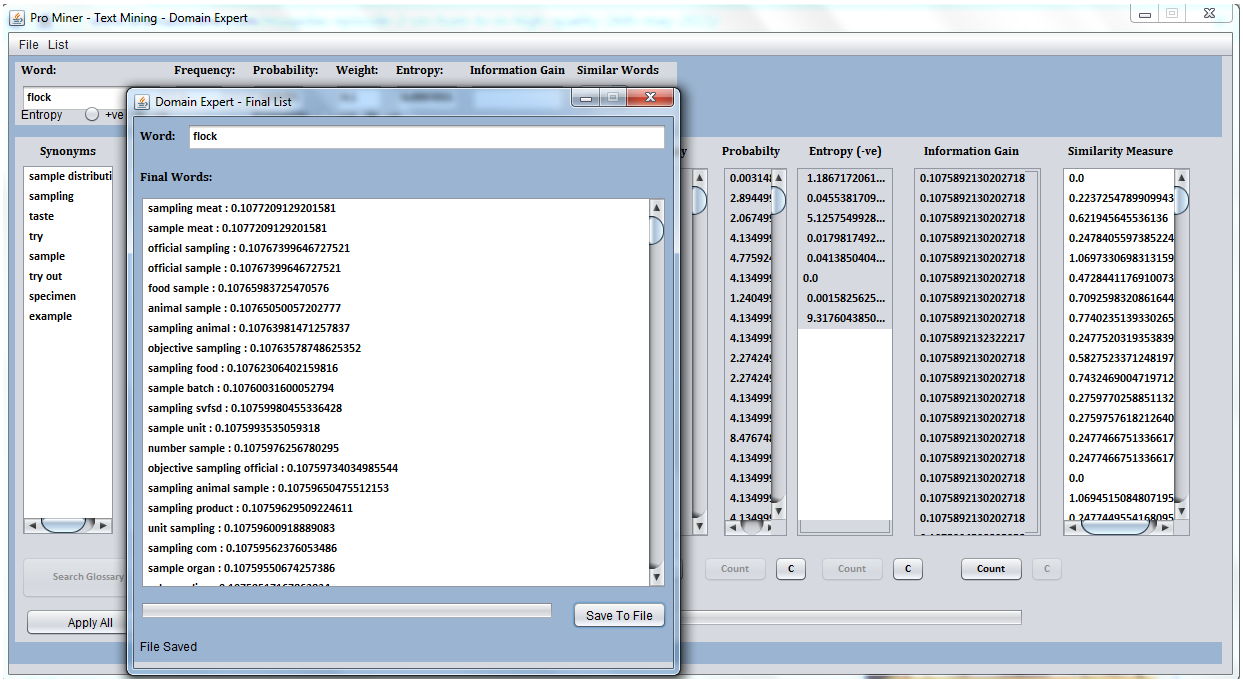

Figure 3-6 Concept Ranking and Selection

At this point some general rules and mathematical lemmas need to be expressed based on the equations aforementioned.

Lemma 1

The depth of lexical tree is monotonically linked to the similarity pair of two semantically related word.

Proof

Given the lexical tree T, there is a root node Nr. The set child nodes of Nr comprised of those concepts (nodes) which are subsumed by the underlying concept of Nr. We denote this set as SNr with elements $C_1, C_2, \dots C_n$. The set and the root node SNr poses two mathematical properties

DOI: 10.14267/phd.2015065

$$SNr \neq \Phi \text{ ---------- (4)}$$

$$SNr = C_1 \cup C_2 \ldots \cup C_n \text{ ---------- (5)}$$

The equations above indicate that the set SNr is essentially a non-empty set. Secondly, the semantic concept of every element of the set is a specialization of the generalized concept of root node. This illustration provides the notion that concepts at upper layers of the lexical hierarchy have more general semantics and less similarity between them, while concepts at lower layers have relatively more concrete (or specific) semantics resulting into stronger similarity. Therefore, the depth of concept plays an important role in identifying the proposed functional projection. Hence it can be concluded that the depth of the hierarchy is directly linked the degree of semantic relationship between two terms.

Lemma 2

In general, statistical information is a function of probability of key term and probability of the semantically related concept.

Proof

From the statistical theory, we know that the statistical information involves the assumption that whatever is proposed as a cause has no effect on the variable being measured. According to this assumption, for a given word W and its related concepts holds a statistical information for semantic interpretation. We know that the information cannot be data less. At the bottom level, it must be composed of a unit datum. A datum is a putative fact regarding some difference or lack of uniformity within some context. There is no concept existing which delivers the idea of lacks of uniformity in the real world out there. This leads to the idea of lack of uniformity

DOI: 10.14267/phd.2015065

between two or more physical states of concepts. An example to such phenomenon is a higher or lower charge in a battery related to the variable electrical signal in a telephone conversation. It indicates that the probability of the key term and probability of semantically related concepts must hold the statistical information.

Lemma 3

Information gain (IG) derived from domain corpus and path length and depth calculated from the lexical database (WordNet) play a non-trivial role in calculating the cloud kernel estimation in the domain of business management. The cloud kernel estimator is always a non-negative function.

Proof

It is useful that we describe some core definitions to prove the claim aforementioned. Cloud kernel estimator can be expressed as a function of following factors

Where:

len = the length of the shortest path from $C_1$and C_2 in the lexical hierarchy of WordNet.

$deep\_max$= The maximum depth of the semantic hierarchy

$depth\ (c1)$= The length of the path to $c1$ from the global root entity in the hierarchy

$LCS(c1, c2)$= The lowest common subsume of C_1 and C_2

IG= Information Gain

DOI: 10.14267/phd.2015065

Given (C_1, C_2), the length is the hierarchical level of derived lexical database contribute towards the formulation of cloud of semantic concepts for a given word C1. We already have defined Path Length Attribute (see definition 1), accordingly the distance between the concepts is characterized by the shortest path between two concepts and $deep\_max$ is the depth of tree T. The second factor $depth\ (c1)$ is the length of the path to $c1$ from the global root entity and $LCS(c1, c2)$ is the lowest common subsume of $c1$and$c2$. The third factor is probabilistic inspired factor which calculates the information gain of same two concepts. Cloud kernel Estimator can be conclusively expressed as the function of three attributes as below.

$$\text{Cloud kernel Estimator (CKE)} = \text{f}(len, depth, IG)$$

This is equivalent to

$$\text{CKE}=\big(2 \times deep_{max} - len(C_1, C_2)\big)/2 \times deep_{max} + \left(\frac{2\times depth(LCS(c1,c2))}{depth(c1)+depth(c2)+2\times depth(LCS(c1,c2))}\right) + \big(-\sum_{i=1}^{m} P_r\ (c_i) logPr(c_i) + P_r(t) \sum_{i=1}^{m} P_r\ (c_i|t) logPr(c_i|t)\big)$$

It is a known fact that if $len(C_1, C_2)$ which is always a non-negative value. The second factor is quite evident to be non-negative, the fact is coined in that depth of the lexical tree can never be negative therefore the LCS value will such a value that is between 0 and 1. In third factor (IG) the probability of every candidate concept is always below 1 but greater than zero. We know that the log of every positive value smaller than 1 is always negative, the summation of all these logs have been multiplied by -1. It is clear that the third factor is also a non-negative function. Hence, it is proved the lexical cloud kernel estimator is always a non-negative function.

DOI: 10.14267/phd.2015065

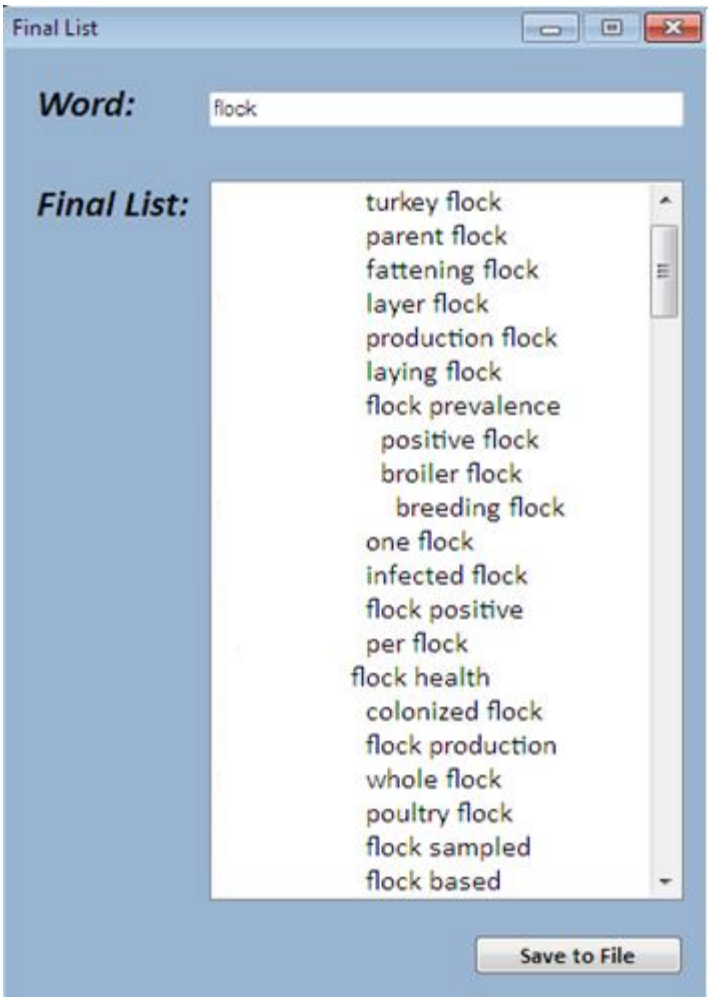

Figure 3-7 Final List of Concepts

3.5 Semantic Concept Categorization

ProMine architecture performs basically two main tasks; one is concept extraction and the second is semantic concept categorization. At the end of third phase, we have a refined list of domain specific concepts against each key term that was selected in the second phase. These extracted concepts are semantically similar to the key term. Now, we want to categorize these concepts in such a manner that ontology to be enriched. For this purpose, it is necessary to find out concept relationships between these words and existing (seed) domain ontology. We

DOI: 10.14267/phd.2015065

proposed a novel semantic concept categorization method to enrich an existing ontology. This method will classify new domain-specific concepts according to the existing taxonomy of the seed ontology. For concept categorization, this method will use the knowledge of existing concept categories (taxonomy of classes) of the ontology with the help of external knowledge resources such as Wiktionary.

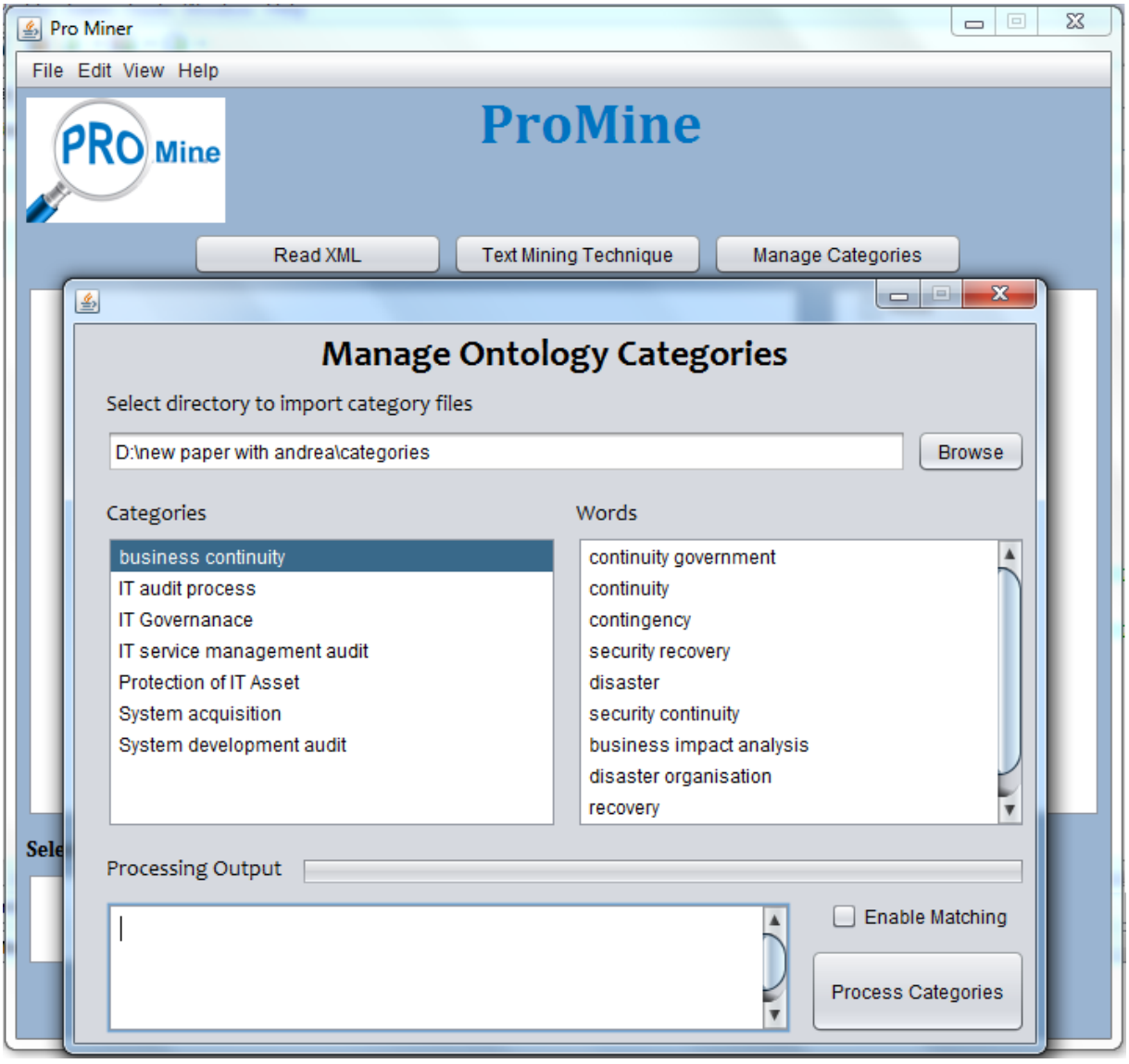

Figure 3-8 ProMine: Semantic Concept Categorization

DOI: 10.14267/phd.2015065

The proposed approach tries to find a semantic similarity between extracted concepts using some fragment of the ontology that describes a certain category. The outline of the approach is presented below:

1. Select the target categories that are defined in the ontology represented as "CATi". Where CAT represents category and I is iterator. We know that every category "CATi" hold an arbitrary number of concepts denoted as CNij where CN represents concepts, "i" is category iterator and "j" is concept iterator. Some of these concepts may belong to more than one category (as a member of the hierarchy). For example, "Risk" is a concept falls under the category of "IT Audit Process", "IT Governance" and "System Acquisition". Such an overlapping begets another challenge of the underlying problem.

2. For each concept of every category, a set of related terms that will include synonyms and derived terms will be realized using Wiktionary. I denote these terms as TR_CNijk where k is iterator of every term in the set. This set represents semantically similar words of the selected concepts. For example, if one category contains ten concepts, then 10 sets of semantically similar words will be prepared at the end of this step.

3. The third step of ProMine is concerned with matching. The system takes every potential candidate from the given set of concepts say PCNx where "x" represents the counter of every Potential CoNcept (PCN). This step checks its relevance or semantic relationship to each concept CNij in every category CATi. Here one essential aspect is the decidability of inclusion of potential concept PCNx in more than one category CATi. Although, it is arguable either to use a strict threshold based criteria or restricting it to the highest relevant category. I justify that the potential candidate can be put into more than one category.

DOI: 10.14267/phd.2015065

```
Let us have a relation "R" that associates every CONCEPT to its related
CATEGORY:
R = {con : CONCEPT; cat : CATEGORY| con ↦ cat}
Let Miscellaneous = {c : CONCEPT| c has unknown category}
Let W = { set of new words } ← process(document)
For all i ∈W{%% Repeat for every new word i
    For all j ∈range(R) { %% Repeat for every category j
       K_match = ϕ
       M_match =ϕ
        For all k ∈ domain ({con ↦ cat : R|cat == j}{
        %% Repeat for every concept k in category j
           M = { set of Synonyms and Related words of k }
           %% Obtained from Wikipedia
           IF (i == k)
             K_match = k_match ⋃{ j }
           ELSEIF (i∈M)
            M_match = M_match⋃{ j }
}
}
  IF (|k_match| ≥ 1)
     do nothing
  ELSEIF (|M_match| == 1)
     R = R⋃{i ↦ j} where j∈M_match
   ELSEIF (|M_match| > 1)
     Apply Similarity measure and ADD in category with most match
ELSE
   Miscellaneous = Miscellaneous ⋃{ i }
}
```

Figure 3-9 A Proposed Semantic Concept Categorization Algorithm

4. It can be noticed that one every single pass of the algorithm, each potential candidate PCNx is put into one or more category. For the next step, this "fresh" concept (like other peer concepts) will also be expanded into its relevant synonyms.

5. If there is any concept observed with no match in either of the categories, then it will be subjected to "Miscellaneous" category. As the whole process is exhausted, domain expert will manually identify their respective categories.

DOI: 10.14267/phd.2015065

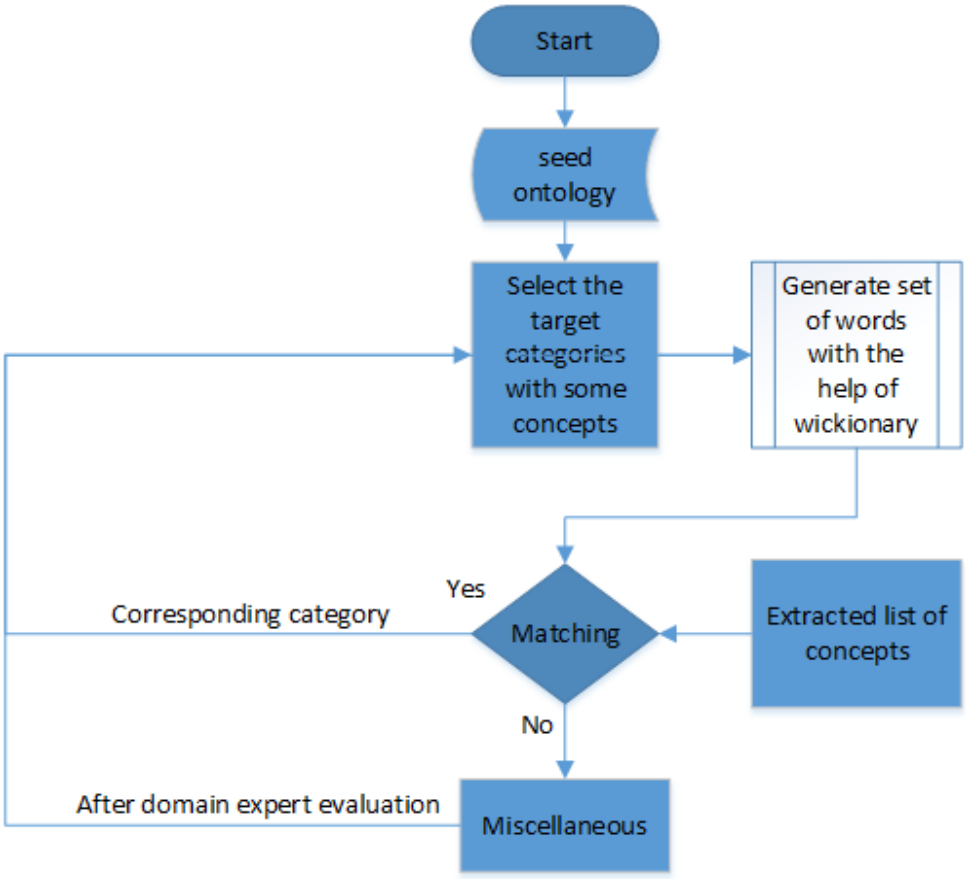

Figure 3-10 ProMine: Semantic Concept Categorization Procedure

The five functional steps ensure that the proposed ontology enrichment algorithm will populate the given taxonomy with new concepts that are extracted from domain corpus and other external sources.

DOI: 10.14267/phd.2015065

Sampling in Controlling Food Safety: Case Study1

As I mentioned earlier, this research is a part of PROKEX project and one objective of this project is to inspect the complex processes of the food supply chain. The reason for selecting this domain is the complexity of the related tasks and the problems occurring during the everyday execution.

4.1 Case Background

This ProMine case focused on the sampling process as a testbed. The reason for selecting this sampling process is; first of all this was the requirement of our pilot partner, the National Food Chain Safety Office and sampling is the most widespread activity of this office. As a national control authority, doing the control of the food safety, one important step is to check regularly the status of food in the distribution network. This is done via collecting samples from different group of articles and according to a previously worked out calendar. Since the commercial units cover the territory of the whole country, at this moment there are over 20 subordinated agency doing the sampling based food control. Sampling is an official procedure, the processes should be very highly standardized. Food safety requirements cannot be different at different places. It is very important to ensure that all the sampling performance within country applies the same knowledge and employees understand the same technical procedure. The rest of the food safety is more concentrated because samples are taken into labs and there are only four or five labs in the country. The complexity of the related tasks and the problems occurring during the everyday execution. NEBIH is the National Food Chain Safety Office in Hungary responsible for all aspects of the food chain safety supervision. NEBIH has its standard sampling procedure, and

DOI: 10.14267/phd.2015065

the annual sampling programme. Thousands of on-site sampling and inspections are carried out each year by the geographically dispersed task forces of NEBIH.

The real complete process starts with the planning of sampling meaning about 3000 product/parameter pairs and 100.000 samples during a year. The planning of sampling is centrally prepared, the execution is done by a staff of approx. 1000 people organized regionally. The detailed timing and task breakdown of the individual sampling events is done by the regional management. The sampling process (main focus process) starts with the pre-set timing and task breakdown. The sampling process includes the preparation of sampling, execution described by specific regulations, shipping of samples to accredited laboratories, related documentation and input in the Public Authority internal IT system. The sampling process ends at the point when the samples are arriving to the laboratories. Complexity of sampling is caused by the small differences depending on the scope. Depending on the goal of the investigation and the parameters of sampling, different portions, methods, devices and documentations have to be used, on top of it the goals and parameters may also change in line with some high-risk events in the food chain. Sampling process is a very good candidate for data and text mining, because the related regulations are various and changing fast. The domain of food safety is a strongly regulated environment: EU legislation, national legislation and in-house Public Authority regulations are deeply described, thereby causing strong difficulty for the sampling staff to have always fully updated and actual information.

4.2 Case Objective and the Related Research Question

The purpose of this case study is to determine whether the proposed solution ProMine can do as well, or better the key knowledge elements' identification within sampling process. This goal

DOI: 10.14267/phd.2015065

is related to research question 2 that whether a text mining based solution can be used to enhance the existing knowledge that captured from the business process and can we used this knowledge for ontology learning?

Through this case study, if we become able to automatically extract knowledge elements then this captured knowledge can be passed more quickly and easily to seed ontology, and provides regular update for the related knowledge base (the ontology) as well. This concept extraction system can be used for other domains as well. To identify these key concepts manually is a tedious and time consuming work for a domain expert and especially when corpus size increases overtime then it becomes impractical to search manually through very large databases. This research aims to overcome this difficulty by automating the process of extracting information and converting it into a useful source of knowledge. Therefore, I have chosen this case study to evaluate the results of concept extraction part of ProMine.

4.3 ProMine Context (or Application of ProMine)

4.3.1 Datasets Description

In order to apply our concept extraction method for ontology learning, I have taken two types of data. The initial input come from business processes which is the description of tasks related to sampling process. This is in the form of XML file. However, this tasks' description is not in big size so we need some other sources as well to enrich our knowledge for ontology learning. As I mentioned above, our focus domain for this case study is food safety sampling process. Therefore, a domain text corpus had to be prepared as well as the relevant domain outsources like AGROVOC multilingual agricultural thesaurus, an agriculture dictionary is also included for capturing more domain related concepts seed ontology with some sample concepts.

DOI: 10.14267/phd.2015065

The text corpus examined is taken from the Official Journal of the European Community (OJEC) (Cheli, Battaglia, Gallo, & Dell'Orto, 2014; Directive, 1979). OJEC is now recognized as OJEU. It is the official gazette of record in which all tenders from the public sector which are valued above a certain financial threshold according to EU legislation are published for the European Union (EU). It is published every working day in all of the official languages of the member states. Some published white papers (Commission of the European Communities, 1999; Daviter, 2009). Books (Alemanno, 2006; Dreyer & Renn, 2009; Schmidt & Rodrick, 2003) are also added in this corpus. Legal acts and regulations published in the EUR-Lex (EUROPA, 2014) that provides free access, in the 24 official EU languages, to the Official Journal of the European Union, EU law (EU treaties, directives, regulations, decisions, consolidated legislation, etc.), preparatory acts (legislative proposals, reports, green and white papers, etc.), EU case-law (judgements, orders, etc.) international agreements, EFTA documents and other public documents are part of the corpus as well. Some material is also taken from the official website of National Food Chain Safety (National Food Chain Safety, March 2012). Text preprocessing techniques including NLP techniques are applied to this corpus as detailed in the following section.

4.3.2 Empirical Evaluation

In this sampling case, from process model the XML file is generated, in which some description about the different tasks (steps) of "execution of sampling" are defined. ProMine take this XML as an input data and extract text from some specified tags that contain a description of the sampling process and save this extracted text into different text files (each for different task of sampling process). After extracting text, preprocessing techniques that are mentioned in section 3.1.1 are applied on this data and find out all key terms of each task. Now each key term is sent

DOI: 10.14267/phd.2015065

to WordNet for expanding the list of related words to find more information elements related to the sampling process. After finding related words of a keyword, then these related words along with a keyword match with domain related documents and find some more relevant single and compound words as mentioned in section 3.2. In this way, a long list of related words of keyword comes as an output. Once a set of key-concepts has been extracted from a domain corpus, the next empirical step is to filter out most domain related words, we used the information gain method to rank all words, and by defining a threshold most informative words can be selected as shown in Figure 4-1.

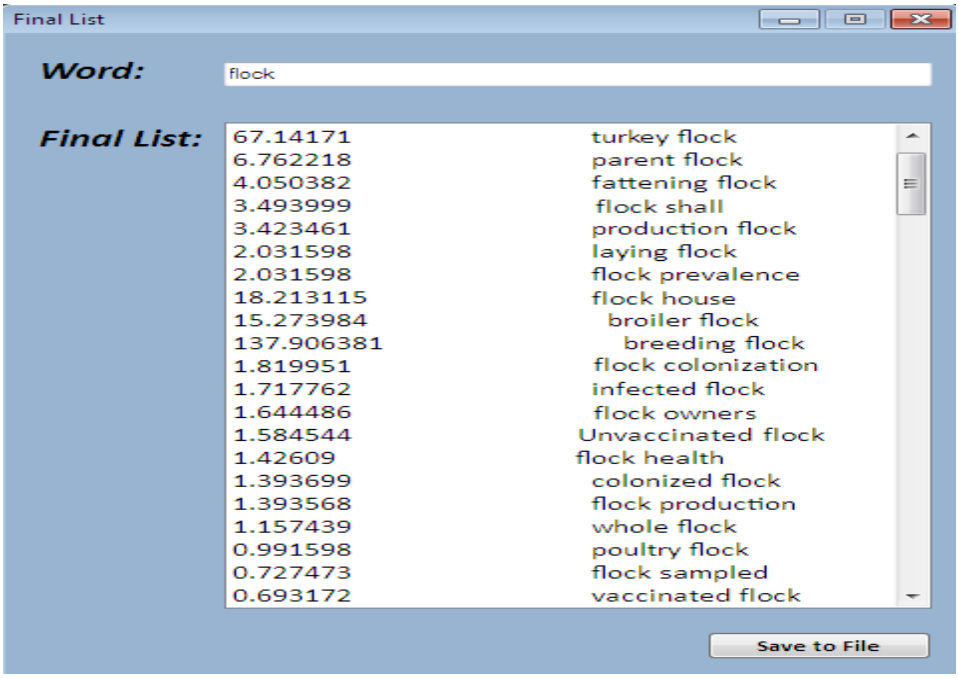

Figure 4-1 ProMine: Final list of knowledge elements

After extracting these concepts list, with the help of domain expert these concepts are categorized according to seed ontology classes. This process showed that ProMine is working in right direction. There were some limitations, like it just extracted two words (compound words) concepts but some three words can also represent a concept so in next iteration of

DOI: 10.14267/phd.2015065

ProMine three words concepts are also extracted. This case study has been done twice; once after first initial iteration of ProMine and second time Wiktionary and AGROVOC multilingual agricultural thesaurus, an agriculture dictionary for capturing more domain related concepts were included. In this case study I was extracting knowledge elements from business process with the help of some outsources and after extraction, I was filtering these knowledge elements by applying statistical measure (Information Gain). Results of this case study are shown in figure 4.2 and table 4.1.figure 4.2 and table 4.1.

4.3.3 Results and Discussion

Table 4.1 shows the evaluation statistics, where check1 and check2 correspond to the two rounds of the checking. Concepts are the total concepts that ProMine extracted, accepted means the meaningful concepts of the food safety domain in extracted list. This column is further categorized in to two columns; filter column and unfiltered column. Filter column shows the concepts that are filtered out by applying statistical measure. Unfiltered column shows the meaningful concepts that are not selected by the statistical method (Information Gain measure rank these concepts in low ranking so by applying threshold they are omitted) though these should be included in the final results and last column is unimportant words with respect to domain ontology.

As shown in the table 4.1, in check2 number of concepts increased because we also included Wiktionary and EGROVOC so ProMine extracted more domain related words. If we see total number of accepted concepts (that can be a part of domain ontology) are more than 70% of the total extracted concepts that shows that our approach of ProMine is working in right direction.

DOI: 10.14267/phd.2015065

Table 4.1. Evaluation of extracted concepts

Evaluation	Total Concepts	Filtered Concepts	Accepted			Unaccepted Filtered
			Filtered (1)	Unfiltered (2)	(1)+(2)	
Check1	272	191	113	81	194	78
Check2	475	333	196	142	338	137

However, some useful concepts are also discarded by ProMine that should be in final list of concepts. Figure 4.2 is showing the manual hierarchical structure of food safety domain ontology that is made from the concepts that are extracted by ProMine, this hierarchical structure shows that ProMine is working in right direction.

DOI: 10.14267/phd.2015065

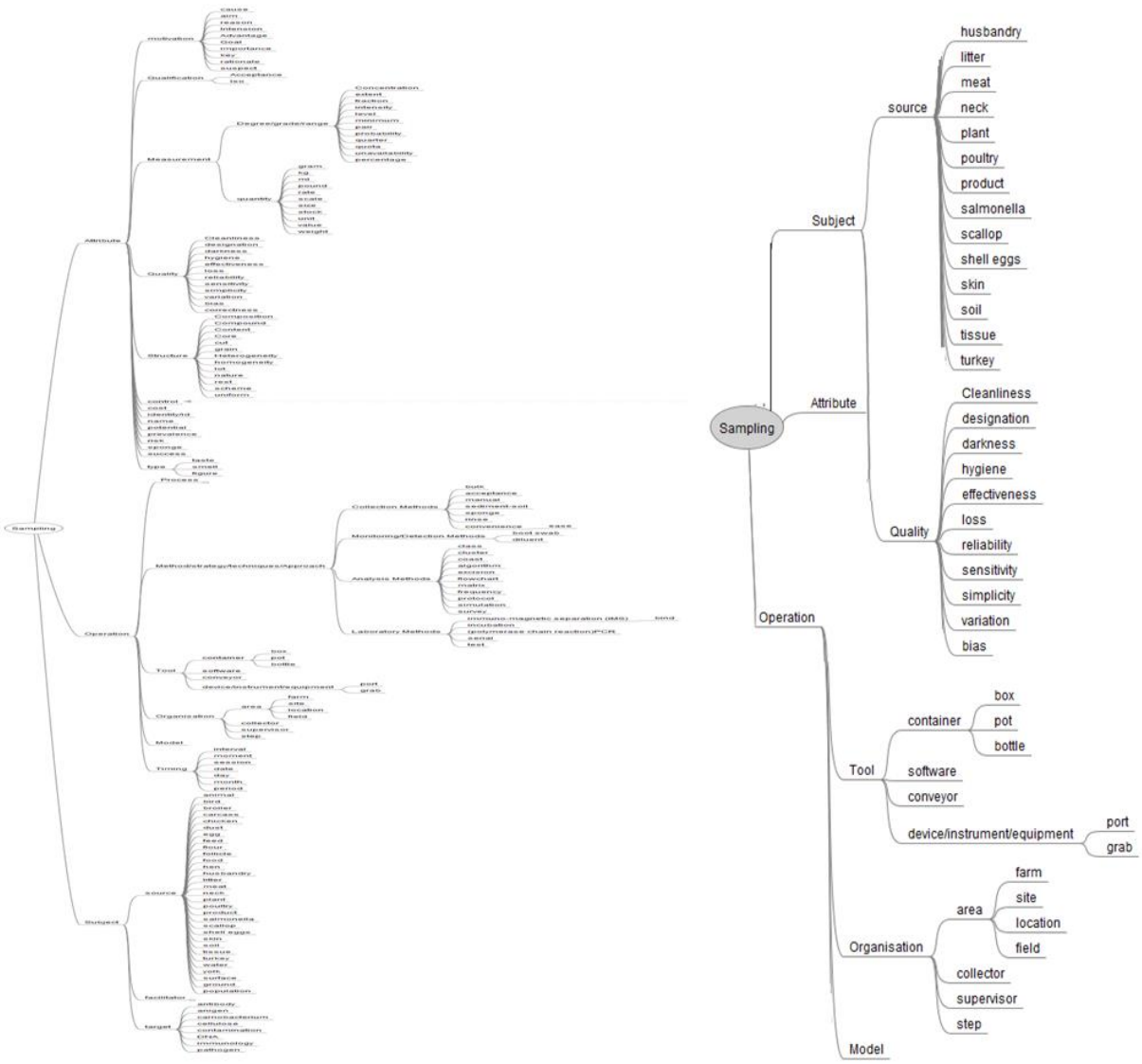

Figure 4-2Manual categorization of extracted knowledge elements

4.4 Case Conclusion

With the help of domain expert evaluation some impressions can be noted from this case study: ProMine concept extraction approach can support domain experts and ontology engineers in building domain specific ontologies efficiently. However, to reduce rate of concepts that should be in final concept list, some other methods besides information gain should be also considered. As a weakness of the IG criterion is that it is biased in favor of features with more values even

DOI: 10.14267/phd.2015065

when they are not more informative (Novaković, Štrbac, & Bulatović, 2011) and as I mentioned earlier this is pure statistical measure so we have to consider some semantic measures as well for better results. Therefore, I propose a new hybrid similarity measure (Cloud Kernel Estimator) for concept filtering process. The next case study is about that measure.

DOI: 10.14267/phd.2015065

Insurance Product Sale: Case Study 2

The second case study is about the insurance process of selling products. In contrast of the previously mentioned case, the complexity of the selling product comes from not the geographical diversity but complexity of the product itself. The objective of our research is to transform the business process knowledge into domain ontology. In this case I use ProMine to ensure the extraction and filtering of most relevant domain related concepts for domain ontology by using our proposed hybrid similarity measure.

5.1 Case Background

In insurance, very basic insurance products are quite simple and they are more or less about the distribution of risk such as accident insurance. But in other cases, the insurance products are more sophisticated and somehow they are link to savings options. So, In this case the insurance is an investment as well, like life insurance combined with investments. Now these are complicated products and to sell these products, the brokers have to know very well the products e.g. what is the saving ratio; what is the investment ratio; what to do with earnings and savings but they also have to be familiar how to understand your customers' requirements, how to customize the product for the specific interest of a specific customer. It is very important to ensure that everybody in the insurance network who are responsible for selling products in office has the same level and highest level knowledge of the products, because if they make mistake then have very serious consequences. The text mining is somehow controlled by a prior knowledge; we have basic idea what we are looking for? The insurance domain is not absolutely unfamiliar. We have some knowledge about insurance business. Therefore, we have selected this domain as a use case.

DOI: 10.14267/phd.2015065

5.2 Case Objective and the Related Research Question

One of main goals of my research is the concept extraction from unstructured data (in form of text) of business processes. The third research question of my research is based on this case that a modified semantic similarity measure will improve significantly the efficiency and quality of a domain ontology enhancement/ enrichment. In first case study, we have examined that by using only statistical measure for filtering most relevant concepts. We didn't get good enough results; so there is need to apply some other methods on our different business domains. For this purpose in this case study for insurance domain I applied statistical as well as lexical taxonomy structure so that the semantic distance between nodes in the semantic space constructed by the taxonomy can be better quantified with the computational evidence derived from information gain of derived concepts. The purpose of the case is to prove that the proposed hybrid measure (Cloud Kernel Estimator) has appropriate concept filtering capability to filter out the most relevant concepts from the candidate list of concepts and provides an effective and efficient (ontology-based e-learning environment) or ontology learning.

5.3 ProMine Context (or Application of ProMine)

5.3.1 Datasets Description

In order to apply the proposed hybrid similarity measure for concept filtering, I have divided our data set into two parts. As I mentioned above, the focus domain is insurance. Therefore, first part of dataset are those XML files which are produced from the process management tool and for second part of dataset I have prepared a domain text corpus.

DOI: 10.14267/phd.2015065

5.3.2 Text corpus description

Insurance domain corpus is mainly consisting of articles, manuals, tutorials in a variety of formats in a given domain. Insurance books are included such as (Bickelhaupt, 1979; Bodla, 2004; Martin C. Pentz, 2009; Vaughan, 1995) and some insurance glossaries (A.M. Best Webinar, 2015; Faulkner, 1960; Green, Osler, & Bickley, 1982; IRMI, 2000; Keim, 1978). Some journal articles related to insurance such as (Brown & Warshawsky, 2013; Dionne, 2009; Giesbert, Steiner, & Bendig, 2011; Smoluk, 2009). Suitable adaptors are then used to bring the information in plain text form such that the documents can be parsed by a parser. After that, text preprocessing techniques are applied to this corpus as detail description is given in Section 3.1.1.

5.3.3 Empirical Evaluation

In the insurance product selling case, from process model, XML file is generated in which some description about different tasks of "insurance product sale" are defined. ProMine take this XML as an input data and its first data extracting module (section 3.1) extracts text from description attribute of different tasks. This description attribute contains some descriptive information about the task. After extraction, data extraction module save this extracted text into different text files according to tasks. After extracting this text, the preliminary phase of ProMine starts in which different preprocessing techniques are applied to this unstructured text as mentioned in section 5.1.2. The output of this phase is a set of unique key words against each task. Now the main processing of concept extraction tool, ProMine starts. First of all, I have selected "insurance" as key word and pass it to concept enrichment module which performs a two-phase process. In first phase to find more information elements related to insurance product

DOI: 10.14267/phd.2015065

development process, set of synonyms from WordNet and as well from Wiktionary are extracted. For example, for the keyword “insurance” a synonym list with the following elements: [policy, insurance policy, indemnity] is produced. In second phase, to make this list richer and domain related, each word of this list including key word is passed through the domain corpus where compound words are compiled according to a procedure that is given in section 5.3. The result of this phase is a huge list of concepts is extracted against “insurance” key word. To filter more relevant terms we applied information gain measure that is described in section 3.4.1. As the results of the case study 1 showed that only information gain method doesn’t produce good results for concept filtering so in this case study I have also applied another semantic method which is based on lexical taxonomy of WordNet. And then I have applied our own proposed hybrid similarity measure for concept filtering. A comparison result of three methods is given in Table 2.1. At the end, we have a filtered list of concepts that is ready to add in the ontology and then from filtered concepts, a taxonomy of insurance domain is also prepared with the help of domain expert that is shown in figure 2.1.

5.3.4 Results and Discussion

Table 2.1 lists the complete results of each similarity rating measure. Table 2.1 shows the number of concepts that were extracted from the keywords and then I have set a threshold for filtering more relevant concept. In this case I set our threshold at 60% of the total obtained results. For example, if ProMine extracted 370 concepts related to key word insurance then applying different filtering methods I have filtered 60% top relevant concepts that are 222 in this case. These filtered concepts are evaluated by domain expert. These filtered concepts are evaluated by two categories; “Accepted” and “Rejected”. The domain expert also examined the

DOI: 10.14267/phd.2015065

low ranked concepts which are rejected during filtering process to check either our filtering method is working properly or not.

Table 5-1 Summary of Experimental Results

	Concepts		Filtered Concepts		Unfiltered
Evaluation	Extracted	Filtered	Accepted	Rejected	Accepted
Information Gain (IG)	370	222	123	99	81
Semantic similarity	370	222	112	110	92
Cloud Kernel Estimator	370	222	183	39	21

From the table we can see that our proposed hybrid measure showed more than 82% accuracy while other methods; IG have 55% accuracy and lexical taxonomy based measure has 50% accuracy. It shows that our proposed hybrid approach outperforms from others.

For further validation of our filtering approach with the help of domain expert these filtered concepts are mapped on seed ontology of insurance and have created a taxonomy. Table 2.2 shows this evaluation. Accepted concepts are those concepts which are recognized by domain expert from filtered list of concepts. This Accepted category is further divided by two categories; “Important” that are fit in to the categories of seed ontology and “Understandable” which are understandable and can be the part of ontology as add more classes or categories in seed ontology. Rejected concepts are considered as invalid or irrelevant concepts. The resulting taxonomy from our extracted concepts is given in figure 2.1.

Table 5-2 Evaluation of Filtered Concepts

Concepts	Accepted			Rejected
	Important (1)	Understandable (2)	(1)+(2)	
222	166	17	183	39

DOI: 10.14267/phd.2015065

The results show that more than 80% of the concepts were accepted by the domain expert. I have selected insurance domain because we have some prior knowledge in the form of seed ontology and we want to enrich this ontology with new concepts. However, due to this prior knowledge, we found that some of the concepts are easily categorized in seed ontology because these fit into the existing classes but for some concepts (understandable), we have faced difficulty to categorize them. These are concepts which belong to the policy attribute. However, on the basis of these "understandable" concepts ontology engineers can add more categories in the ontology. Thus, ProMine ontology extraction tool can provide a great help in ontology population and enrichment as well.

5.4 Case Conclusion

This case presented ProMine as an ontology extraction and filtering tool for ontology learning. Besides concept extraction ProMine also addresses ranking and filtering relevant terms by using a new hybrid similarity measure. The novelty of this extracting tool is that 1) it extracts concepts from a very little knowledge that are embedded in organizational processes and with the help of outsources enrich this knowledge and extracts a huge number of new concepts automatically without human interaction; 2) its filtering approach uses deep syntactic and semantic analysis to filter important concepts. The proposed new hybrid similarity measure can be used for other applications of artificial intelligence, psychology and cognitive science. This case study illustrated that the performance of ProMine was assessed using a human evaluation and the results showed that many new concepts were successfully extracted and later on used for ontology population.

DOI: 10.14267/phd.2015065

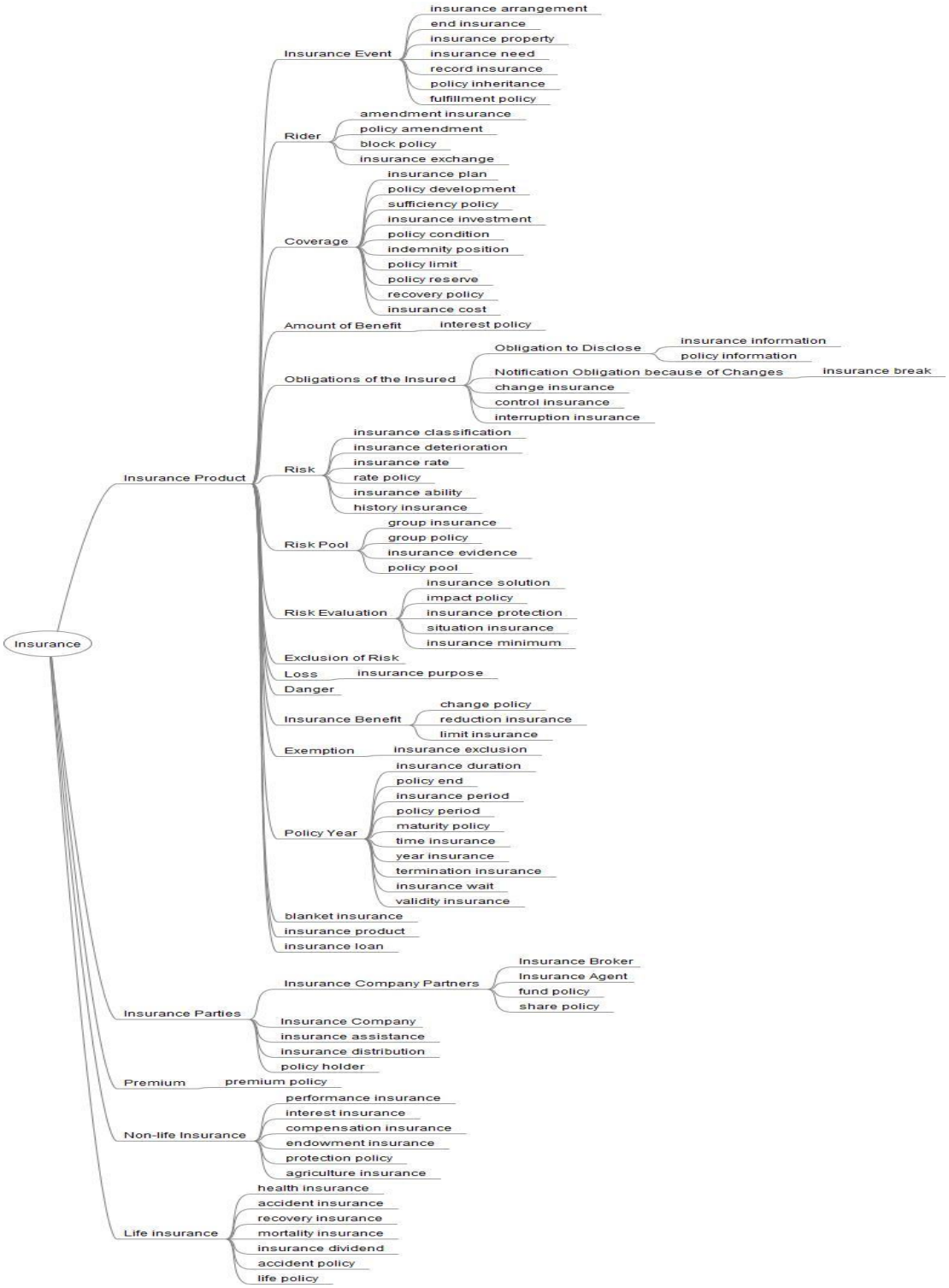

Figure 5-1 Manual categorization of extracted knowledge elements

DOI: 10.14267/phd.2015065

ProMine for IT Audit Ontology Population: Case Study 3

ProMine ontology enrichment capability was tested in an ontology-based e-learning environment applied for IT auditors' training in CISA preparation courses.

6.1 Case Background

As more companies realize the importance of their intellectual assets, especially the employees' competencies, e-learning solutions are gaining in popularity. Therefore, enterprises turn towards the development and/or application of innovative ICT systems to meet their diverse training needs. The most innovative e-learning solutions not only enable employees to easily access, understand and apply even complex training materials, but also can be easily integrated with the already existing knowledge management solutions and systems of the organization.

Corvinno Ltd. developed a complex ontology driven e-learning solution, called Studio system that actively supports the whole learning cycle, independently from its form (e.g. workstation- or mobile phone-based learning). The Studio system is a platform independent e-learning environment that enables the development of customized qualification programs, based on the individual's pervious qualifications, completed levels, corporate trainings and practical experiences. The potential application areas of this approach are wide, both public and business sector can enjoy its benefits. Through its main components – educational ontology, content management system, adaptive testing system, learning management system –, the system is also capable to tackle the challenges of communication, collaboration, content delivery regardless of time and space.

DOI: 10.14267/phd.2015065

Studio system can be applied in formal and non-formal education; in higher education and in company trainings. Studio has an ontology component which covers various fields such as IT audit, business information systems, business intelligence. The focus is IT audit in this case, because Studio is actively applied in CISA (Certified Information System Auditor) exam preparation training in Hungary. CISA training is a combination of face-to-face consultation and e-learning methods. Face to face consultations are dedicated to discuss IT audit domains and the related concepts, theories, required for the exam, while Studio, the e-learning environment has a key role in providing customized learning support for the participants to identify their knowledge bottleneck. Studio system helps to discover the user`s knowledge gaps through the adaptive testing approach and provide access to instant learning material..

6.2 Case Objective and the Related Research Question

Ontology maintenance is a complex and critical issue in general and in our case too. The fourth research question of my research is based on this case that is whether taking top categories from the existing ontology will improve the result of text mining solution to help ontology enrichment process or not? It determines the quality of the Studio system and our educational service as well. The most important challenges, which have to face are the following:

- Business Informatics and IT audit fields are changing fast
- New knowledge areas appear, which relate to IT audit, like data science/big data management
- Students require support during the whole educational cycle (from knowledge level assessment through customized learning material)
- The regulatory environment in higher education is volatile

DOI: 10.14267/phd.2015065

- New competencies are required from the labour market
- MSc curriculums are renewed often.

The purpose of the case is to prove that ProMine has appropriate ontology enrichment capability to provide effective and efficient ontology-based e-learning environment.

6.3 ProMine Context (or Application of ProMine)

6.3.1 Datasets Description

In order to apply the proposed ontology based categorization method for ontology enrichment, we have taken two types of datasets. As we mentioned above, our focus domain is IT audit. Therefore, a domain text corpus had to be prepared as well as the relevant domain seed ontology with some sample concepts.

6.3.1.1 Text corpus description

The text corpus examined is taken from the CISA Review manual 2014 and ISACA® Model Curriculum for IS Audit and Control, 3rd Edition, 2012 that is available from www.isaca.org. ISACA was incorporated in 1969 by a small group of individuals to provide a centralized source of information and guidance in the growing field of auditing controls for computer systems. ISACA provides practical guidance, benchmarks and disseminate best practices in IT audit domain. Some other IT audit related books and articles are also collected from ISACA Journal and web sources. Today's de facto standard for IT auditing is COBIT, therefore COBIT 4.1 has been included in this corpus. TOGAF (Version, 2009)is also a part of this domain corpus that is a framework for developing an enterprise architecture to achieve the technology convergence and application rationalization implied in the audit analysis. Many white papers and published

DOI: 10.14267/phd.2015065

articles about IT audit are also included in the corpus (Brand, 2014; ISACA, 2012, 2015; Rolling Meadows, 2007; Sajay Rai, 2014). Text preprocessing techniques including NLP techniques are applied to this corpus as detailed in Section 3.1.1.

6.3.1.2 Seed Ontology

IT audit seed ontology is structured by the suggestion of ISACA IT audit sample curriculum. ISACA divides the IT audit field for five main areas. The "process of auditing information systems" area deals with risk-based audit planning; audit project management techniques, control objectives and controls related to information systems, ISACA IT audit and assurance standards, guidelines, and tools and techniques. "Governance and management of IT" area has ten subtopics, each that focus on the management of process IT areas such as human resources (HR), IT organizational structure legal issues, and standards and monitoring of assurance practices. "Information systems acquisition, development and implementation" area has six topic areas that focus on business case development, information systems implementation and migration, project management and controls. "Information systems operation, maintenance and support" area have ten subtopics, namely service level management, maintenance of information systems, problem and incident management, change and configuration management, and backup and restoration of systems. "Protection of information assets" area has five subareas; design and implementation of system and security controls, data classification, physical access, and the process of retrieving and disposing of information assets. Based on ISACA suggested categorization five main knowledge areas/category are distinguished in the seed ontology as shown in Figure 5-1.

DOI: 10.14267/phd.2015065

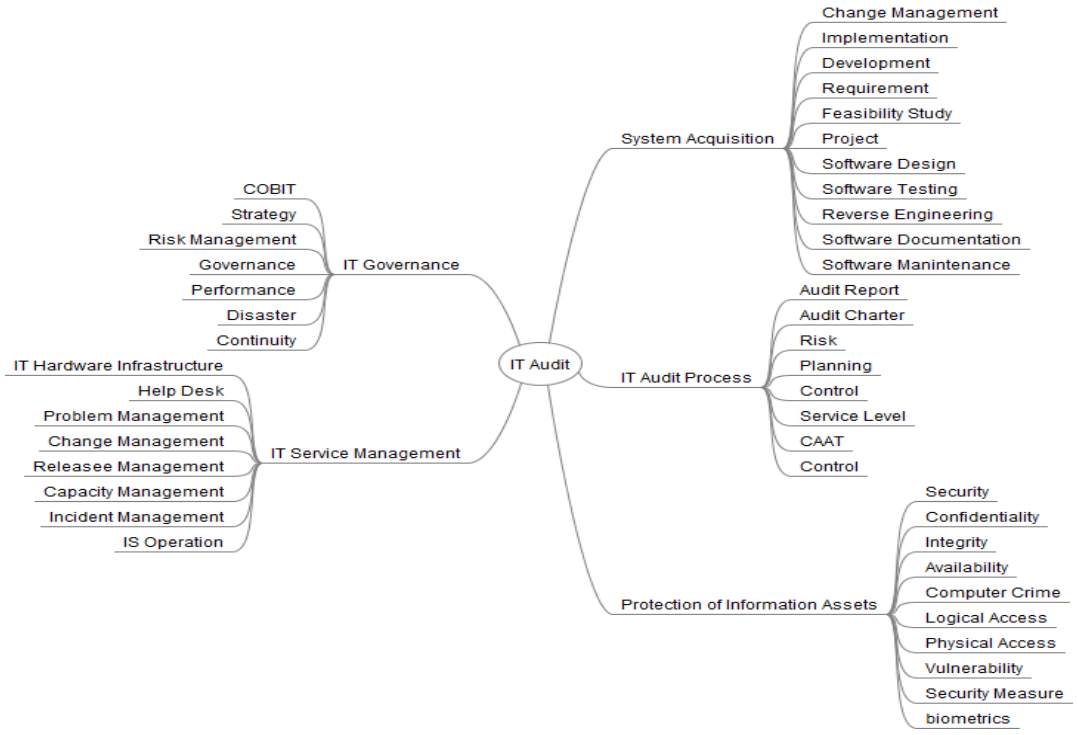

Figure 6-1 IT Audit Seed Ontology used in our categorization experiments

Our ontology enrichment process focuses on protection of information assets and IT governance categories. These two categories are emphasized, because their knowledge content becomes outdated fast protection of information assets has decisive role in CISA certification job practice areas (30%) and IT governance has special importance nowadays in companies' life.

6.3.2 Empirical Evaluation

In order to test the ProMine system for ontology enrichment, I described the empirical evaluation procedure in an incremental iterative process with four essential steps as illustrated in Figure5-2. The focus of the evaluation was to define an experiment to enrich the domain seed ontology by using extracted concepts from ProMine.

Aforementioned domain corpus is used to extract lists of concepts with the help of WordNet and Wiktionary. After extraction of these concepts, concepts relatedness with domain is also

DOI: 10.14267/phd.2015065

checked by statistical and semantic measures. I found a number of domain related concepts at the end of ProMine Concept extraction step. Point-by-point description of the ProMine concept extraction procedure is given in chapter3.

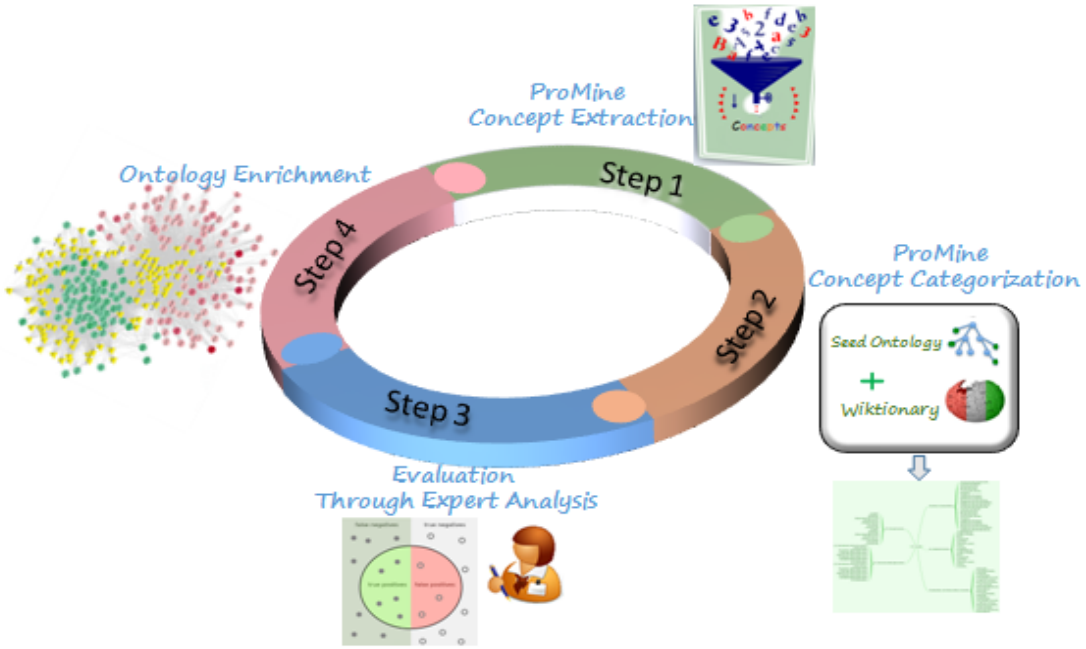

Figure 6-2Empirical Evaluation Process

Once a set of key-concepts has been extracted from a domain corpus, the next empirical step is the categorization of these concepts according to ontology categories. For this purpose, I have taken a seed ontology of IT Audit that have five main categories and each category contains few concepts. To enrich these categories with new domain concepts algorithm is used as in Figure 3-4. According to this algo a set of related concepts against each category's concept is made and then each candidate concept from concept list and is matched against the concept sets of all ontology categories. If the concept is matched with any word of these concept sets, then the system puts that candidate concept to into that specified ontology category. The matching is performed by applying normalized matching techniques to concepts. If the candidate concept

DOI: 10.14267/phd.2015065

is not matched with any word of categories concept sets, then ProMine puts this concept to a new category that is labeled as miscellaneous. One by one, all candidate concepts of a list are checked through this procedure and when a list of concepts is completed, then all words of this list have must in any category.

The third empirical phase is related to evaluating the precision of the ontology enrichment process. To do this, at the end of each concept list categorization, the experimental results are evaluated by the domain expert.

A domain expert analyzed concepts of each category and point out errors. In this phase expert also assigns categories to concepts of miscellaneous category. Here interesting point of ProMine system is that if the expert identifies such concepts on the basis which he/she wants to define a new category in the ontology and in next cycle ProMine will also consider this new category as well. For example, when I used "architecture" as an input concept in ProMine and during concept categorization, in miscellaneous category I got these concepts (amongst others) as result: architecture design, architecture specification, baseline architecture, architecture decomposition, architecture. The results of this step are shown in Table 5-2. I used precision and recall measures to demonstrate the effectiveness of our proposed semantic concept categorization technique. After this phase, numerous new concepts are ready for ontology enrichment. It is important to mention that after completing first cycle, ProMine starts step one again with new concept list and repeat whole empirical procedure until all extracted concept lists are categorized. As shown in Table 5-2, I was able to obtain promising precision results in general for most of the categories. Since ProMine system showed good precision in enriching the ontology at the end of the first cycle, but results showed that every next cycle gave more

DOI: 10.14267/phd.2015065

precisions in the categorization process. The reason is as categories gradually have more concepts, chance to match candidate concept with any category is increased. Therefore, in after coming cycles, in the miscellaneous category contained less false negatives (that should be categorized by the system) as compared to earlier cycles. This difference is shown in Figure 5-3, 5-4 and 5-5.

6.3.3 Results and Discussion

Results are shown in Figure 5-3, 5-4 and 5-5. To find out these results I used following contingency Table 5-1. In this table; TP are those concepts which are correctly classified and ProMine system has put them in their corresponding categories, FP are those concepts that are wrongly categorized, FN are those concepts that are not selected by our categorization system for any category and therefore put them in miscellaneous category, while those should put in some category and TN are those concepts that are extracted from the ProMine extraction module but system could not categorize them for any category and put them in miscellaneous category (a new category that system generates itself in each cycle). This miscellaneous category will be analysed by domain expert who will manually put FN into the corresponding category. I used this table to find precision and recall indicators to measure the performance of our system.

Table 6-1 Contingency table

	Categorized	Non-categorized
Categorized	True Positives (TP)	False Positives (FP)
Non-categorized	False Negatives (FN)	True Negatives (TN)

DOI: 10.14267/phd.2015065

In Figure 5-3, TP, FP and FN trends illustrated how the ProMine categorization system improves its functionality during the course of 10 cycles of automated iterative analysis. TP curve is illustrating that as no. of cycles increases, the true positive values (new concepts in their relevant category) increase gradually. It is because at the start of the process, each category has few concepts and when system matches new concepts with these seed concepts according to semantic concept categorization algorithm, only few concepts can matched by the system so we can see in Figure 5-3 that in first cycle the number of false negatives is high.

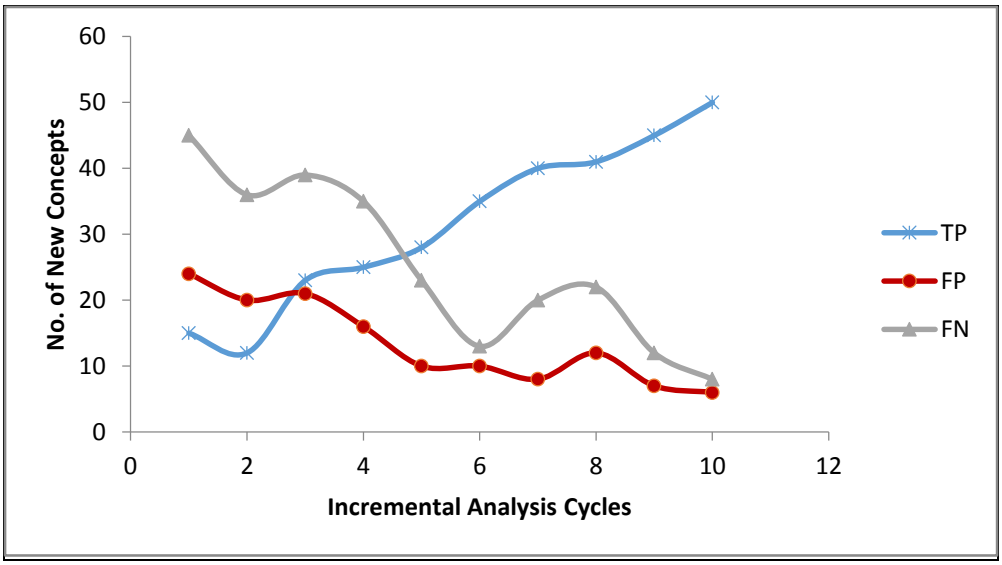

Figure 6-3Evaluation Indicators of Incremental Analysis

At the end of each cycle, domain master selects FN from the miscellaneous category and these concepts are added to their corresponding category and therefore, in the next cycle system has more concepts for matching. In Figure 6-3, the second line shows false positive (FP) values that refers to concepts that are assigned in incorrect categories but these are not high values. It shows our system perform correctly and puts a new concept in its right category and otherwise puts in the miscellaneous category. As the number of cycles increases false positives, further decrease as we can see at tenth cycle, less than ten false positives are showing. The third line in

DOI: 10.14267/phd.2015065

Figure 6-3 illustrating false negative (FN) trend of the ProMine categorization system. False negatives are representing those concepts which should be categorized in their related category, but system couldn't do this and put them in the miscellaneous category. To show the good performance of the system, false negative should be decreased because when false negatives are high then this shows we need manual analysis. The positive point of our system is to gradually minimize false negatives as cycles are increasing. This shows that our system gradually need less effort of domain expert. All these results are shown in table 6-2.

Table 6-2 Contingency table for Concept Categorization

Concepts	Categorized		Non Categorized	
	True Positives(TP) Truly Categorized	False Positives (FP) Incorrectly Categorized	False Negatives (FN) Missed Concepts	True Negatives (TN) Correctly non-categorized
1	15	24	45	113
2	12	20	36	242
3	23	21	39	150
4	25	16	35	295
5	28	10	23	64
6	35	10	13	226
7	40	8	20	297
8	41	12	22	146
9	45	7	12	96
10	50	6	8	76

However, in second cycle, we observe a change in behavior of TP curve that is showing decrease which seems against our expectations, but detailed view analysis shows that this decrease is not the malfunctioning of our system rather the less number of concepts are extracted from the ProMine concept extraction module and therefore a clear increase in TP values is not appearing. To justify this reason, Figure 6-4 represents true-positive rate that is:

$$\text{TP rate} = {}^{\text{TP}}\!\!\diagup\!\!{}_{\text{TP}+\text{FN}}$$

DOI: 10.14267/phd.2015065

The true-positive rate, which is also known as sensitivity or recall measures the proportion of actual positives which are correctly identified. If we see in Figure 6-4, the trend of the line will show that the proportion of positives has not decreased in cycle 2 rather it is gradually increasing and this high recall relates to a low false negative rate that indicates our proposed technique is returning a majority of all positive results (high recall). However, this will happen gradually by increasing number of cycles as categories will be enriched with new concepts. Recall is a measure to determine that how many truly relevant results are returned.

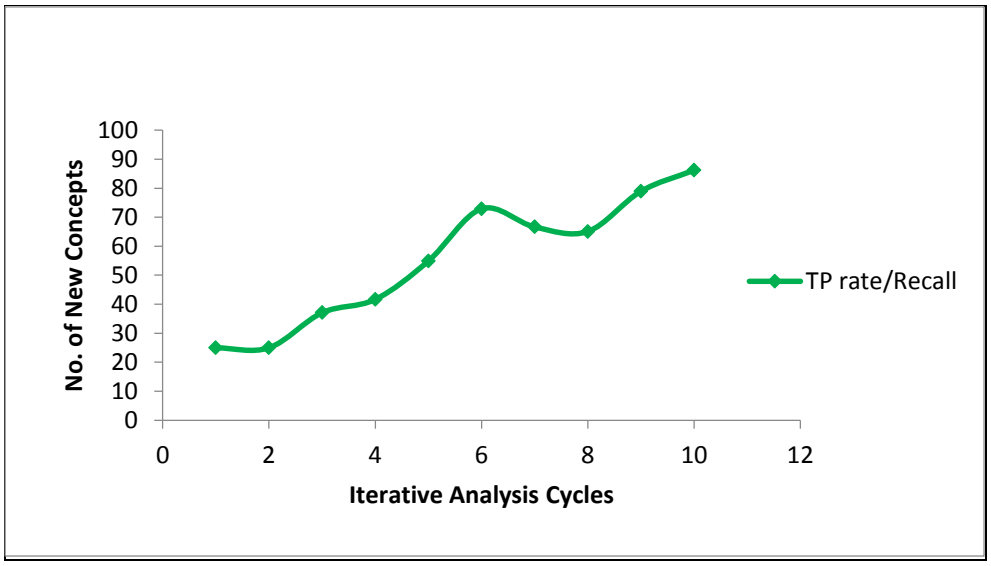

Figure 6-4 TP rate/Recall of semantic concept categorization

Another deviation can be seen in the trend of FN during seventh and eighth cycle where this line goes up and then in the ninth cycle it is showing normal behavior. In every new cycle ProMine selects a new key word and extracts a list of new concepts (there is not limit of this list) relevant to that key term as detailed are mentioned in chapter 3. If we see our these concept lists, we have come to know that during these cycles the key term are more general and therefore extracted more general concepts from corpus and most of which cannot be fit in any category. Domain expert analysis confirms this reason; because during these cycles a large number of

DOI: 10.14267/phd.2015065

concepts have been generated and consequently, the number of false negatives is high as well irrelevant concepts (true negatives) are also great in numbers.

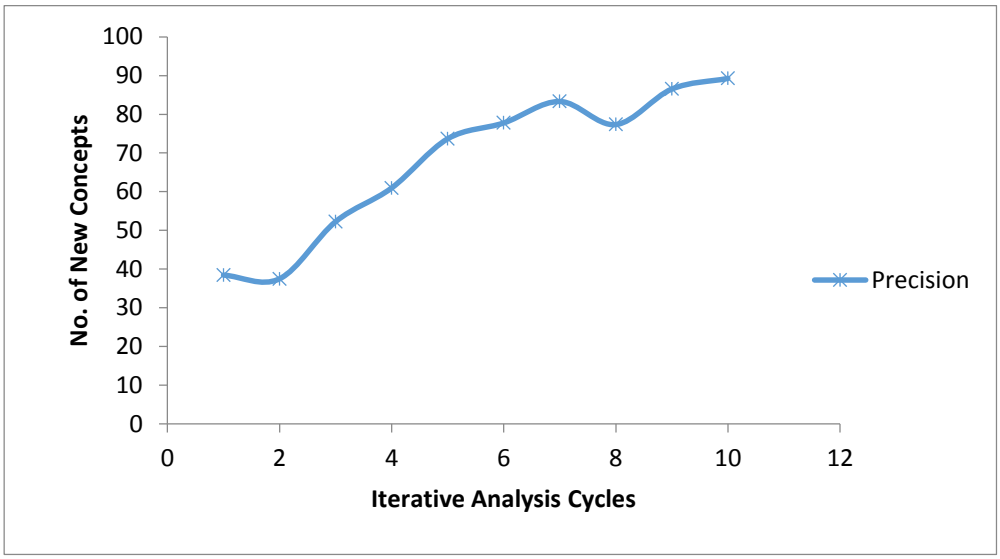

Figure 6-5 Precision of semantic concept categorization

We have also used the precision indicator in order to measure the quality of our results as shown in Figure 6-5. Precision is the fraction of correctly matched concepts (true positives) over the total number of extracted concepts (true positives and false positives). In this context, correctly enriched entities are the concepts and instances that are in the correct category. I used precision to measure the correctness of my proposed technique.

$$Precision = TP/(TP + FP)$$

Figure 6-5 shows the precision of the proposed categorization method. Precision is a measure of result relevancy where high precision relates to a low false positive rate.

Cycle	New Concepts	Retrieval Measure

DOI: 10.14267/phd.2015065

	Total Observation	Manual Identification	Automatic Identification	Precision	Recall
1	80	45	15	38	25
2	128	36	12	37	25
3	190	39	23	52	37
4	250	35	25	61	41
5	301	23	28	74	54
6	349	13	35	78	72
7	409	20	40	83	66
8	472	22	41	77	65
9	529	12	45	87	79
10	587	8	50	89	86

Table 6-3 Evaluation of the overall semantic concept categorization method

6.4 Case Conclusion

An ideal system with high precision and high recall will return many results, with all results labelled correctly. As the number of cycles passed, ProMine system showed good precision in enriching the ontology, a perspective of the work is to extend the number of cycles. If we see precision and recall both then we will come to know that both are increasing gradually. High scores for both showed that the proposed categorization method is returning accurate results (high precision), as well as returning a majority of all positive results (high recall). The experimental evaluation using precision and recall measures as well human judgments showed that ProMine method categorized new concepts with high precision and recall. Empirical results show that ProMine approach can support domain expert ontology engineers in building domain specific ontologies efficiently.

DOI: 10.14267/phd.2015065

Conclusions and Future Work

In this dissertation, I presented a new paradigm of text mining with respect to business processes/organizations. Through this text mining solution, I have to identify and mine the hidden knowledge elements from business processes and put them into a context (domain ontology) to make it knowledge.

The novelty of this extracting tool is that it extracts concepts from a very little knowledge that are embedded in organizational processes and with the help of outsources (WordNet, Wiktionary and domain corpus) enriches this knowledge and extracts a huge number of new concepts automatically without human interaction. The filtering approach of ProMine uses deep syntactic and semantic analysis methods to filter important concepts from the bulk of extracted concepts. For filtering process, a new hybrid similarity measure has been proposed that is the combination of statistic and semantic measures. This method is being implemented in filtering module of ProMine. A novel aspect of this research is the discovery of a semantic process of concept categorization of the extracted concepts. This procedure helps in ontology enrichment and population.

From three case studies, one by one, I strived to answer all research questions. Results of first case study showed that ProMine concept extraction approach can support domain experts and ontology engineers in building domain specific ontologies efficiently. From this case study I have realized that besides statistic measure some semantic measures have to be considered to improve concept filtering process. Therefore, I proposed a new hybrid similarity measure (Cloud Kernel Estimator) for concept filtering process that is the combination of statistic and semantic methods. In the second case study, filtering method of ProMine has been evaluated.

DOI: 10.14267/phd.2015065

The proposed hybrid similarity measure gave good results so this similarity measure might be used for other applications of artificial intelligence, psychology and cognitive science. Many new concepts were successfully extracted and later on used for ontology population. I have conducted evaluations of the concept categorization method. This method categories the extracted knowledge elements by using Wiktionary. The proposed system is evaluated in a cyclic way with the help of domain expert analysis. From this evaluation, it is concluded that ProMine generates a large number of concepts and at the same time categorize these concepts according to given ontology categories or classes. This categorization is good quality and with the passage of time its precision and recall increases. The experiments proved the applicability of automatic concept extraction, filtering and automatic concepts categorization.

However, following points can be considered to improve the proposed methodology which are part of the future work:

The ProMine solution has more potential for the future development. Comparison analyses of different concept extracting and filtering methods have not been considered during the current research work, but this could also help to improve the methodological development of the proposed framework, as a part of future research work. The current research has focused on concept extraction methods whereas the relation (between concept) extraction could be used to improve the ontology enrichment as a part of future research work.

DOI: 10.14267/phd.2015065

Summary

Organizations are struggling with the challenges coming from the regulatory, social and economic environment which are complex and changing continuously. They cause increase demand for the management of organizational knowledge, like how to provide employees, the necessary job-specific knowledge in right time and in right format. Employees have to update their knowledge, improve their competencies continuously. Knowledge repositories have key roles from knowledge management aspects, because they contain primarily the organizations' intellectual assets (it is explicit knowledge) while employees have tacit knowledge, which is difficult to extract and codify. Business processes are also important from the management of organizational knowledge aspects, they have explicit and tacit knowledge elements as well. One of the key questions is how to handle this hidden knowledge in order to improve the organizational knowledge especially employees' knowledge by providing the most appropriate learning and/or training materials and how can we ensure that the knowledge in business processes are the same as in knowledge repositories and employees' head. These are the major themes in this thesis.

The thesis investigates three main research questions: (1) how can we use text mining to extract knowledge from processes in order to enhance or populate the existing ontology; and (2) what methodology and concept extraction methods are available to enhance the existing knowledge that captured from the business process? Whether a text mining based solution can be used for ontology learning in the context of machine learning? (3) A modified semantic similarity measure will improve significantly the efficiency and quality of a domain ontology enhancement/ enrichment; and (4) whether taking top categories from the existing ontology will improve the result of text mining solution to help ontology enrichment process or not? The aim

DOI: 10.14267/phd.2015065

of this research is to examine how knowledge can be extracted more accurately and automatically from the business processes and how automatically categorized this extracted knowledge for domain ontology enrichment and population.

To answer for the research questions above, I have developed a research methodology that consists of existing processes and findings from design research. This methodology combines qualitative and quantitative research methods and detailed in Research Methodology chapter.

In this thesis, I have proposed a text mining framework, ProMine for process-based knowledge extraction, filtering and categorization of this knowledge according to seed ontology. ProMine is used as a text mining component in Prokex project (EUREKA_HU_12-1-2012-0039). ProMine supports to extract, organize and preserve knowledge embedded in organizational processes to enrich the organizational knowledge base in a systematic and controlled way, support employees to easily acquire their job role specific knowledge. This framework extracts knowledge elements from business processes with the help of domain corpus and lexical databases. It provides an automated system doing text mining to extract domain related new concepts and categories these concepts to enrich and populate a domain ontology. For gathering domain related concepts among text data, ProMine uses a semantic similar measure that is the combination of statistical and semantic approaches. In addition to concept extraction, a new procedure is developed for categorization of the extracted concepts. This procedure helps in ontology enrichment and population.

The novelty of the ProMine solution is based on the analysis of business processes' tasks with the help of text mining techniques to extract knowledge from them and in order to connect these business processes to organizational knowledge base, where the process structure will be used

DOI: 10.14267/phd.2015065

for building up the knowledge structure. In this approach knowledge base is the ontology, which provides the conceptualization of a certain domain. The primary innovation lies in new algorithms for the extraction and enrichment of new domain concepts and integration of the static and dynamic process knowledge.

Finally, for the evaluation of ProMine, three case studies are conducted. A case study is performed of the concept extraction procedure by selecting one particular domain, food safety. The results of this case study can help to automatically extract new concepts to support domain ontology learning process. On the basis of results of first case study, some changes have made in concept extraction method and now involving further resources to enrich new concepts and also working on semantic similarity measure to rank most relevant domain concepts. Another case study is conducted for the evaluation of categorization method of ProMine. This method categories the extracted knowledge elements by using Wiktionary. The proposed system is evaluated in a cyclic way with the help of domain expert analysis. From this evaluation, it is concluded that ProMine generates a large number of concepts and at the same time categorize these concepts according to given ontology categories or classes. This categorization is good quality and with the passage of time its precision and recall increases. The experiments proved the applicability of automatic concept extraction and automatic concepts categorization. I intend to conduct additional testing of our proposed system, especially test the relevancy of extracted concepts to the domain because more relevant concepts would be more accurately categorized. These case studies can be considered a preliminary investigation of ProMine framework.

The thesis has four main contributions. Based on the analysis of state-of-the-art text mining and NLP techniques are used in information extraction solution, I suggested and developed a generic

DOI: 10.14267/phd.2015065

text mining solution/framework that build bridges between two different approaches; process modeling that is procedural in nature and context/ontology is declarative in nature. The second contribution is concept extraction and enrichment with the help our resources such as WordNet, Wiktionary and domain corpus. For concept filtration, I proposed a new semantic similarity measure which is a combination of statistical and semantic measures. The third important contribution is to design a concept categorization method for ontology population. Besides these primary contributions, an algorithm is also designed for compound word extraction from text to make concept learning more effective.

Finally, this thesis concludes the work in the epilogue. The involvement of text mining provides a new perspective to handle business processes' semi-structured data in a common framework. This text mining solution together with NLP techniques is expected to be a promising tool in ontology learning for different organizations in the future.

DOI: 10.14267/phd.2015065

References

A.M. Best Webinar. (2015). Glossary of Insurance Terms. Retrieved from http://www.ambest.com/resource/glossary.html

Abdel-moneim, W. T., Abdel-Aziz, M. H., & Hassan, M. M. (2013). Clinical Relationships Extraction Techniques from Patient Narratives. *arXiv preprint arXiv:1306.5170*.

Aduriz, I., Alegria, I., Arriola, J. M., Artola, X., Ezeiza, N., Gojenola, K., & Maritxalar, M. (1995). Different issues in the design of a lemmatizer/tagger for Basque. *arXiv preprint cmp-lg/9503020*.

Ahonen-Myka, H. (1999). *Finding all maximal frequent sequences in text.* Paper presented at the Proc. of the ICML99 Workshop on Machine Learning in Text Data Analysis.

Ahonen-Myka, H. (2002). Discovery of frequent word sequences in text *Pattern Detection and Discovery* (pp. 180-189): Springer.

Ahonen-Myka, H., Heinonen, O., Klemettinen, M., & Verkamo, A. I. (1999). *Finding co-occurring text phrases by combining sequence and frequent set discovery.* Paper presented at the Proceedings of 16th International Joint Conference on Artificial Intelligence IJCAI-99 Workshop on Text Mining: Foundations, Techniques and Applications.

Ahonen, H. (1999). Knowledge Discovery in Documents by Extracting Frequent Word Sequences. *Library trends, 48*(1), 160-181.

Al-Shammari, E., & Lin, J. (2008). *A novel Arabic lemmatization algorithm.* Paper presented at the Proceedings of the second workshop on Analytics for noisy unstructured text data.

Alemanno, A. (2006). Food safety and the single European market. *What's the Beef*, 237-258.

Alfred, R., Mujat, A., & Obit, J. H. (2013). A ruled-based part of speech (RPOS) tagger for malay text articles *Intelligent Information and Database Systems* (pp. 50-59): Springer.

Ananiadou, S., Pyysalo, S., Tsujii, J. i., & Kell, D. B. (2010). Event extraction for systems biology by text mining the literature. *Trends in biotechnology, 28*(7), 381-390.

Antony, P., Mohan, S. P., & Soman, K. (2010). *SVM Based Part of Speech Tagger for Malayalam.* Paper presented at the Recent Trends in Information, Telecommunication and Computing (ITC), 2010 International Conference on.

Attia, M. A. (2007). *Arabic tokenization system.* Paper presented at the Proceedings of the 2007 workshop on computational approaches to semitic languages: Common issues and resources.

Auer, S. (2005). *Powl–a web based platform for collaborative semantic web development.* Paper presented at the Proceedings of the Workshop Scripting for the Semantic Web.

Balbach, E. D. (1999). Using case studies to do program evaluation. *Sacramento, CA: California Department of Health Services*.

Barforush, A. A., & Rahnama, A. (2012). Ontology learning: revisted. *Journal of Web Engineering, 11*(4), 269-289.

Bekkerman, R., El-Yaniv, R., Tishby, N., & Winter, Y. (2001). *On feature distributional clustering for text categorization.* Paper presented at the Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval.

DOI: 10.14267/phd.2015065

Bembenik, R., Skonieczny, L., Rybinski, H., Kryszkiewicz, M., & Niezgodka, M. (2013). *Intelligent Tools for Building a Scientific Information Platform: Advanced Architectures and Solutions* (Vol. 467): Springer.

Bickelhaupt, D. L. (1979). *General insurance*: McGraw-Hill/Irwin.

Blanchard, A. (2007). Understanding and customizing stopword lists for enhanced patent mapping. *World Patent Information, 29*(4), 308-316.

Bodla, B. S. (2004). *Insurance: Fundamentals, Environment and Procedures*: Deep and Deep Publications.

Brand, D. (2014). A Global Look at IT Audit Best Practices. *4th Annual IT Audit Benchmarking Survey*. Retrieved from 4th Annual IT Audit Benchmarking Survey website: http://www.isaca.org/Knowledge-Center/Research/Documents/A-Global-Look-at-IT-Audit-Best-Practices_whp_Eng_1114.pdf?regnum=260570 Retrieved from http://www.isaca.org/Knowledge-Center/Research/Documents/A-Global-Look-at-IT-Audit-Best-Practices_whp_Eng_1114.pdf?regnum=260570

Brewster, C., Ciravegna, F., & Wilks, Y. (2002). User-centred ontology learning for knowledge management *Natural Language Processing and Information Systems* (pp. 203-207): Springer.

Brown, J., & Warshawsky, M. (2013). The Life Care Annuity: A New Empirical Examination of an Insurance Innovation That Addresses Problems in the Markets for Life Annuities and Long-Term Care Insurance. *Journal of Risk and Insurance, 80*(3), 677-704.

Buitelaar, P., Olejnik, D., & Sintek, M. (2003). *OntoLT: A protege plug-in for ontology extractionfrom text.* Paper presented at the Proceedings of the International Semantic Web Conference (ISWC).

Buitelaar, P., & Sacaleanu, B. (2001). *Ranking and selecting synsets by domain relevance.* Paper presented at the Proceedings of WordNet and Other Lexical Resources: Applications, Extensions and Customizations, NAACL 2001 Workshop.

Bunescu, R., Ge, R., Kate, R. J., Marcotte, E. M., Mooney, R. J., Ramani, A. K., & Wong, Y. W. (2005). Comparative experiments on learning information extractors for proteins and their interactions. *Artificial intelligence in medicine, 33*(2), 139-155.

Burgess, C., Livesay, K., & Lund, K. (1998). Explorations in context space: Words, sentences, discourse. *Discourse processes, 25*(2-3), 211-257.

Chang, C. H., Kayed, M., Girgis, M. R., & Shaalan, K. F. (2006). A survey of web information extraction systems. *Knowledge and Data Engineering, IEEE Transactions on, 18*(10), 1411-1428.

Cheli, F., Battaglia, D., Gallo, R., & Dell'Orto, V. (2014). EU legislation on cereal safety: An update with a focus on mycotoxins. *Food Control, 37*, 315-325.

Chifu, E. T., & Le Ia, I. A. (2008). Text-based ontology enrichment using hierarchical self-organizing maps.

Cimiano, P. (2005). *Ontology Learning and Population: Algorithms, Evaluation and Applications.* PhD thesis, University of Karlsruhe, 2005. forthcoming.

Cimiano, P., & Völker, J. (2005). Text2Onto *Natural language processing and information systems* (pp. 227-238): Springer.

Commission of the European Communities. (1999). White Paper on Food Safety *Brussels: Commission of the European Communities*, 1-37.

Corvinno Center, T. T. (2011). STUDIO - Ontology Driven Learning Environment. Retrieved from http://corvinno.hu/web.nsf/do?open&lang=en&page=proj-studio#

DOI: 10.14267/phd.2015065

Cunningham, H. (2005). Indexing and querying linguistic metadata and document content. *Recent Advances in Natural Language Processing IV: Selected papers from RANLP, 292*, 35.

Cunningham, H., Maynard, D., Bontcheva, K., & Tablan, V. *GATE: A framework and graphical development environment for robust NLP tools and applications, 2002.* Paper presented at the Proc. 40th Anniversary Meeting of the Association for Computational Linguistics (ACL).

Cunningham, H., Maynard, D., Bontcheva, K., & Tablan, V. (2002). *GATE: A framework and graphical development environment for robust NLP tools and applications.* Paper presented at the Proc. 40th Anniversary Meeting of the Association for Computational Linguistics (ACL).

Dagan, I., Pereira, F., & Lee, L. (1994). *Similarity-based estimation of word cooccurrence probabilities.* Paper presented at the Proceedings of the 32nd annual meeting on Association for Computational Linguistics.

Daviter, F. (2009). Schattschneider in Brussels: how policy conflict reshaped the biotechnology agenda in the European Union. *West European Politics, 32*(6), 1118-1139.

Dawson, J. (1974). Suffix removal and word conflation. *ALLC bulletin, 2*(3), 33-46.

Delgado, M., Martín-Bautista, M. J., Sánchez, D., & Vila, M. (2002). Mining text data: special features and patterns *Pattern Detection and Discovery* (pp. 140-153): Springer.

Dennis, S., Landauer, T., Kintsch, W., & Quesada, J. (2003). *Introduction to latent semantic analysis.* Paper presented at the Slides from the tutorial given at the 25th Annual Meeting of the Cognitive Science Society, Boston.

Dionne, G. (2009). Introduction to the Special Issue on Long-Term Care Insurance and Health Insurance. *Journal of Risk and Insurance, 76*(1), 1-4.

Directive, B. (1979). Council Directive 79/409/EEC of 2 April 1979 on the conservation of wild birds. *Official Journal L, 103*(25/04), 0001-0018.

Dreyer, M., & Renn, O. (2009). *Food safety governance*: Springer.

Ercan, G., & Cicekli, I. (2007). Using lexical chains for keyword extraction. *Information Processing & Management, 43*(6), 1705-1714.

Etzioni, O., Cafarella, M., Downey, D., Popescu, A.-M., Shaked, T., Soderland, S., . . . Yates, A. (2005). Unsupervised named-entity extraction from the web: An experimental study. *Artificial intelligence, 165*(1), 91-134.

EUREKA. (2013). ProKEx -Integrated Platform for Process-based Knowledge Extraction (EUREKA project). Hungary: Research and Technology Innovation Fund, New Szécsényi Plan, Hungary.

EUROPA. (2014). Access to European Union law. *EUR-Lex.* Retrieved from http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32012L0013

Euzenat, J., & Shvaiko, P. (2007). *Ontology matching* (Vol. 333): Springer.

Farquhar, A., Fikes, R., & Rice, J. (1997). The ontolingua server: A tool for collaborative ontology construction. *International journal of human-computer studies, 46*(6), 707-727.

Faulkner, E. J. (1960). *Health insurance*: McGraw-Hill.

Feldman, R., Dagan, I., & Hirsh, H. (1998). Mining text using keyword distributions. *Journal of Intelligent Information Systems, 10*(3), 281-300.

DOI: 10.14267/phd.2015065

Feldman, R., & Hirsh, H. (1996). *Mining Associations in Text in the Presence of Background Knowledge.* Paper presented at the KDD.

Ferraro, J. P., Daumé, H., DuVall, S. L., Chapman, W. W., Harkema, H., & Haug, P. J. (2013). Improving performance of natural language processing part-of-speech tagging on clinical narratives through domain adaptation. *Journal of the American Medical Informatics Association, 20*(5), 931-939.

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse processes, 25*(2-3), 285-307.

Formica, A. (2008). Concept similarity in formal concept analysis: An information content approach. *Knowledge-Based Systems, 21*(1), 80-87.

Fortuna, B., Grobelnik, M., & Mladenic, D. (2007). *OntoGen: semi-automatic ontology editor*: Springer.

Friedman, C., Kra, P., Yu, H., Krauthammer, M., & Rzhetsky, A. (2001). GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics, 17*(suppl 1), S74-S82.

Gábor, A., & Szabó, Z. (2013). Semantic Technologies in Business Process Management *Integration of practice-oriented knowledge technology: trends and prospectives* (pp. 17-28): Springer.

Gacitua, R., Sawyer, P., & Rayson, P. (2008). A flexible framework to experiment with ontology learning techniques. *Knowledge-Based Systems, 21*(3), 192-199.

Gal, A., Modica, G., & Jamil, H. (2004). *Ontobuilder: Fully automatic extraction and consolidation of ontologies from web sources.* Paper presented at the Data Engineering, 2004. Proceedings. 20th International Conference on.

Gelfand, B., Wulfekuler, M., & Punch, W. (1998). *Automated concept extraction from plain text.* Paper presented at the AAAI 1998 Workshop on Text Categorization.

George, P., Vangelis, K., Anastasia, K., Georgios, P., & Constantine, S. D. Semi-automated ontology learning: the BOEMIE approach.

Ghadfi, S., Béchet, N., & Berio, G. (2014). *Building ontologies from textual resources: A pattern based improvement using deep linguistic information*. Paper presented at the Proceedings of the 5th Workshop on Ontology and Semantic Web Patterns (WOP2014), Riva del Garda, Italy.

Giesbert, L., Steiner, S., & Bendig, M. (2011). Participation in micro life insurance and the use of other financial services in Ghana. *Journal of Risk and Insurance, 78*(1), 7-35.

Gillani, S. A., & Kő, A. (2014). Process-based knowledge extraction in a public authority: A text mining approach *Electronic Government and the Information Systems Perspective* (pp. 91-103): Springer.

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., . . . Smith, N. A. (2011). *Part-of-speech tagging for twitter: Annotation, features, and experiments.* Paper presented at the Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2.

Goldsmith, J., & Messarovitch, Y. (1994). *The trap*: Macmillan London.

Gómez-Pérez, A., & Manzano-Macho, D. (2003). A survey of ontology learning methods and techniques. *OntoWeb Deliverable D, 1*, 5.

Green, T. E., Osler, R. W., & Bickley, J. S. (1982). *Glossary of Insurance Terms*: Merritt.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge acquisition, 5*(2), 199-220.

DOI: 10.14267/phd.2015065

Guo, W., & Diab, M. (2012). *A simple unsupervised latent semantics based approach for sentence similarity.* Paper presented at the Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation.

Hajic, J., Panevová, J., Hajicová, E., Sgall, P., Pajas, P., Štepánek, J., . . . Razımová, M. Š. (2006). Prague dependency treebank 2.0. *CD-ROM, Linguistic Data Consortium, LDC Catalog No.: LDC2006T01, Philadelphia, 98.*

Hasan, F. M., UzZaman, N., & Khan, M. (2007). Comparison of different POS Tagging Techniques (N-Gram, HMM and Brill's tagger) for Bangla *Advances and Innovations in Systems, Computing Sciences and Software Engineering* (pp. 121-126): Springer.

Hassanpour, S., O'Connor, M. J., & Das, A. K. (2013). A semantic-based method for extracting concept definitions from scientific publications: evaluation in the autism phenotype domain. *J. Biomedical Semantics, 4*, 14.

Hassler, M., & Fliedl, G. (2006). Text preparation through extended tokenization. *Data Mining VII: Data, Text and Web Mining and their Business Applications, 37*, 13-21.

Hepp, M., Leymann, F., Domingue, J., Wahler, A., & Fensel, D. (2005). *Semantic business process management: A vision towards using semantic web services for business process management.* Paper presented at the e-Business Engineering, 2005. ICEBE 2005. IEEE International Conference on.

Hobbs, J. R., Appelt, D. E., Bear, J., & Tyson, M. (1992). *Robust processing of real-world natural-language texts.* Paper presented at the Proceedings of the third conference on Applied natural language processing.

Hotho, A., Nürnberger, A., & Paaß, G. (2005). *A Brief Survey of Text Mining.* Paper presented at the Ldv Forum.

Huang, C.-R., Šimon, P., Hsieh, S.-K., & Prévot, L. (2007). *Rethinking Chinese word segmentation: tokenization, character classification, or wordbreak identification.* Paper presented at the Proceedings of the 45th annual meeting of the acl on interactive poster and demonstration sessions.

IRMI. (2000). Insurance and Risk Management Terms. Retrieved from http://www.irmi.com/online/insurance-glossary/terms.aspx

ISACA, C. (2012). 5: A Business Framework for the Governance and Management of Enterprise IT. *Rolling Meadows: ISACA.*

ISACA, C. (2015). *Getting Started With Governance of Enterprise IT (GEIT)* (0736-6981). Retrieved from USA http://www.isaca.org/Knowledge-Center/Research/Documents/Getting-Started-With-GEIT_whp_Eng_0314.pdf?regnum=260572

Islam, N., Siddiqui, M. S., & Shaikh, Z. (2010). *TODE: A dot net based tool for ontology development and editing.* Paper presented at the Computer Engineering and Technology (ICCET), 2010 2nd International Conference on.

ITGI. (2007). COBIT 4.1.

Janik, M., & Kochut, K. J. (2008). *Wikipedia in action: Ontological knowledge in text categorization.* Paper presented at the Semantic Computing, 2008 IEEE International Conference on.

DOI: 10.14267/phd.2015065

Jiang, X., & Tan, A. H. (2010). CRCTOL: A semantic-based domain ontology learning system. *Journal of the American Society for Information Science and Technology, 61*(1), 150-168.

Jivani, A. G. (2011). A comparative study of stemming algorithms. *Int. J. Comp. Tech. Appl, 2*(6), 1930-1938.

Kang, Y.-B., Haghighi, P. D., & Burstein, F. (2014a). CFinder: An Intelligent Key Concept Finder from Text for Ontology Development. *Expert Systems with Applications, 41*(9), 4494–4504.

Kang, Y.-B., Haghighi, P. D., & Burstein, F. (2014b). CFinder: An intelligent key concept finder from text for ontology development. *Expert Systems with Applications, 41*(9), 4494-4504.

Kanis, J., & Müller, L. (2004). *Using the lemmatization technique for phonetic transcription in text-to-speech system.* Paper presented at the Text, speech and dialogue.

Kanis, J., & Müller, L. (2005). *Automatic lemmatizer construction with focus on OOV words lemmatization.* Paper presented at the Text, speech and dialogue.

Kara, S., Alan, Ö., Sabuncu, O., Akpınar, S., Cicekli, N. K., & Alpaslan, F. N. (2012). An ontology-based retrieval system using semantic indexing. *Information Systems, 37*(4), 294-305.

Karanikas, H., Tjortjis, C., & Theodoulidis, B. (2000). *An approach to text mining using information extraction.* Paper presented at the Proc. Knowledge Management Theory Applications Workshop,(KMTA 2000).

Kardan, A. A., Farahmandnia, F., & Omidvar, A. (2013). A novel approach for keyword extraction in learning objects using text mining and WordNet. *Global Journal of Information Technology, 3*(1).

Keim, M. T. (1978). *Running Press Glossary of Insurance Language*: Running Press.

Kemper, B., Matsuzaki, T., Matsuoka, Y., Tsuruoka, Y., Kitano, H., Ananiadou, S., & Tsujii, J. i. (2010). PathText: a text mining integrator for biological pathway visualizations. *Bioinformatics, 26*(12), i374-i381.

Kok, S., & Domingos, P. (2008). Extracting semantic networks from text via relational clustering *Machine Learning and Knowledge Discovery in Databases* (pp. 624-639): Springer.

Labadié, A., & Prince, V. (2008). Lexical and semantic methods in inner text topic segmentation: a comparison between C99 and Transeg *Natural Language and Information Systems* (pp. 347-349): Springer.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes, 25*(2-3), 259-284.

Leaman, R., & Gonzalez, G. (2008). *BANNER: an executable survey of advances in biomedical named entity recognition.* Paper presented at the Pacific Symposium on Biocomputing.

Lefever, E., & Hoste, V. (2010). *Semeval-2010 task 3: Cross-lingual word sense disambiguation.* Paper presented at the Proceedings of the 5th International Workshop on Semantic Evaluation.

Li, Y., Bandar, Z., & McLean, D. (2003). An approach for measuring semantic similarity between words using multiple information sources. *Knowledge and Data Engineering, IEEE Transactions on, 15*(4), 871-882.

DOI: 10.14267/phd.2015065

Li, Y., & Chung, S. M. (2005). *Text document clustering based on frequent word sequences.* Paper presented at the Proceedings of the 14th ACM international conference on Information and knowledge management.

Lo, R. T.-W., He, B., & Ounis, I. (2005). *Automatically building a stopword list for an information retrieval system.* Paper presented at the Journal on Digital Information Management: Special Issue on the 5th Dutch-Belgian Information Retrieval Workshop (DIR).

Lovins, J. B. (1968). *Development of a stemming algorithm*: MIT Information Processing Group, Electronic Systems Laboratory.

Maedche, A., & Staab, S. (2000). *The text-to-onto ontology learning environment.* Paper presented at the Software Demonstration at ICCS-2000-Eight International Conference on Conceptual Structures.

Maedche, A., & Staab, S. (2004). Ontology learning *Handbook on ontologies* (pp. 173-190): Springer.

Majumder, P., Mitra, M., Parui, S. K., Kole, G., Mitra, P., & Datta, K. (2007). YASS: Yet another suffix stripper. *ACM Transactions on Information Systems (TOIS), 25*(4), 18.

Manning, C. D. (2011). Part-of-speech tagging from 97% to 100%: is it time for some linguistics? *Computational Linguistics and Intelligent Text Processing* (pp. 171-189): Springer.

March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. *Decision support systems, 15*(4), 251-266.

Martin C. Pentz, E. a. J. A. M. E., Esq. (2009). *Obligations of Insurer and Policyholder*: Foley Hoag eBook.

Mathiak, B., & Eckstein, S. (2004). *Five steps to text mining in biomedical literature.* Paper presented at the Proceedings of the second European workshop on data mining and text mining in bioinformatics.

Mayfield, J., & McNamee, P. (2003). *Single n-gram stemming.* Paper presented at the Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval.

Meknavin, S., Charoenpornsawat, P., & Kijsirikul, B. (1997). Feature-based thai word segmentation. *Proceedings of of NLPRS*.

Melucci, M., & Orio, N. (2003). *A novel method for stemmer generation based on hidden markov models.* Paper presented at the Proceedings of the twelfth international conference on Information and knowledge management.

Meng, L., Huang, R., & Gu, J. (2013). A review of semantic similarity measures in wordnet. *International Journal of Hybrid Information Technology, 6*(1), 1-12.

Meyer, C. M., & Gurevych, I. (2012). OntoWiktionary: Constructing an Ontology from the. *Semi-Automatic Ontology Development: Processes and Resources: Processes and Resources*, 131.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM, 38*(11), 39-41.

Miyao, Y., Ohta, T., Masuda, K., Tsuruoka, Y., Yoshida, K., Ninomiya, T., & Tsujii, J. i. (2006). *Semantic retrieval for the accurate identification of relational concepts in massive textbases.* Paper presented at the Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics.

Morita, T., Fukuta, N., Izumi, N., & Yamaguchi, T. (2006). DODDLE-OWL: a domain ontology construction tool with OWL *The Semantic Web–ASWC 2006* (pp. 537-551): Springer.

Mykowiecka, A., Marciniak, M., & Kupść, A. (2009). Rule-based information extraction from patients' clinical data. *Journal of biomedical informatics, 42*(5), 923-936.

Nagar, A., & Al-Mubaid, H. (2008). *A new path length measure based on go for gene similarity with evaluation using sgd pathways.* Paper presented at the Computer-Based Medical Systems, 2008. CBMS'08. 21st IEEE International Symposium on.

Nahm, U. Y., & Mooney, R. J. (2002). *Text mining with information extraction.* Paper presented at the AAAI 2002 Spring Symposium on Mining Answers from Texts and Knowledge Bases.

National Food Chain Safety. (March 2012). Introduction of the National Food Chain Safety Office (NFCSO). Retrieved from https://www.nebih.gov.hu/en/

Naz, F., Anwar, W., Bajwa, U. I., & Munir, E. U. (2012). Urdu part of speech tagging using transformation based error driven learning. *World Applied Sciences Journal, 16*(3), 437-448.

Nie, X., & Zhou, J. (2008). *A domain adaptive ontology learning framework.* Paper presented at the Networking, Sensing and Control, 2008. ICNSC 2008. IEEE International Conference on.

Novaković, J., Štrbac, P., & Bulatović, D. (2011). Toward optimal feature selection using ranking methods and classification algorithms. *Yugoslav Journal of Operations Research ISSN: 0354-0243 EISSN: 2334-6043, 21*(1).

Noy, N. F., & Musen, M. A. (2003). The PROMPT suite: interactive tools for ontology merging and mapping. *International journal of human-computer studies, 59*(6), 983-1024.

Noy, N. F., Sintek, M., Decker, S., Crubézy, M., Fergerson, R. W., & Musen, M. A. (2001). Creating semantic web contents with protege-2000. *IEEE intelligent systems*(2), 60-71.

Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N., & Smith, N. A. (2013). *Improved part-of-speech tagging for online conversational text with word clusters.* Paper presented at the Proceedings of NAACL-HLT.

Packer, T. L., & Embley, D. W. (2013). *Cost effective ontology population with data from lists in ocred historical documents.* Paper presented at the Proceedings of the 2nd International Workshop on Historical Document Imaging and Processing.

Paice, C. D. (1990). Another stemmer. *ACM SIGIR Forum, 24*(3), 56-61. doi:10.1145/101306.101310

Paik, J. H., Mitra, M., Parui, S. K., & Järvelin, K. (2011). GRAS: An effective and efficient stemming algorithm for information retrieval. *ACM Transactions on Information Systems (TOIS), 29*(4), 19.

Paik, J. H., & Parui, S. K. (2011). A fast corpus-based stemmer. *ACM Transactions on Asian Language Information Processing (TALIP), 10*(2), 8.

Park, J., Cho, W., & Rho, S. (2010). Evaluating ontology extraction tools using a comprehensive evaluation framework. *Data & Knowledge Engineering, 69*(10), 1043-1061.

DOI: 10.14267/phd.2015065

Pedersen, T., Pakhomov, S. V., Patwardhan, S., & Chute, C. G. (2007). Measures of semantic similarity and relatedness in the biomedical domain. *Journal of biomedical informatics, 40*(3), 288-299.

Peng, W., Li, T., & Ma, S. (2005). *Mining logs files for computing system management.* Paper presented at the Autonomic Computing, 2005. ICAC 2005. Proceedings. Second International Conference on.

Pennete, K. C. (2014). *Leveraging Open Source Tools for Web Mining.* Paper presented at the Proceedings of the 9th International Symposium on Linear Drives for Industry Applications, Volume 2.

Perera, P., & Witte, R. (2005). *A self-learning context-aware lemmatizer for German.* Paper presented at the Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing.

Petasis, G., Karkaletsis, V., Paliouras, G., Krithara, A., & Zavitsanos, E. (2011). *Ontology population and enrichment: State of the art.* Paper presented at the Knowledge-driven multimedia information extraction and ontology evolution.

Pirró, G. (2009). A semantic similarity metric combining features and intrinsic information content. *Data & Knowledge Engineering, 68*(11), 1289-1308.

Plisson, J., Lavrac, N., & Mladenić, D. (2004). A rule based approach to word lemmatization.

Ponzetto, S. P., & Strube, M. (2007). *Deriving a large scale taxonomy from Wikipedia.* Paper presented at the AAAI.

Popowich, F. (2005). Using text mining and natural language processing for health care claims processing. *ACM SIGKDD Explorations Newsletter, 7*(1), 59-66.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program: electronic library and information systems, 14*(3), 130-137.

Porter, M. F. (2001). Snowball: A language for stemming algorithms.

Qin, P., Lu, Z., Yan, Y., & Wu, F. (2009). *A new measure of word semantic similarity based on wordnet hierarchy and dag theory.* Paper presented at the Web Information Systems and Mining, 2009. WISM 2009. International Conference on.

Rada, R., Mili, H., Bicknell, E., & Blettner, M. (1989). Development and application of a metric on semantic nets. *Systems, Man and Cybernetics, IEEE Transactions on, 19*(1), 17-30.

Raunich, S., & Rahm, E. (2011). *ATOM: Automatic target-driven ontology merging.* Paper presented at the Data Engineering (ICDE), 2011 IEEE 27th International Conference on.

Rehman, Z., Anwar, W., Bajwa, U. I., Xuan, W., & Chaoying, Z. (2013). Morpheme Matching Based Text Tokenization for a Scarce Resourced Language. *PloS one, 8*(8), e68178.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007*.

Rogati, M., McCarley, S., & Yang, Y. (2003). *Unsupervised learning of arabic stemming using a parallel corpus.* Paper presented at the Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1.

Rolling Meadows, I. (2007). *Cobit 4.1: Framework; Control Objectives; Management Guidelines; Maturity Models*: IT Governance Institute.

Rose, S. J. (2013). Automatic generation of stop word lists for information retrieval and analysis: Google Patents.

Ruiz-Martınez, J. M., Minarro-Giménez, J. A., Castellanos-Nieves, D., Garcıa-Sánchez, F., & Valencia-Garcia, R. (2011). Ontology population: an application for the E-tourism domain. *International Journal of Innovative Computing, Information and Control (IJICIC), 7*(11), 6115-6134.

Sajay Rai, C., CISSP, CISM. (2014). *Cybersecurity: What the Board of Directors Needs to Ask* (5036.dl). Retrieved from Florida:

Saleena, B., & Srivatsa, S. (2015). Using concept similarity in cross ontology for adaptive e-Learning systems. *Journal of King Saud University-Computer and Information Sciences, 27*(1), 1-12.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management, 24*(5), 513-523.

Sánchez, D., Batet, M., & Isern, D. (2011). Ontology-based information content computation. *Knowledge-Based Systems, 24*(2), 297-303.

Santoso, H. A., Haw, S.-C., & Abdul-Mehdi, Z. T. (2011). Ontology extraction from relational database: Concept hierarchy as background knowledge. *Knowledge-Based Systems, 24*(3), 457-464.

Sarnovský, M., Butka, P., & Paralič, J. (2009). Grid-based support for different text mining tasks. *Acta Polytechnica Hungarica, 6*(4), 5-27.

Schmidt, R. H., & Rodrick, G. E. (2003). *Food safety handbook*: John Wiley & Sons.

Schutz, A., & Buitelaar, P. (2005). Relext: A tool for relation extraction from text in ontology extension *The Semantic Web–ISWC 2005* (pp. 593-606): Springer.

Selig, G. J. (2008). *Implementing IT Governance-A Practical Guide to Global Best Practices in IT Management*: Van Haren.

Shailer, G. E. (2004). *Introduction to Corporate Governance in Australia*: Pearson Education Australia.

Shamsfard, M., & Abdollahzadeh Barforoush, A. (2003). The state of the art in ontology learning: a framework for comparison. *The Knowledge Engineering Review, 18*(04), 293-316.

Slimani, T. (2013). Description and evaluation of semantic similarity measures approaches. *arXiv preprint arXiv:1310.8059*.

Smirnov, I. (2008). Overview of stemming algorithms. *Mechanical Translation, 52*.

Smith, H., & Fingar, P. (2003). *Business process management: the third wave* (Vol. 1): Meghan-Kiffer Press Tampa.

Smoluk, H. (2009). Long-Term Disability Claims Rates and the Consumption-to-Wealth Ratio. *Journal of Risk and Insurance, 76*(1), 109-131.

Sohail, S., & Hassanain, E. (2012). *Arabic Email Spam Detection Techniques and Related Arabic Text Preprocessing Options: A Survey.* Paper presented at the The International Conference on Computing, Networking and Digital Technologies (ICCNDT2012).

Stanford, U. (2014). The Stanford parser: a statistical parser (version 1.6). Retrieved from http://nlp.stanford.edu/software/tagger.shtml

Suchanek, F. M., Ifrim, G., & Weikum, G. (2006). *LEILA: Learning to extract information by linguistic analysis.* Paper presented at the Proceedings of the 2nd Workshop on Ontology Learning and Population: Bridging the Gap between Text and Knowledge.

Suchanek, F. M., Kasneci, G., & Weikum, G. (2008). Yago: A large ontology from wikipedia and wordnet. *Web Semantics: Science, Services and Agents on the World Wide Web, 6*(3), 203-217.

DOI: 10.14267/phd.2015065

Sure, Y., Angele, J., & Staab, S. (2002). OntoEdit: Guiding ontology development by methodology and inferencing *On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE* (pp. 1205-1222): Springer.

Sussna, M. J. (1997). Text retrieval using inference in semantic metanetworks.

Torunoglu, D., Cakirman, E., Ganiz, M. C., Akyokus, S., & Gurbuz, M. (2011). *Analysis of preprocessing methods on classification of Turkish texts.* Paper presented at the Innovations in Intelligent Systems and Applications (INISTA), 2011 International Symposium on.

Turney, P. (2001). Mining the web for synonyms: PMI-IR versus LSA on TOEFL.

Vas, R. (2007). Educational ontology and knowledge testing. *The Electronic Journal of Knowledge Management of, 5*(1), 123-130.

Vaughan, E. J. (1995). *Essentials of Insurance: A risk management perspective*: John Wiley & Sons Incorporated.

Version, T. (2009). 9, The Open Group Architecture Framework (TOGAF). *The Open Group, 1*.

Voutilainen, A. (2003). Part-of-speech tagging. *The Oxford handbook of computational linguistics*, 219-232.

Wang, G., Yu, Y., & Zhu, H. (2007). *Pore: Positive-only relation extraction from wikipedia text*: Springer.

Wang, Y. (2004). Various approaches in text pre-processing. *TM Work Paper No, 2*.

Wang, Y. (2012). " Novel Approaches to Pre-processing Documentbase in Text Classification: Citeseer.

Weber, N., & Buitelaar, P. (2006). *Web-based ontology learning with isolde.* Paper presented at the Proc. of the ISWC Workshop on Web Content Mining with Human Language Technologies.

Wetzstein, B., Ma, Z., Filipowska, A., Kaczmarek, M., Bhiri, S., Losada, S., . . . Cicurel, L. (2007). *Semantic Business Process Management: A Lifecycle Based Requirements Analysis.* Paper presented at the SBPM.

Wilks, Y. (1997). Information extraction as a core language technology *Information Extraction A Multidisciplinary Approach to an Emerging Information Technology* (pp. 1-9): Springer.

Wong, W., Liu, W., & Bennamoun, M. (2007). Tree-traversing ant algorithm for term clustering based on featureless similarities. *Data Mining and Knowledge Discovery, 15*(3), 349-381.

Wong, W., Liu, W., & Bennamoun, M. (2012). Ontology learning from text: A look back and into the future. *ACM Computing Surveys (CSUR), 44*(4), 20.

Wu, F., & Weld, D. S. (2007). *Autonomously semantifying wikipedia.* Paper presented at the Proceedings of the sixteenth ACM conference on Conference on information and knowledge management.

Wu, X., & Bolivar, A. (2008). *Keyword extraction for contextual advertisement.* Paper presented at the Proceedings of the 17th international conference on World Wide Web.

Wu, Z., & Palmer, M. (1994). *Verbs semantics and lexical selection.* Paper presented at the Proceedings of the 32nd annual meeting on Association for Computational Linguistics.

Xu, H., Stenner, S. P., Doan, S., Johnson, K. B., Waitman, L. R., & Denny, J. C. (2010). MedEx: a medication information extraction system for clinical narratives. *Journal of the American Medical Informatics Association, 17*(1), 19-24.

DOI: 10.14267/phd.2015065

Xu, J., & Croft, W. B. (1998). Corpus-based stemming using cooccurrence of word variants. *ACM Transactions on Information Systems (TOIS), 16*(1), 61-81.

Yang, Y., & Pedersen, J. O. (1997). *A comparative study on feature selection in text categorization.* Paper presented at the ICML.

Yates, A. (2004). *Web-scale information extraction in Knowitall.* Paper presented at the Proceedings of the 13th International.

Zablith, F. (2008). Dynamic ontology evolution.

Zhang, C. (2008). Automatic keyword extraction from documents using conditional random fields. *Journal of Computational Information Systems, 4*(3), 1169-1180.

Zhang, Z., & Ciravegna, F. (2011). Named entity recognition for ontology population using background knowledge from Wikipedia. *Ontology Learning and Knowledge Discovery Using the Web: Challenges and Recent Advances. IGI Global*.

Zouaq, A. (2011). An overview of shallow and deep natural language processing for ontology learning. *Ontology learning and knowledge discovery using the web: Challenges and recent advances, 2*, 16-37.

Zouaq, A., Gasevic, D., & Hatala, M. (2011). Towards open ontology learning and filtering. *Information Systems, 36*(7), 1064-1081.

DOI: 10.14267/phd.2015065

Acronyms and terminology

The following glossary is a collection of acronyms and terms with explanations used throughout this dissertation:

Acronyms:

Task	Elementary activity, which has the following attributes • there is always one or more input and one or more output • during the activity, there is a transformation (value creation, but at least something happen, some parameter of the task changes due to the activity • always requires resources (tangible and intangible) • the granularity limited to the human resource required
Process	A set of activities which belong together by any organizational- functional logic. Processes are grouped into process groups, iteratively. Usually we work with no more than four levels. Processes are connected to each other through trigger events.
Job-role	Belongs to one or more position (m:n relation). The activity, what and how one should practices within a task. The same job-role may belong to more than one position.
Job description	A job precisely can be described by • *task description* (procedural description, what to do?) • *competencies* (what to know?) • knowledge • skill • attitude

DOI: 10.14267/phd.2015065

	• *responsibility* (how to make decision, whom to report?)
Information	Data which has *meaning* for the *receiver*
Knowledge	Special set of *information* which gives the *context* from which meaning is derivate
Skill	Ability to do something • hard skills – we hardly can do anything with hard skill soft skills – communication, etc.
Domain Ontology	The domain ontology provides vocabulary of concepts and their relationships captures the activities performed on the theories and elementary principles governing that domain. It is not a glossary of terms, it is what defines the company sphere and represents what the company does.
Job-role knowledge	It is based on the customized sub-domain knowledge. Customization may take place according to actual needs of the given job-role for different purposes (training needs, selection, human resource allocation, etc.)
What is concept?	"A concept is an abstraction or generalization from experience or the result of a transformation of existing concepts. The concept reifies all of its actual or potential instances whether these are things in the real world or other ideas (http://en.wikipedia.org/wiki/Concept)". We use it in connection with the Studio ontology as synonym of node.
What is the difference between concept and information?	In our understanding information is a relation between the data and the observer. It emphasize is there any meaning of data for the observer (receiver) and if yes, what is the meaning. Another question is whether we can measure it, e.g. in terms of information gain. It follows from the definition, there are many

DOI: 10.14267/phd.2015065

	preconditions to evaluate information, not exhaustively: availability, readability of data, the previous, background knowledge (articulated, semi-articulated or hidden character of previous experiences, briefly: context) of the observer (receiver). In other word the ability of the observer (receiver) to process data. Hence come into the picture the ontology which may be a handful auxiliary tool to process (=interpret) data in order to transform information.
Why we want to extract concepts?	The answer more or less relates to the previous question. Extracted pieces of data from the processes handled (interpreted) as concepts in order to build or enhance the ontology. Ontology building is not for torture naive users, but used as a tool for the sake of interpretation, understanding, helping (whom? the user!) to build up and make use of context (cf. answer to Q2) in order to be able to interpret data (=transform to information).
What is the relation of concept with semantics?	"The word semantics itself denotes a range of ideas—from the popular to the highly technical. It is often used in ordinary language for denoting a problem of understanding that comes down to word selection or connotation. This problem of understanding has been the subject of many formal enquiries, over a long period of time, especially in the field of formal semantics. (http://en.wikipedia.org/wiki/Semantics)" I think - at least in our approach - the most important is the problem of understanding, and semantics in almost every definition centers on of this phenomenon. When we build up "our world" from concepts (that is when concepts arranged hierarchically (taxonomy, thesaurus) and are connected through some or many

DOI: 10.14267/phd.2015065

	relations), we do it in order to have a broad and as much as possible exhaustive contextual background. Concept itself nothing to do with semantics, it is assumed as building block. But how the buildings block are arranged, it gives more or different meaning, and this is semantics (in my understanding). With a primitive example: Every single brick looks the same. But you can put together the bricks to form the most wonderful palace - or an ugly prison. The brick is still a brick.
KeyTerms/Words	Key terms are basic terms that are extracted from the input files (XML, TXT). We assume this is the short list of basic terms that are related to a specific domain.
Ontology learning	Ontology learning refers to the process of creating/modifying an ontology in a semi-automatic way.
Ontology enrichment	Ontology enrichment is a process to enrich an existing ontology structure.
Ontology Population	Adding new concepts as new instances to an ontology